# Essays on Parameter Heterogeneity and Model Uncertainty

ISBN: 978 90 3610 213 1

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

This book is no. **489** of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

# Essays on Parameter Heterogeneity

# and

# Model Uncertainty

Essays over parameter heterogeniteit en modelonzekerheid

Thesis

to obtain the degree of Doctor
from Erasmus University Rotterdam
by command of the rector magnificus

Prof.dr. H.G. Schmidt

and in accordance with the decision of Doctoral Board

The public defense shall be held
Thursday 4 November 2010 at 9:30 hours
by
Nalan Baştürk
born in Kocaeli, Turkey.

Doctoral Committee

| | |
|---|---|
| Promoters: | Prof.dr. R. Paap |
| | Prof.dr. D.J.C. van Dijk |
| Other members: | Dr. D. Fok |
| | Prof.dr. G. Koop |
| | Prof.dr. H.K. van Dijk |

# Acknowledgements

Several people deserve my most sincere thanks for their invaluable support during my PhD years and for their direct contribution on this thesis.

First, I would like to thank my promoters Richard Paap and Dick van Dijk for their supervision and contribution in this thesis. Richard has been a very encouraging and understanding supervisor. Despite the complicated research tasks, he always provided motivation and several useful suggestions for the ongoing research. Dick, on the other hand, has provided valuable insights in several parts of this thesis, and he has been a positive and effective supervisor by setting clear goals and deadlines in a very productive manner. I am grateful to Richard and Dick also for encouraging me to work with other researchers, and for their guidance during my learning process of writing papers which (hopefully) do not 'read like a telegram'.

I thank Herman K. van Dijk and Lennart F. Hoogerheide for being co-authors of parts of this thesis and for their valuable guidance and advice. Herman has been an encouraging and inspiring mentor, who not only helped me with research but also with future career plans. I thank Lennart for his extensive contribution to this thesis, for his help in several computational issues, and for his own PhD thesis which I benefited a lot from. I also thank David Ardia for co-authoring parts of this thesis and his fruitful cooperation.

I am indebted to the other members of the inner committee, Dennis Fok and Gary Koop, for reviewing the earlier draft of this thesis given their busy schedule and for their feedback and comments including ideas for future research.

I would like to take this opportunity to thank the faculty and the staff of Tinbergen and Econometric Institutes for being so approachable and willing to help in several research and administration related matters. I am honored to have the opportunity to continue working in such a hospitable and friendly environment at the Econometric Institute after my PhD.

I thank my colleagues Anne, Cem and Lukazs for our productive discussions about the similar topics we worked on. I also thank my office-mate Rianne and I am sorry for my accidental attempts to lock her in the office.

I am grateful to my parents Selma and Arif, and my brother Eyüp for supporting me not only during my PhD years, but throughout my life. It is priceless to know that I have a welcoming family back home, regardless of the circumstances. (*Doktora yıllarında ve hayatım boyunca beni destekleyen annem, babam ve kardeşime teşekkür ederim. İyi ki varsınız.*)

During my PhD years, I was lucky to have many friends around. Eda, Burcu, Duygu, Bengu and Selen, thank you for your moral support from far away. I would like to thank Umut, Milan, Cem, Simon, Anil, Viorel, Nuno, Carlos, Ayça, Murat, Nüfer, Çerağ, Esen, Victor and all my collegues for making these demanding years cheerful. Thanks to Simon for helping me prepare the summary of this thesis in Dutch, Viorel for the inspiring lunch conversations, and Ayça for always reminding me that there is more to life than my computer screen.

Last but not least, I thank my paranymphs Duygu and Rui for their valued friendship. Duygu, I am amazed that you are willing to come all the way from Turkey to be with me on this important day. You continue surprising me, even after knowing you for 14 years. Rui, thanks for your invaluable help and support on several occasions, for putting up with my occasional grumpiness throughout the process, and for many needed distractions.

Nalan Başẗurk
Rotterdam, September 2010

# Contents

# Chapter 1

# Introduction and Outline

## 1.1 Introduction

The choice of a particular model in quantitative economic analysis reflects the economic question analyzed, jointly with related economic theory and the specific structure of the given data being analyzed. The degree to which economic theory or the data dominates the analysis is an important strategic decision that the researcher has to face. In the first strategy, the model is based mainly on *a priori* economic theory. Several contributions in the economics literature, in particular those that occurred in the period just after the second World War, are based on this strategy, suggesting explicit links between economic theory, mathematics and statistics (see e.g. the contributions of the Cowles Foundation for Research in Economics at Yale University. In the second strategy, which became more popular during late nineteen seventies and early nineteen eighties, modeling is based more on the data information, see e.g. Sims (1980). In the time series context, the advantages of this data-based approach are addressed and it is mentioned that economic theory often does not provide precise information on functional relationships between variables. A good survey of this approach is given by Zellner and Palm (2004). These latter authors conclude that the use of data information for discovering and repairing the defects of proposed models are of crucial importance.

Common practice in empirical research is to combine these strategies in a meaningful way, i.e. the constructed model is based on economic theory and the data information at the same time. This combination of strategies is motivated by two arguments: On the one hand, data information may not be informative enough. On the other hand, too strong assumptions may affect the reliability of results and the forecasting performance. This thesis considers the relatively more data-based approach in analyzing economic relationships and provides alternative methods to avoid very strong assumptions in the analysis. Note

that the focus on data-based methods by no means neglects the importance of economic theory in providing guidelines for the empirical analysis.

One approach to avoid strong assumptions in modeling is to employ flexible models which allow for heterogeneities in certain model parameters relying on the data information instead of a priori specifications. The use of such flexible models has been considered in several areas of economic research. A specific example is the analysis of countries' economic development measured by real Gross Domestic Product (GDP) per capita, where it is well documented that GDP patterns of countries show substantial heterogeneity both in terms of GDP levels and GDP growth rates (Quah, 1996; Paap and Van Dijk, 1998).

A second approach to reduce the effects of underlying model assumptions is to consider different models that are possibly suitable for the data, and to determine the extent to which each of these models adequately describes the data. The methods to assess model performance are rather different from models allowing for parameter heterogeneities, in the sense that the performance of each candidate model can be taken into account. Further, these alternative models can be combined in order to account for the inherent model uncertainty explicitly. In economics, this approach has been applied mainly in the forecasting literature, see e.g. Min and Zellner (1993).

This thesis consists of two parts. The first part develops new econometric models with a sufficient degree of flexibility to accommodate various forms and degrees of heterogeneity in (the relations among) economic variables. The second part considers model uncertainty issues providing new tools for evaluating to what extent one (or more) model is suitable to the economic data at hand.

The first part of this thesis, comprising Chapters 2 and 3, builds on existing work employing relatively data-driven models for analyzing countries' economic performance. Several papers, such as Paap and Van Dijk (1998), Hobijn and Franses (2000), Bloom *et al.* (2003), and Phillips and Sul (2009) document that homogeneity assumptions are unrealistic when analyzing economic growth. These chapters develop novel methodologies to infer differences between countries' growth experiences, and the sources of these differences. The second part of the thesis, comprising Chapters 4 and 5, focuses on issues related to assessing model fit and model comparison, with applications to several datasets. In particular, these chapters develop new tools to accurately assess model fit from a Bayesian perspective.

Section 1.2 briefly summarizes the methods employed in this thesis. Section 1.3 provides a detailed outline of the individual chapters, together with a summary of the main theoretical and empirical findings and the main contributions of the research presented in this thesis.

## 1.2 Methodology

Different methodologies are employed in separate chapters of this thesis, according to the research question. Despite this heterogeneity, one common method in all chapters is the use of finite mixture distributions.

A finite mixture distribution is a probabilistic approach for density estimation using a finite number of densities (see Frühwirth-Schnatter (2006) for detailed explanations and extensions of these models). Finite mixtures can be used to analyze parameter heterogeneity as well as model uncertainty. In terms of assessing parameter heterogeneity, mixture models can be used to define the hidden segments in the data (Feller, 1943). These hidden segments may be defined by different effects of the conditioning variables across segments, as well as heterogenous unobserved paths that the dependent variable follows. This interpretation of finite mixtures is used in Chapters 2 and 3, where finite mixture models are utilized to identify heterogeneity in the effects of conditioning variables and unobserved variables, respectively.

Another common area where finite mixtures are shown to be useful is density approximation. The purpose in density estimation using finite mixtures is rather different from segmenting the data: Even if the data generating process is not exactly a finite mixture process, a non-trivial density resulting from the data can be approximated using mixtures of known densities, provided that the number of components is not limited (Ferguson, 1973). The main focus in density estimation using finite mixtures is approximating a target density, instead of identifying the members of each component of the mixture process. In Chapter 4, we use finite mixture models for this purpose. Specifically, we consider the AdMit approach (Hoogerheide *et al.*, 2007b), which is defined as a finite mixture of Student-$t$ densities, for approximating the marginal likelihood.

A further possibility to utilize finite mixtures is related to the case when two or more alternative models can be suitable for the data. Bayesian Model Averaging (BMA) is one such method where the uncertainty in the model is explicitly taken into account using mixtures of models (Leamer, 1978). In Chapter 5, BMA and finite mixture distributions are utilized in order to account for model uncertainty.

Finally, the inference in most of the analyses in this thesis makes use of Bayesian methods. One motivation for employing these methods is the natural interpretation of the Bayesian approach to parameter and model uncertainty: Parameters or models do not have to be fixed, but rather their distributions are defined in a probabilistic way. Second, from a practical point of view, estimation of some of the complex models proposed in this thesis are rather straightforward using simulation based Bayesian methods (and

conditional distributions of parameters). For detailed discussions on the advantages of the Bayesian approach, see Zellner (1971); Kim and Nelson (1999); Koop (2003) among others.

Section 1.3 provides a detailed summary of the main theoretical and empirical findings, together with the implementation of finite mixtures in each chapter.

## 1.3 Outline and Summary

This thesis consists of four self-contained chapters on recent methods for modeling parameter heterogeneity and model comparison tools. These tools are applied to a variety of economic questions varying from macroeconomic issues, such as heterogeneity in growth determinants across countries, to microeconomic issues, such as the effects of education on labor income of individuals.

Chapter 2 is based on Baştürk, Paap, and Van Dijk (2008). This chapter provides a systematic analysis of possible heterogeneity in the effects of determinants of real GDP per capita growth rates across countries. This approach is motivated by previous research documenting differences in growth determinants across countries (e.g. Barro, 1991; Easterly and Levine, 1997; Collier and Gunning, 1999). Despite this consensus on existing differences across countries, neither theoretical nor empirical studies pinpoint a preferred method to address these differences. In the literature, it is common to capture heterogeneity by defining groups (*clusters*) of countries with (presumably) different growth characteristics a priori, for example, based on geographical location or other country-specific characteristics. The results of these studies differ depending on the preferred method.

Chapter 2 proposes a novel methodology that groups the countries according to the differing effects of growth determinants themselves. The data is allowed to determine which countries belong to which cluster and also how many clusters there are. Consequently, a priori groupings using regional dummies or other variables are special cases of possible models compared in this method.

The approach to handle parameter heterogeneity in growth determinants builds on the endogenous clustering approach of Hobijn and Franses (2000), and the finite mixture modeling approach developed by Paap, Franses, and Van Dijk (2005). Similar to the latter analysis, a finite mixture panel data model is specified for the annual real GDP per capita growth rates of countries. This approach is extended by clustering countries not only according to their average growth rates but also according to the effects of growth determinants.

For a set of 59 countries in Asia, Latin and Middle America, and Africa for the period 1971–2000, the estimation results indicate two groups (clusters) of countries in terms of different effects of conventional determinants of economic growth. This result is in line with the literature documenting differences across countries. Furthermore, the identified segmentation of countries are compared with conventional clustering variables used in the literature, that is, initial GDP levels, initial human capital measures, geographical divisions and initial openness measures. None of these variables are found to provide a clear relationship with the data-based cluster memberships, confirming the importance of forming country clusters endogenously.

Regarding the effects of growth determinants, the results show that the structural differences between the countries in the two clusters are given by different effects of investment measures (gross domestic investment, and price of investment), openness measures (total trade as a percentage of GDP, and real exchange rate distortions), and government share of the economy. On the other hand, conditional on the covariates, the mean growth rates are not found to be significantly different for the two clusters of countries.

Chapter 3 is based on Baştürk, Paap, and Van Dijk (2010). This chapter focuses on the growth differences across countries from a different perspective, and presents a model to analyze real GDP per capita levels for a panel of countries accounting for changing intra-distributional dynamics. A regime switching model is proposed to identify groups of countries (convergence clubs) that show similar GDP structures, while allowing for changes in club memberships over time. Further, the effects of initial conditions and financial development on the formation of these clubs are analyzed.

Our methodology for clustering the data is related to the mixture model employed in Chapter 2 with the distinction of clustering of GDP levels, instead of GDP growth rates, as well as allowing for certain covariates to affect the changes in the club memberships and short-run GDP fluctuations.

In terms of the methodology employed, Frühwirth-Schnatter and Kaufmann (2008) and Hamilton and Owyang (2009) are the two papers closest to this paper. Both studies propose regime switching models to assess the subgroups of the data that show similar characteristics. Our model builds on these models by modeling the time-dependent data characteristics. In order to estimate the model, a Bayesian approach is pursued, and the Gibbs Sampling algorithm is used to obtain posterior results (Geman and Geman, 1984b).

In terms of convergence clubs, three distinct groups are detected, and club memberships are found to be quite persistent, but still group compositions change substantially over time. In particular, several EU member countries and East Asian countries are found to belong to a higher GDP club in recent times compared to the beginning of the

1970s. Regarding the determinants of club memberships, initial conditions are found to be important factors. For the effects of financial development indicators on the GDP process, our results confirm the theoretical basis for different effects of financial development indicators in the short-run and the long-run. In the long-run, financial development is found to affect the countries' GDP level positively, i.e. the probability of belonging to a club with a higher GDP path increases with the level of financial development. In the short-run however, the estimated effect of stock market development on GDP levels are in general negative.

Chapter 4 is based on Ardia, Baştürk, Hoogerheide, and Van Dijk (2010). This chapter considers common Monte Carlo methods to estimate marginal likelihood, and addresses various details in implementation, as well as the robustness and efficiency of these methods. The chapter builds on the overviews of Han and Carlin (2001), Frühwirth-Schnatter (2004) and Miazhynskaia and Dorffner (2006) on Monte Carlo estimators of marginal likelihoods. These studies are extended by focusing on the particular case of non-elliptical posterior distributions.

Regarding the marginal likelihoods, a set of common estimators are compared: Chib and Jeliazkov (2001) (CJ), the importance sampling (IS) estimator (Hammersley and Handscomb, 1964; Kloek and Van Dijk, 1978; Van Dijk and Kloek, 1980; Geweke, 1989), bridge sampling (BS) estimator (Meng and Wong, 1996) and the reciprocal importance sampling (RIS) estimator (Gelfand and Dey, 1994). The performance of these estimators are evaluated in combination with two other techniques: the approach of Hoogerheide *et al.* (2007b) that constructs an adaptive mixture of Student-t distributions (AdMit) to obtain an appropriate candidate density, and the *warping* method (Meng and Schilling, 2002).

Marginal likelihood estimation methods are compared for two different model structures leading to non-elliptical posterior shapes: a non-linear regression model and a mixture GARCH model. The first application considers the biochemical oxygen demand (BOD) data from Marske (1967) that are analyzed by Bates and Watts (1988) and Ritter and Tanner (1992). The second application considers data from Ausín and Galeano (2007), concerning daily returns on the Swiss Stock Market Index (SMI).

For both models, it is shown that the IS estimator provides gains in efficiency compared to the alternative methods. These gains are more pronounced when an appropriate candidate distribution is chosen using the AdMit approach of Hoogerheide *et al.* (2007b). Combining the estimation methods with the 'warping' method of Meng and Schilling (2002) is found to improve the efficiency of the estimates. However, employing the AdMit

approach is found to provide substantially higher increase in efficiency compared to the 'warping method'.

Chapter 5 is based on Baştürk, Hoogerheide, and Van Dijk (2010). This chapter analyzes monetary returns to education in two datasets, accounting for uncertainty in the model that best characterizes the income-education relationship. It is argued that the simple regression of years of education on individuals' earnings can lead to erroneous results because of *endogeneity* issues, which may stem from omitted intellectual capabilities, measurement errors in reported earnings and education, or simultaneity as individuals can determine the amount of education they receive judging the possible monetary returns (Angrist *et al.*, 1996). The typical analysis of the income-education relationship makes use of Instrumental Variable (IV) models.

IV models rely on a set of instruments (proxies) that explain the so-called endogenous variable. In the income-education relationship, several instruments are proposed for the endogenous variable, education. For example Angrist and Krueger (1991) proposes quarter of birth of individuals as instruments for the level of education in the US. Their choice of instruments is based on schooling laws on compulsory education, which allows individuals to end their education only at a certain age. As an alternative, family background variables have been proposed as instruments for education, as these variables are expected to affect the decisions of an individual regarding education (Parker and Van Praag, 2006; Trostel *et al.*, 2002; Hoogerheide *et al.*, 2010).

The purpose of this chapter is to assess the evidence for alternative models for the income-education relationship, addressing some of the underlying assumptions in standard IV models, such as the degree of endogeneity in the dataset, the strength of instruments in explaining education, validity of proposed instruments, and possibly differing effects of education on income across subsets of the data.

The methodology employed in this chapter is closest to Bayesian methods to deal with model uncertainty, such as (Clyde and George, 2004). In particular, the *predictive likelihood* approach is used to document evidence for alternative models, instead of the conventional methods based on marginal likelihoods. This approach has been successfully applied in forecasting (Min and Zellner, 1993; Eklund and Karlsson, 2007; Geweke and Amisano, 2010). These studies are extended by employing predictive likelihoods in assessing model fit in IV models. The proposed method is applied to two datasets on the income-education relationship, namely Angrist and Krueger (1991) data and data from German Socio-Economic Panel Survey (SOEP), recently analyzed in Hoogerheide *et al.* (2010).

The contributions of this study are as follows: First, it is demonstrated that the evidence for alternative IV models can be assessed using predictive likelihoods instead of marginal likelihoods, and BMA using predictive likelihoods provides a tool for efficient measurement of the marginal effects of regressors.

Second, for the US data on the income-education relationship, the issues of weak instruments and the degree of endogeneity are addressed. Two alternative models are considered: one which assumes that education levels do not suffer from endogeneity, and a second model proposing quarter of birth as instruments for education. Based on the evidence of these models, it is shown that choosing one of these models can be problematic, and a Bayesian Model Averaging (BMA) approach based on predictive likelihoods can provide gains in estimating the effect of education on income. Furthermore, it is documented that the data shows significant heterogeneity across states and regions both in terms of strength of instruments and in terms of the degree of endogeneity in the education levels.

Finally, for the SOEP data on the income-education relationship, two issues are addressed: the issue of possible endogeneity of the instruments, namely father's education level, and possible heterogenous effects of education on income. The effects of education on income are found to be different across male and female respondents, leading to a negative correlation between education level and gender discrimination in earnings. On the other hand, monetary returns to education are found to be higher for the period after the year 2001. It is also shown that these findings are relatively insensitive to the possibility of endogenous instruments.

# Chapter 2

# Structural Differences in Economic Growth: An Endogenous Clustering Approach

*Chapter 2 is based on Baştürk, Paap, and Van Dijk (2008).*

## 2.1 Introduction

The empirical literature on the analysis of growth determinants has provided substantial evidence for the existence of variations in growth patterns across countries. Despite this common finding, there is no agreed way of incorporating these variations in econometric models. There are several possibilities to allow for cross-country heterogeneity in the effects of growth determinants, and existing growth theories do not pinpoint a preferred method. It is common to capture heterogeneity by defining groups of countries with (presumably) different growth characteristics a priori, for example, based on geographical location. Using this approach, a number of empirical studies have examined whether growth patterns are different for countries in sub-Saharan Africa and East Asia (e.g. Barro, 1991; Easterly and Levine, 1997; Collier and Gunning, 1999), for landlocked countries (Sachs and Warner, 1997; Bloom *et al.*, 2003), or for former colonies (Barro, 1999).

Durlauf (2000) suggests that modeling cross-country heterogeneity is one of the main challenges in the current empirical growth literature, and points out two problems arising from the failure to control for heterogeneity in the right way. First, ad-hoc country groupings may simply be incorrect, in the sense that they may differ substantially from the true grouping. Second, most studies only allow variation in the intercept (which

corresponds with the mean growth rate conditional on the included regressors), while restricting the effects of variables such as inflation and investment to be the same across (groups of) countries. Obviously, this assumption is quite restrictive. In fact, one of the most interesting questions is whether and how the effects of such growth determinants are country specific. Hence, it is emphasized that empirical research should focus on analyzing and documenting the heterogeneities in the countries' growth processes, see also Durlauf (2007).

In line with Durlauf's arguments, several recent empirical studies refrain from grouping countries beforehand, and mainly let the data classify the countries into clusters with distinct growth patterns. Two main approaches can be distinguished within these data-based clustering methods.

In the first approach, countries are grouped according to the values of one or more selected covariates. In the simplest possible case with two groups and a single covariate, the countries are assigned to one of the two groups depending on whether the value of the so-called splitting variable is below or above a certain threshold (see e.g. Durlauf and Johnson, 1995a; Kalaitzidakis *et al.*, 2001; Hansen, 2000; Cuaresma and Doppelhofer, 2007). A possible drawback of this approach is that the splitting variable(s) still has (have) to be determined a priori. This is avoided in the second approach, called endogenous clustering (see e.g. Hobijn and Franses, 2000; Paap, Franses, and Van Dijk, 2005; Davis, Owen, and Videras, 2009). In this approach the clustering essentially is assumed to be a latent endogenous process and hence it is completely data-driven. The only prior assumption that needs to be made is that each country has some probability of getting assigned to a cluster. Within such a cluster countries have the same economic growth pattern, while this is different across the clusters. The data is allowed to determine which countries belong to which cluster and also how many clusters there are. Therefore the existence and the identification of heterogeneity of growth determinants is done without any prior specifications: We note that, consequently, any classification using regional dummies or other splitting variables are special cases of possible models compared in endogenous clustering.

Using the endogenous clustering approach, this chapter aims to examine whether there are structural differences across countries in Asia, Latin and Middle America, and Africa in terms of their growth determinants, and if so, to identify the sources of these differences. We extend the clustering approach that is typically used in the convergence literature and in previous analysis of growth rate differences. Specifically, the countries are clustered not only according to their average growth rates but also according to the effects of growth

determinants. Our method provides a systematic analysis of heterogeneity in growth patterns by allowing all regressors to have different marginal effects across clusters.

Our methodology is close to the endogenous clustering approach of Alfo *et al.* (2008) and Davis *et al.* (2009), where countries are sorted into regimes depending on unobserved characteristics. Their approach is more general than the other endogenous clustering approaches such as Durlauf and Johnson (1995a) and Ardic (2006) in the sense that the country groupings and parameter heterogeneities are analyzed simultaneously, rather than identifying the country groups beforehand.

In particular, similar to Alfo *et al.* (2008) and Davis *et al.* (2009) we use a finite mixture model to examine the growth rate differences between countries. Both studies use 5 year non-overlapping averages of data in order to prevent the effects of strong correlation in the dependent variable following Bond *et al.* (2001)[1]. However, it is well documented that averaging the data might lead to accumulated error terms, and amplify the autocorrelation problem. One of the main differences in our study is the use of annual growth rates instead of 5 year averages in order to refrain from accumulated error terms in the dependent variable. In order to account for the cross-country correlations in growth rates, we follow the novel approach of (Paap *et al.*, 2005) to eliminate the cross-country correlation in the error terms. Note that the degrees of freedom problem is partially avoided using annual data, and contrary to the existing clustering studies, we can compare the finite mixture models with country-specific growth models. Hence our analysis does not necessarily restrict the growth patterns to a finite mixture model.

We apply our clustering approach to an unbalanced panel of annual growth rates for 59 countries from Asia, Africa, and Latin and Middle America for the period 1971 to 2000. We find 2 clusters of countries in terms of different marginal effects of growth determinants. The resulting finite mixture panel model outperforms a homogenous growth regression model for all countries, as well as country-specific growth regressions.

Our estimation results show that the structural differences between the countries in the two clusters are caused by different marginal effects of investment measures (gross domestic investment, and price of investment), openness measures (total trade as a percentage of GDP, and real exchange rate distortions), and government share of the economy. On the other hand, conditional on the covariates, the mean growth rates are not found to be significantly different for the two clusters of countries.

We compare the identified cluster memberships of the countries with conventional clustering variables used in the literature, that is, initial GDP levels, initial human capital

[1]Alfo *et al.* (2008) also consider 10 and 25 year averages of the data and stress the problem of country heterogeneity being masked particularly in the latter case.

measures, and initial openness measures. None of these variables are found to provide a clear relationship with the data-based cluster memberships. Furthermore, the clusters do not show a clear geographical division either.

The remainder of this chapter is organized as follows. In Section 5.2 we discuss the panel finite mixture panel model which we use for endogenous clustering. We also discuss several important aspects of the empirical model specification procedure, as well as the algorithm for parameter estimation. In Section 2.5 we present the data, while we discuss the empirical results in Section 3.4. We conclude in Section 5.5.

## 2.2 Finite Mixture Panel Model

Our approach to handle parameter heterogeneity builds upon the finite mixture modeling approach developed by Paap, Franses, and Van Dijk (2005). They propose a model in which all regressors are assumed to have different parameters across clusters. We extend their model in order to allow for a subset of regressors to have common marginal effects across clusters. This formulation allows us to use Likelihood Ratio (LR) tests to check for overparametrization, i.e. to test common versus heterogenous effects of regressors across clusters. In addition, if some regressors are found to have common effects across clusters, more efficient estimates can be obtained as incorporating this restriction into the model reduces the number of parameters to be estimated.

The growth rates of real GDP per capita for $N$ countries are assumed to be a mixture of $J$ distributions or clusters, each defined by a homogenous model. Let $s_i \in \{1, \ldots, J\}$ denote the cluster which country $i$ belongs to, for $i = 1, \ldots, N$. We assume that $s_i$ is unknown and has to be estimated from the data. A priori, there is a constant probability that country $i$ belongs to cluster $j$. For $j \in \{1, \ldots, J\}$, this cluster membership probability is given by $p_j = \Pr[s_i = j]$, where $p_j \in (0, 1)$ and $\sum_{j=1}^{J} p_j = 1$ by definition.

Given $J$ and $s_i$, we consider the following regression model for the growth rate of real GDP per capita $g_{i,t}$ of country $i = 1, \ldots, N$ in year $t = 1, \ldots, T$:

$$g_{i,t} = w_{i,t}'\gamma + x_{i,t}'\beta_{s_i} + z_{i,t}'\alpha_i + \varepsilon_{i,t}, \tag{2.1}$$

where $\varepsilon_{i,t} \sim NID\,(0, \sigma_i^2)$.

Unlike conventional growth equations, which mainly define models for cross-country data and fixed effects for certain groups of countries, (2.1) defines a panel data model with slope heterogeneity depending on three sets of regressors, that is, $x$, $w$ and $z$. First, the regressors in the $k_w \times 1$ vector $w_{i,t}$ have the same marginal effects across both clusters and countries. Second, the regressors in the $k_x \times 1$ vector $x_{i,t}$ have different effects across

clusters, but the same effects for all countries within a given cluster. Hence, the parameters associated with these variables specify the structural differences in the distribution of the dependent variable across clusters. Third, the variables in the $k_z \times 1$ vector $z_{i,t}$ have different marginal effects across countries even within the same cluster. The vectors $w_{i,t}$, $x_{i,t}$ and $z_{i,t}$ are said to contain the common variables, cluster-dependent variables, and country-specific variables, respectively.

At first sight, it seems that the slope heterogeneity in (2.1) is defined by the marginal effects of the regressors in the $x_{i,t}$ and $z_{i,t}$ vectors. However, it should be noted that the country-specific regressors in the vector $z_{i,t}$ in fact do not aim to capture such heterogeneity. Although growth determinants could possibly have different effects for all countries, the finite mixture model can capture such heterogeneities completely through the regressors in $x_{i,t}$, without the vector $z_{i,t}$, by taking the number of clusters $J$ equal to the number of countries $N$. Instead, in the general model in (2.1) we include the vector $z_{i,t}$ to capture the cross-country error correlations in the growth regression. The way to implement this is to define a regressor in $z$ that has the same values for all countries within a time period, i.e. $z_{i,t} = z_t$ for all $i = 1, \ldots, N$.

## 2.3 Model specification

There are two important issues when using the model in (2.1). First, the number of mixture components $J$ has to be determined from the data in order to deal with parameter heterogeneity in a general way. Second, one has to classify the regressors into the three different types, that is, the vectors $w_{i,t}$, $x_{i,t}$ and $z_{i,t}$.

To determine $J$, standard tests are not applicable. It is well documented that in case of finite mixture models, the number of clusters cannot be selected using standard tests, due to the presence of unidentified nuisance parameters. For example, when testing the null hypothesis of $J$ clusters against the alternative of $J+1$ clusters, the unrestricted log-likelihood function for such a test is not bounded under the null and, consequently, the asymptotic distribution of the LR statistic is not $\chi^2$. The two most common ways to deal with this problem in the literature are to use parametric or non-parametric bootstraps (see e.g. Turner, 2000; Wedel, 2002, for discussion), or to rely upon information criteria. Examining several information criteria on simulated data, Jedidi *et al.* (1997) show that the consistent Akaike Information Criterion - CAIC (Bozdogan, 1987) and Bayesian Information Criterion - BIC (Schwarz, 1978) have the best performance in case of mixture models. Following their results, we use CAIC and BIC for determining the number of clusters. Therefore, the model parameters are first estimated for fixed $J$, and the informa-

tion criteria for different values of $J$ are compared to decide upon the appropriate number of clusters.

For a given value $J$ we can use standard tests to classify the regressors into the $w_{i,t}$, $x_{i,t}$ and $z_{i,t}$ vectors. Given the data-based nature of the finite mixture modeling approach, we suggest to follow a data-driven procedure to define these vectors.

Given a particular choice of covariates to be included in the model, one possibility is to initialize the model in (2.1) by including all covariates in the country-specific covariate vector $z_{i,t}$, and then testing for common coefficients across countries. This would lead to the endogenous clustering approach followed by Hobijn and Franses (2000). Note that this approach essentially means that we first estimate the model for $J = N$, and then try to reduce the number of clusters by imposing suitable parameter restrictions. This approach would require a substantial data set if the number of regressors is fairly large, as in our case. Furthermore, it is difficult to control for the overall size of the sequential testing procedure.

An alternative approach is to start from a homogenous linear model, and then test for different marginal effects of the regressors. In this case, all regressors are put in the vector $w_{i,t}$ initially, which is then tested against more general models with part of the regressors put in the vector(s) $x_{i,t}$ (or $z_{i,t}$). However, the choice of the cluster-specific regressors in $x$ (or $z$) is not obvious, and as a result, all restricted models potentially have omitted variable bias in this case.

Given these considerations, we use a general-to-specific approach in terms of parameter heterogeneity. All covariates are initially assumed to be cluster-dependent and enter the $x_{i,t}$ vector (i.e. we start with an empty vector $w_{i,t}$ vector for all $i$ and $t$), and the optimal number of clusters in this model is determined using the information criteria. In a second step, given the number of clusters we test for common marginal effects for each of the regressors separately. Using the test results we impose the appropriate parameter restrictions and we estimate a restricted model with part of the regressors moved from $x_{i,t}$ to $w_{i,t}$.

## 2.4 Parameter estimation

The parameters of the finite mixture panel model can be estimated using Maximum Likelihood (ML). Since we are dealing with a finite mixture model and cluster memberships of the individual countries are unknown, the Expectation Maximization (EM) method by Dempster *et al.* (1977) is a convenient way to maximize the likelihood function. To derive the steps of the EM algorithm, we first consider the complete data likelihood function, for

which the cluster indicators $s_i$ are assumed to be observed. The complete data likelihood function is given by

$$l(g, s; \theta) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left( p_j \prod_{t=1}^{T} \frac{1}{\sigma_i} \phi \left( \frac{\varepsilon_{i,t}^{(j)}}{\sigma_i} \right) \right)^{I(s_i=j)}, \tag{2.2}$$

where $I(.)$ is the indicator function which takes the value 1 if the argument is true, and zero otherwise. $\phi(.)$ is the standard normal density function, and the cluster-specific error term is given by

$$\varepsilon_{i,t}^{(j)} = g_{i,t} - w'_{i,t}\gamma - x'_{i,t}\beta_j - z'_{i,t}\alpha_i. \tag{2.3}$$

The EM algorithm is an iterative algorithm which consists of two steps, that is, an expectation step followed by a maximization step. In the expectation step, the expected value of the complete data log-likelihood function with respect to the missing or unobserved data is computed. In the finite mixture model, the cluster indicators, $s_i$ for $i = 1, ..., N$, are unobserved. Hence, in this case the expectation of the log of the complete data likelihood function (2.2) with respect to these latent variables (conditional on the observed variables) is given by

$$L(g; \theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij}^{*} \left( \ln (p_j) - \frac{T}{2} \ln \sigma_i^2 - \frac{T}{2} \ln 2\pi - \sum_{t=1}^{T} \frac{\left( \varepsilon_{i,t}^{(j)} \right)^2}{2\sigma_i^2} \right), \tag{2.4}$$

where the expected cluster probabilities $p_{ij}^{*}$ are defined as follows:

$$p_{ij}^{*} = \frac{p_j \prod_{t=1}^{T} \frac{1}{\sigma_i} \phi \left( \frac{\varepsilon_{i,t}^{(j)}}{\sigma_i} \right)}{\sum_{l=1}^{J} p_l \prod_{t=1}^{T} \frac{1}{\sigma_i} \phi \left( \frac{\varepsilon_{i,t}^{(l)}}{\sigma_i} \right)}. \tag{2.5}$$

In the maximization step, the expected log-likelihood function in (2.4) is maximized with respect to the model parameters $p_j$ and $\beta_j$ for $j = 1, ..., J$, $\alpha_i$ and $\sigma_i^2$ for $i = 1, ..., N$, and $\gamma$. The first-order conditions for maximization are derived in 2.A.2. The E- and M-steps are repeated until convergence. The resulting values of the parameters are the ML estimates.

The ML parameters can be used to estimate the value of $s_i$ given the data, for $i = 1, \ldots, N$. This estimate is equal to the expected cluster membership probability (2.5) evaluated at the ML estimates. Hence, $p_{ij}^{*}$ provides the posterior probability that country $i$ belongs to cluster $j$. It can be seen from (2.4) that each observation is weighted according to these posterior probabilities in the objective function. Hence the estimated cluster

memberships are not taken as fixed while estimating the regression parameters, unlike the exogenous clustering methods used in the growth literature. The uncertainty in the estimated cluster memberships is also taken into account in parameter estimation and inference.

## 2.5 Data

Our data set consists of annual observations for an unbalanced panel of 59 countries in Asia, Latin and Middle America and Africa covering the period 1971-2000. The countries are selected according to data availability, where we require observations to be available for at least half of the sample period. The list of included countries is given in 2.A.1.

The regressors included in (2.1) cover variables that have traditionally been considered as important determinants of economic growth. Specifically, the explanatory variables we use are (i) human capital, measured by the logarithm of secondary school enrollment as a percentage of the population over 25 years; (ii) the annual growth rate of the population between 15 and 65 years; (iii) the logarithm of total trade as a percentage of GDP and real exchange rate distortions as proxies for openness measures; (iv) annual inflation as proxy for macroeconomic stability; (v) the government share of GDP in percent; (vi) the logarithm of the price of investment and Gross Domestic Investment (GDI) as a percentage of GDP.[2] The last variable is not included in some studies for endogeneity reasons. We include this regressor but do make sure to employ endogeneity checks following the approach of Barro (1996).

The dependent variable is the annual growth rate of real GDP per capita, which is obtained from the Penn World Tables version 6.2 (PWT 6.2). The government share of GDP, price of investment and GDI variables are also taken from PWT 6.2. Real exchange rate distortions, trade percentage and inflation variables are obtained from the Global Development Network Growth Database, which in turn uses the World Development Indicators (WDI), and Global Development and Finance databases. For the labor force growth, we use the WDI database for population between 15-65 years. Secondary school enrollment percentages in the population over 25 years are taken from Barro and Lee (2000). Their educational data is available mostly for 5-year intervals, and we obtain annual observations by using spline interpolation.

[2]This choice of regressors is by no means exclusive. Some regressors used in several growth regressions such as population density or squared inflation are not included in the model as a result of data availability or the presence of high multicollinearity with the other regressors.

We emphasize that we do not include any dummy variables or country-specific factors in this model. These variables are commonly employed in growth regressions to capture the heterogeneity in the mean growth rates. Instead, in our finite mixture approach, any heterogeneity in mean growth rates as well as in the marginal effects of regressors are completely determined by the data. We will however investigate if the endogenously determined clusters of countries correspond to, for example, regional dummies or country-specific factors.

## 2.6 Empirical Results

We estimate the finite mixture panel data model presented in Section 5.2 for the annual real GDP per capita growth rates of 59 countries over the period 1971-2000, including the growth determinants discussed in the previous section. For model specification we follow the general-to-specific approach outlined before. Hence, in terms of the notation in Section 5.2, initially all regressors are assumed to have cluster-dependent marginal effects, and are included in the vector $x_{i,t}$. We refer to this specification, for which $w_{i,t}$ is empty, as the 'general model'. As discussed in Section 5.2, the variables $z_{i,t}$ can be used to capture any remaining cross-country correlation in the annual growth rates. Here we follow Paap, Franses, and Van Dijk (2005), and include US real GDP per capita growth rate in $z_{i,t}$ for this purpose. As they note, this variable can be seen as representing the "world business cycle". Finally, the regressors are demeaned such that the intercepts correspond with average growth rates.

The first step in the analysis is to determine the number of clusters, $J$. For this purpose, we estimate the finite mixture model for 2 to 7 clusters,[3] as well as a linear model ($J = 1$) where the growth equation is homogenous for all countries. Finally, we also consider a model where all countries are analyzed separately rather than making any parameter homogeneity assumptions across countries. This last case, where the growth equation is different for all countries, corresponds to $J = 59$.

The results in Table 2.1 show that both BIC and CAIC indicate a clear preference for a model with $J = 2$ clusters. Note in particular that the finite mixture model is preferred over a homogenous growth rate equation for the included countries ($J = 1$). This result is in line with other studies on heterogeneity of growth determinants, such as Kalaitzidakis *et al.* (2001), Hansen (2000), and Cuaresma and Doppelhofer (2007). Furthermore, we see

[3]The EM algorithm may converge to a local maximum. To prevent reporting local maximum results, we use 4000 different random starting cluster probabilities. For all considered models, the estimation results belonging to the highest log-likelihood value are reported.

that the finite mixture model with 2 clusters also performs better than the country-specific growth regressions ($J = 59$) in terms of the information criteria.

Table 2.1: Information criteria for different number of clusters

| $J$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 59 |
|---|---|---|---|---|---|---|---|---|
| *Based on number of cross-sections (59 observations)* | | | | | | | | |
| BIC | -65.67 | -66.63* | -66.50 | -66.36 | -66.18 | -65.88 | -65.50 | -44.26 |
| CAIC | -63.52 | -64.30* | -64.00 | -63.70 | -63.35 | -62.89 | -62.33 | -33.26 |
| *Based on total number of observations (1482 observations)* | | | | | | | | |
| BIC | -2.34 | -2.35* | -2.33 | -2.30 | -2.27 | -2.24 | -2.20 | -0.35 |
| CAIC | -2.25 | -2.26* | -2.23 | -2.19 | -2.16 | -2.21 | -2.07 | 0.09 |

*Note*: The table presents values of the Bayesian Information Criterion (BIC) and consistent Akaike Information Criterion (CAIC) for the finite mixture model (2.1) with $J$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. For simplification, information criteria are divided by the number of observations in all cases. The model with smallest information criteria is to be preferred. An asterisk indicates the minimum value of the information criteria.

Table 2.2 reports the parameter estimates for the finite mixture model with $J = 2$ clusters, along with the LR tests for the joint significance of the coefficients of a particular regressor in both clusters (4th column) and for equal marginal effects across clusters (5th column). Both the LR tests for joint significance and the individual $t$-statistics indicate that except for the population growth rate all variables have statistically significant coefficients at the 5% level in at least one of the clusters. Hence, apart from the population growth rate, we find that all included regressors are important determinants of economic growth.

The most interesting aspect of the model of course concerns the differences in the marginal effects of the regressors across clusters. Recall that the variables with distinct coefficients identify the structural differences between the countries in the two clusters. Based on the $p$-values of the LR tests for common marginal effects reported in the final column of Table 2.2, we find that the parameters differ significantly across clusters for investment, real exchange rate distortions, trade percentage, government share and investment price.

Several structural differences in growth patterns between the countries in the two clusters are apparent from the estimation results. First, growth is much more sensitive to investment for countries in cluster 2 compared to those in cluster 1. Although GDI has a positive effect on growth in both clusters, the effect is almost twice as large in cluster 2.

Furthermore, we do not find a significant coefficient for the price of investment in cluster 1, whereas this variable has a significant positive effect on growth for countries in cluster 2.

The second difference between the clusters is in terms of the marginal effects of openness variables. For both openness measures, that is, total trade percentage and real exchange rate distortions, the marginal effects clearly differ across the clusters. For trade, we find significantly positive and negative effects on growth for the first and second clusters, respectively. Hence, trade openness is beneficial for economic growth in cluster 1

Table 2.2: Estimation results for the general finite mixture model with $J = 2$ clusters

| Variable | Cluster 1 | Cluster 2 | $p$-value[a] (joint significance) | $p$-value[b] (common marg. effects) |
|---|---|---|---|---|
| intercept | 4.211** | 3.164** | 0.000 | 0.322 |
| | (0.667) | (0.687) | | |
| investment | 2.820** | 5.470** | 0.000 | 0.004 |
| | (0.640) | (0.670) | | |
| school enr. | -2.929** | -1.684** | 0.000 | 0.104 |
| | (0.508) | (0.431) | | |
| pop. growth | 29.028 | -7.450 | 0.193 | 0.382 |
| | (21.247) | (35.103) | | |
| RER distor. | 0.000 | -0.004** | 0.007 | 0.000 |
| | (0.000) | (0.001) | | |
| trade% | 3.846** | -1.435** | 0.000 | 0.000 |
| | (0.742) | (0.653) | | |
| inflation | 0.000 | -0.002** | 0.035 | 0.111 |
| | (0.000) | (0.001) | | |
| govt. share | -6.314** | 1.710* | 0.000 | 0.000 |
| | (0.903) | (0.999) | | |
| invest. price | -1.168 | 3.281** | 0.019 | 0.000 |
| | (0.739) | (0.796) | | |

*Note*: The table shows parameter estimates with standard errors in parentheses for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. All regressors are allowed to have different marginal effects across clusters. The regressors are demeaned such that the intercepts correspond with average growth rates. Parameter estimates and standard errors are multiplied by 100. * and ** denote significance at 10% and 5% levels, respectively.

[a] Asymptotic $p$-values for the LR tests for the joint significance of the parameters in both segments.

[b] Asymptotic $p$-values for the LR tests for equal marginal effects in both clusters.

countries, but it depresses economic growth for cluster 2 countries. For the exchange rate distortions on the other hand, the marginal effect on growth is not significant for cluster 1, while we find a significantly negative coefficient for the countries in cluster 2.

Third, fiscal policy, measured by the government share in GDP, also has different effects across clusters. For the countries in cluster 1, an increase in the government share in the economy has a significantly negative effect on growth, indicating that the government sector in these countries is relatively less efficient compared to the private sector. For cluster 2 on the other hand, government share does not have a significant effect on growth.

The human capital variable, that is, secondary school enrollment rates in the population over 25, has a negative and significant coefficient for both clusters. This result is rather surprising since we would expect schooling in the working age population to stimulate growth through human capital accumulation[4].

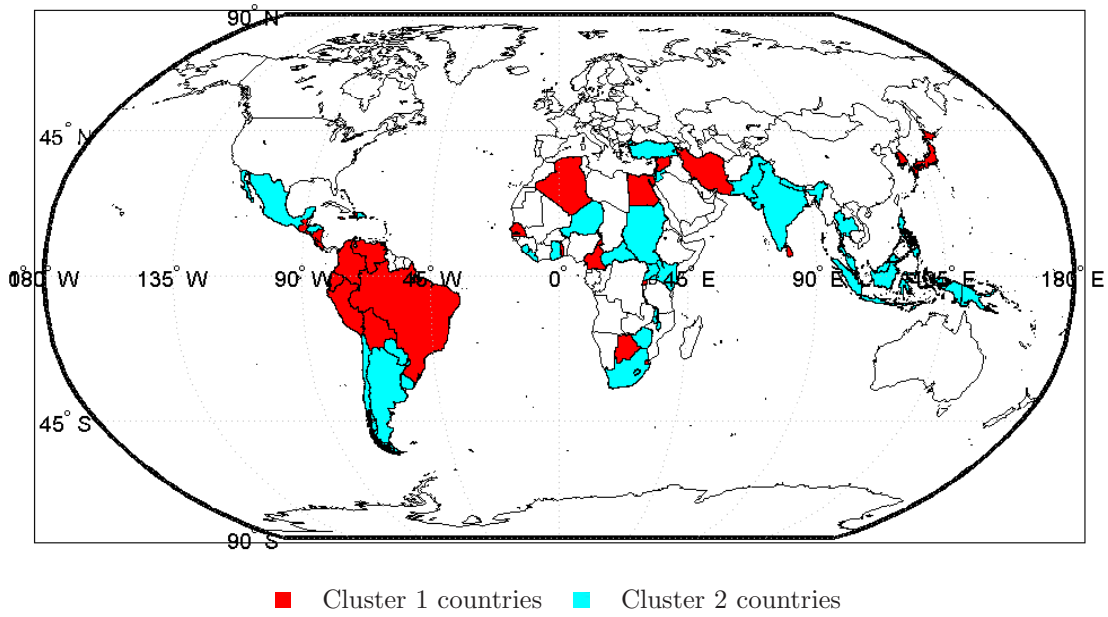
Finally, it is interesting to point out that, although the estimated intercept in Table 2.2 is higher for cluster 1, the difference with the mean growth rate in cluster 2 is not statistically significant. This result is quite different from conventional growth studies, which only allow the mean growth rate to vary across countries but restrict the marginal effects of other regressors to be the same. Hence, the significant differences in mean growth rates reported in such studies may in fact be due to the different marginal effects of growth determinants, as uncovered by our finite mixture model.

Figure 2.1 and Table 2.3 show the cluster memberships for the countries included in our data set based on the posterior cluster membership probabilities. Here, country $i$ is said to belong to cluster $j$ if its estimated posterior probability of being in this cluster based on (2.5) is greater than 0.5. The figure and table show that the division based on posterior cluster membership probabilities does not match with regional specifications, especially for Africa and Latin and Middle America. Most Asian countries are in cluster 2, for which returns to investment are relatively higher, and monetary instability measured by annual inflation has a negative effect on economic growth.

A further comparison of the clustering implied by the finite mixture model and the regional segmentation often applied in the literature is given in Table 2.4, showing the average posterior cluster probabilities for the total sample, as well as for the countries in the three geographical regions that are represented in our sample. The average probabilities per region are calculated from (2.5) averaged over the countries in each region.

[4]The correlation between the human capital and real GDP growth rate is also negative when these two variables are analyzed in a separate regression, that is, when real GDP per capita growth rates are regressed on a constant and secondary school enrollment rates.

Figure 2.1: Posterior cluster memberships for the general model




*Note*: The figure shows posterior cluster membership in the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates over the period 1971-2000. All regressors are allowed to have different marginal effects across clusters. Posterior cluster membership probabilities are given in Table 2.A.3.1. Each country is assigned to the cluster with the highest posterior probability.

Asian countries have the highest probability to belong to the second cluster, while the reverse holds for the countries in Latin and Middle America.[5] Furthermore, there is no clear pattern for the countries in sub-Saharan Africa, as their average probability to belong to cluster 2 is close to 50% in Table 2.4. We conclude that there are parameter heterogeneities across the countries considered, but these heterogeneities do not match with conventional regional divisions.

Next, we consider the possibility to make the model more parsimonious by imposing the restriction of common marginal effects across clusters for some regressors. The results of the individual LR tests on common marginal effects displayed in the final column of Table 2.2 suggests that this restriction may be imposed for school enrollment, population growth and inflation. The joint LR test for common marginal effects of all three regressors

[5]From the countries considered, one can argue that the performance of Japan over the time span considered is rather different from the remaining countries. For this reason we also estimated the models excluding Japan from the sample. The general results in terms of the optimal number of clusters, the signs and the significance of the explanatory variables remain the same.

Table 2.3: Posterior clustering for the general finite mixture model with $J = 2$ clusters

| | |
|---|---|
| | *Africa* |
| Cluster 1 | Algeria, Botswana, Cameroon, Egypt, Gambia, Lesotho, Mauritius, Rwanda, Senegal, Swaziland, Togo |
| Cluster 2 | Central African Rep., Ghana, Kenya, Liberia, Malawi, Niger, Sierra Leone, South Africa, Sudan, Uganda, Zimbabwe |
| | *Latin and Middle America* |
| Cluster 1 | Barbados, Bolivia, Brazil, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Haiti, Jamaica, Nicaragua, Peru, Trinidad&Tobago, Venezuela |
| Cluster 2 | Argentina, Chile, Dominican Rep., Honduras, Mexico, Paraguay, Uruguay |
| | *Asia* |
| Cluster 1 | Iran, Japan, Korea Rep., Sri Lanka, Syria |
| Cluster 2 | India, Indonesia, Israel, Jordan, Malaysia, Nepal, Pakistan, Philippines, Papua New Guinea, Thailand, Turkey |

*Note*: The table presents posterior clustering for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. All regressors are allowed to have different marginal effects across clusters. The clustering is based on posterior cluster membership probabilities shown in Table 2.A.3.1. Each country is assigned to the cluster with the highest posterior probability.

Table 2.4: Average posterior cluster probabilities per region for the general model

| | *Cluster probabilities* | |
|---|---|---|
| Sample | cluster 1 | cluster 2 |
| All countries | 0.49 | 0.51 |
| | (0.08) | |
| Africa | 0.46 | 0.54 |
| Asia | 0.36 | 0.64 |
| Latin and Middle America | 0.64 | 0.36 |
| Sub-Saharan Africa | 0.41 | 0.59 |

*Note*: The table presents posterior clustering for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. All regressors are allowed to have different marginal effects across clusters. The average cluster probabilities per region are calculated using the posterior cluster membership probabilities shown in Table 2.A.3.1. The standard error for cluster 1 probability is shown in parentheses.

equals 5.32 with a $p$-value of 0.15. Hence, we do not reject common marginal effects for these regressors jointly.

Based on these results, we estimate a restricted model where school enrollment, population growth and annual inflation have the same marginal effects for all countries considered. Using the notation in Section 5.2, we now put these variables in the $w_{i,t}$ vector. The parameter estimates for this 'restricted model' are given in Table 2.5.

Table 2.5: Estimation results for the restricted model with $J = 2$ clusters

| *Cluster dependent variables* | | | *Cluster independent variables* | |
|---|---|---|---|---|
| Variable | Coefficient Estimates | | Variable | Coefficient Estimate |
| | Cluster 1 | Cluster 2 | | |
| intercept | 3.688 * | 3.642* | school enr. | -2.379* |
| | (0.555) | (0.595) | | (0.368) |
| investment | 2.419* | 5.745* | pop. growth | 19.574 |
| | (0.604) | (0.693) | | (16.958) |
| RER distor. | 0.000 | -0.004* | inflation | 0.000 |
| | (0.000) | (0.001) | | (0.000) |
| trade% | 3.575* | -1.225* | | |
| | (0.692) | (0.617) | | |
| govt. share | -6.277* | 1.721** | | |
| | (0.900) | (1.053) | | |
| invest. price | -1.222** | 3.630* | | |
| | (0.627) | (0.855) | | |

*Note*: The table shows parameter estimates with standard errors in parentheses for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates over the period 1971-2000. Secondary school enrollment, population growth rate and inflation are imposed to have the same marginal effects across clusters. All regressors are demeaned prior to analysis. Parameter estimates and standard errors are multiplied by 100. * and ** denote significance at 5% and 10% levels, respectively.

The results in terms of the cluster-dependent variables are in line with the general, unrestricted model: The estimated mean growth rate is larger for the countries in cluster 1 and returns to investment are lower. Countries in cluster 2 are negatively affected by real exchange rate distortions, while this variable does not have a significant effect for countries in cluster 1. Similar to previous results, cluster 1 is characterized by a negative effect of government size in the economy on growth.

Next, we consider the regressors with homogenous effects in the restricted regression displayed in the final columns of Table 2.5. The common effect of schooling is negative and

significant. Similar to the previous $t$-tests and LR tests, we do not find a significant effect of population growth. In terms of monetary policy stability, when inflation is analyzed as a common regressor, the marginal effect on growth is negligible for all countries in the data set.

Figure 2.2 shows the country clusters based on the posterior cluster membership probabilities for the restricted model. Although the parameter estimates for the clusters are similar, five countries, namely Gambia, Haiti, Nepal, Papua New Guinea and Zimbabwe, are in different clusters compared to the general model. Note that for all these countries except Zimbabwe, the posterior probabilities reported in Table 2.A.3.1 are relatively close to 0.5. Hence the cluster membership probabilities for these countries are not very informative.

Table 2.6 reports the average cluster probabilities for geographical groups for the restricted model. The results in terms of the average regional patterns hold in the restricted model as well: Although Asian countries have the highest probability to belong to cluster 2, the estimated clusters do not match with geographical divisions.

Table 2.6: Average cluster membership probabilities per region for the restricted model

| | *Cluster probabilities* | |
|---|---|---|
| Sample | cluster 1 | cluster 2 |
| All countries | 0.47 | 0.53 |
| | (0.08) | |
| Africa | 0.52 | 0.48 |
| Asia | 0.41 | 0.59 |
| Latin and Middle America | 0.62 | 0.38 |
| Sub-Saharan Africa | 0.48 | 0.52 |

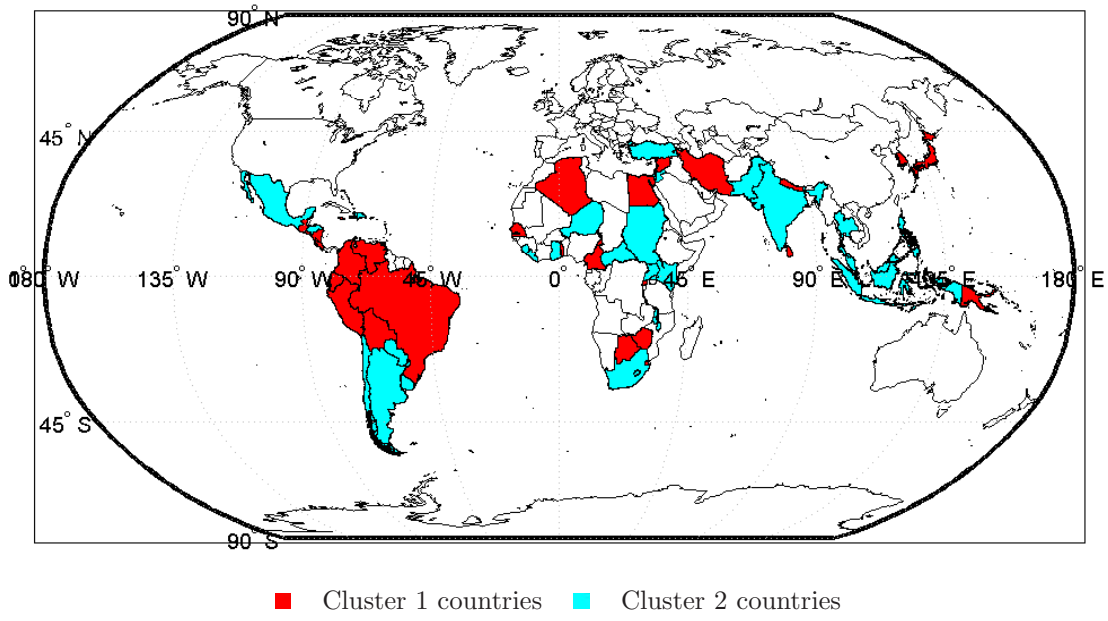
*Note*: The table presents posterior clustering for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. Secondary school enrollment, population growth rate and inflation are assumed to have the same marginal effects across clusters. The average cluster probabilities per region are calculated using the posterior cluster membership probabilities shown in Table 2.A.3.2. Each country is assigned to the cluster with the highest posterior probability. The standard error for cluster 1 probability is shown in parentheses.

## 2.7 Robustness checks

Starting from the general model with $J = 2$ clusters, we perform three additional checks to examine the robustness of our results. First, we test for endogeneity of investment. Second, we check whether there are regional patterns that our results cannot cover. Finally, we compare the estimated clusters with the threshold variables used in the literature in order to see whether the finite mixture model is just an approximation for a model with a threshold specification on the regressors.

For the endogeneity of investment, we follow the approach of Barro (1996). He proposes a simple comparison to check for reverse causality using the lagged value of investment as a regressor. If the model with lagged investment does not lead to significant parameter estimates, we should conclude that the causality is from growth to investment, and there is an endogeneity problem in the estimation of the finite mixture model. Ta-

Figure 2.2: Posterior cluster memberships for the restricted model




*Note*: The figure shows posterior cluster membership in the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates over the period 1971-2000. Secondary school enrollment, population growth rate and inflation are assumed to have the same marginal effects across clusters, while the rest of the explanatory variables have different marginal effects across clusters. Posterior cluster membership probabilities are given in Table 2.A.3.2. Each country is assigned to the cluster with the highest posterior probability.

ble 2.7 presents the estimation results with 2 clusters where lagged value of Gross Domestic Investment is used as a regressor instead of the contemporaneous value.

Table 2.7: Estimation results for the model containing lagged investment

| Variable | Coefficient Estimates Cluster 1 | Cluster 2 | Variable | Coefficient Estimates Cluster 1 | Cluster 2 |
|---|---|---|---|---|---|
| intercept | 3.438 | 2.575 | trade% | 8.113 | -0.898 |
| | (0.718) | (0.569) | | (0.999) | (0.566) |
| investment$_{-1}$ | 1.898* | 2.013* | inflation | 0.000 | -0.002 |
| | (0.728) | (0.561) | | (0.000) | (0.001) |
| school enr. | -4.721 | -1.197 | govt. share | -6.368 | -0.228 |
| | (0.607) | (0.435) | | (0.181) | (0.916) |
| pop. growth | -73.327 | 64.050 | invest. price | 1.021 | 0.623 |
| | (34.914) | (19.020) | | (0.880) | (0.643) |
| RER distor. | 0.000 | -0.004 | | | |
| | (0.000) | (0.001) | | | |

*Note*: The table shows parameter estimates with standard errors in parentheses for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. Parameter estimates and standard errors (in parentheses) are multiplied by 100. The Gross Domestic Investment value in the previous period is denoted by investment$_{-1}$. * indicates significance at 5% level for lagged investment.

Table 2.7 shows that the marginal effect of the lagged investment variable is positive and significant in both clusters. Hence the analysis does not indicate an endogeneity problem in investment.

In order to check whether there are regional patterns that our results cannot cover, we estimate a model including the conventional regional dummy variables in the literature. Specifically, we estimate the model including dummy variables for the East Asian and sub-Saharan African countries. Parameter estimates for this model are given in Table 2.8. Note that a significant coefficient for the dummy variables would indicate that there are regional patterns in growth that the finite mixture model with 2 clusters cannot uncover. The results in Table 2.8 show that both dummy variables do not have a significant effect on growth. Therefore, we do not find any indications for unexplained regional patterns in terms of sub-Saharan Africa or East Asia.

Finally, we examine the relation of the endogenous clustering results with some threshold variables. The literature using exogenous clustering for growth rates mainly considers three threshold variables for heterogeneity in growth rates, namely initial GDP per capita, openness and schooling measures, see e.g. Durlauf and Johnson (1995a) and Cuaresma and

Table 2.8: Estimation results for the model containing regional dummies

| *Cluster dependent explanatory variables* | | | | | |
|---|---|---|---|---|---|
| Variable | Cluster 1 | Cluster 2 | Variable | Cluster 1 | Cluster 2 |
| intercept | 4.357 | 3.557 | trade% | 3.994 | -1.284 |
| | (0.684) | (0.769) | | (0.779) | (0.657) |
| investment | 2.827 | 5.323 | inflation | 0.000 | -0.002 |
| | (0.647) | (0.728) | | (0.000) | (0.001) |
| school enr. | -3.020 | -1.691 | govt. share | -6.370 | 1.512 |
| | (0.532) | (0.427) | | (0.906) | (1.035) |
| pop. growth | 27.355 | -4.700 | invest. price | -1.123 | 3.366 |
| | (21.045) | (35.308) | | (0.746) | (0.783) |
| RER distor. | 0.000 | -0.004 | | | |
| | (0.000) | (0.001) | | | |

| *Cluster independent explanatory variables* | |
|---|---|
| Variable | Coefficient estimate |
| East Asia dummy | -0.676 |
| | (1.282) |
| Sub-Saharan Africa dummy | -0.747 |
| | (1.082) |

*Note*: The table shows parameter estimates with standard errors in parentheses for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. Regional dummy variables for Sub-Saharan Africa and East Asia are added as cluster-independent regressors. Parameter estimates and standard errors (in parentheses) are multiplied by 100.

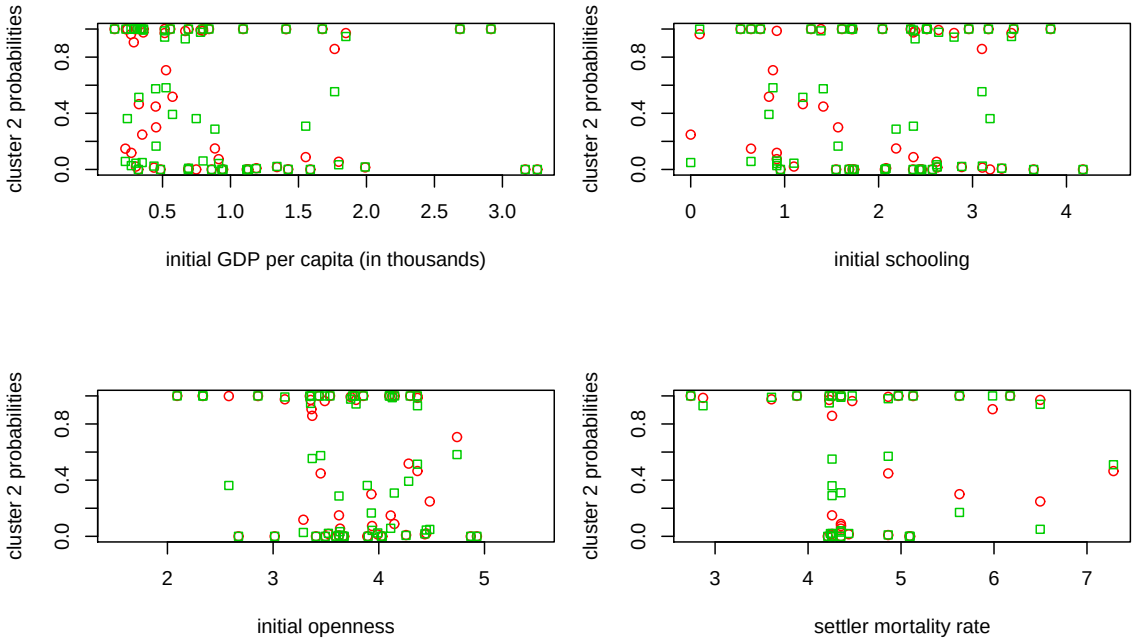
Doppelhofer (2007). Furthermore, Davis *et al.* (2009) argue that the quality of institutions can be seen as the factor grouping the countries. Hence we also check the relationship between institutional quality variable and the estimated clusters. Figure 2.3 shows scatter diagrams of the estimated cluster probabilities and these threshold variables for the models with two clusters. An accurate threshold variable would imply that the cluster probabilities below a certain threshold are smaller than 0.5 while cluster probabilities above the threshold are larger than 0.5, or vice versa.[6]

[6]Initial openness measure is the trade percentage at the beginning of the sample period, and for initial schooling we use secondary school enrollment rates in the population over 25. Similar to Davis *et al.* (2009), we use European settler mortality rates by Acemoglu *et al.* (2001) (revised by Albouy (2008)) as a proxy for institutional quality.

The scatter diagrams do not show a clear relationship between the threshold variables and the cluster probabilities. Hence, the finite mixture model is not just an approximation for a model with threshold variables. In other words, the thresholds of initial GDP per capita, initial openness, initial schooling, and institutional quality do not capture country heterogeneities accurately for this data set[7][8].

We conclude that within the data set we have considered, there are two groups of countries with different marginal effects of the variables affecting growth. The estimated country classification is different from conventional segmentations based on geographical location or threshold variables. The resulting model does not seem to suffer from an endogeneity problem in terms of investment or omitted parameter heterogeneities.

Figure 2.3: Comparisons of cluster probabilities with threshold variables




*Note*: The figures show cluster 2 probabilities for the general model (circles) and the restricted model (squares) on the y-axes against the threshold variables on the x-axes. Initial schooling and settler mortality measures are in natural logarithms.

[7]The comparisons with other institutional quality proxies in Albouy (2008), such as latitude, also do not show a clear relationship with the cluster memberships.

[8]The methodology by Lee and Kim (2009) can also be seen as an exogenous clustering approach using income levels as a threshold variable. They consider different growth patterns across countries depending on the World Bank criterion for the upper middle income countries. However our results do not indicate such a classification of countries either.

## 2.8 Conclusion

Using a finite mixture model and endogenous clustering, we analyze the structural differences in economic growth rates for 59 countries in Latin and Middle America, Asia and Africa for the period 1971-2000. The countries are not grouped beforehand according to, for example, geographical location or the (relative) value of certain covariates. The structural differences and the country groups are rather determined endogenously. The model allows for heterogeneities in the marginal effects of all considered variables affecting growth.

The analysis leads to two important conclusions. First, in line with many previous studies, the results indicate structural differences in growth patterns across countries. The included countries are optimally divided into two groups according to these structural differences. The optimal groups do not match with the conventional regional classifications in the literature. For all three regions that are distinguished, we find a substantial number of countries in both clusters. In particular, we find evidence against treating African countries as a homogenous group which is different from the rest of the developing world. Moreover, we find that threshold variables such as initial GDP levels, human capital levels, openness measures, or proxies for institutional quality, also do not explain the heterogeneities for the included countries accurately.

Second, the results show that the structural differences between the countries are in terms of the marginal effects of several regressors: investment (both gross domestic investment and the price of investment), openness measures (total trade as a percentage of GDP, and real exchange rate distortions), and the government share in GDP all have heterogenous effects between the country clusters. In addition, we do not find significant differences in the mean growth rate across clusters.

In our future work, we intend to account for model uncertainty in economic growth while using this systematic way to deal with parameter heterogeneity. Specifically, we aim to investigate the model uncertainties without assigning a priori groups of countries with homogenous growth structures.

# 2.A Appendices

## 2.A.1 List of Included Countries

**Africa**: Algeria, Botswana, Cameroon, Central African Rep., Egypt, Gambia, Ghana, Kenya, Liberia, Lesotho, Malawi, Mauritius, Niger, Rwanda, Senegal, Sierra Leone, South Africa, Sudan, Swaziland, Togo, Uganda, Zimbabwe.

**Latin and Middle America**: Argentina, Barbados, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Paraguay, Peru, Trinidad&Tobago, Uruguay, Venezuela.

**Asia**:[9] India, Indonesia, Iran, Israel, Japan, Jordan, Korea Rep., Malaysia, Nepal, Pakistan, Philippines, Papua New Guinea, Sri Lanka, Syria, Thailand, Turkey.

## 2.A.2 EM Algorithm

As starting point of the algorithm we take the complete data likelihood function (2.2). Hence, the complete data log-likelihood function is

$$L\left(g,s;\theta\right)=\sum_{i=1}^{N}\sum_{j=1}^{J}I\left(s_i=j\right)\left(\ln\left(p_j\right)+\sum_{t=1}^{T}\ln\left(\frac{1}{\sigma_i}\phi\left(\frac{\varepsilon_{i,t}^{(j)}}{\sigma_i}\right)\right)\right). \tag{2.6}$$

The conditional (on the data and parameters) probability for country $i$ to be included in cluster $j$ is given by the ratio of country $i$'s likelihood contributions to the $J$ segments, that is,

$$p_{ij}^{*}=\frac{p_j\prod_{t=1}^{T}\frac{1}{\sigma_i}\phi\left(\frac{\varepsilon_{i,t}^{(j)}}{\sigma_i}\right)}{\sum_{l=1}^{J}p_l\prod_{t=1}^{T}\frac{1}{\sigma_i}\phi\left(\frac{\varepsilon_{i,t}^{(l)}}{\sigma_i}\right)}\qquad\text{for } j=1,...,J. \tag{2.7}$$

Hence, the expected value of the complete data log-likelihood function [E-step] is

$$L\left(g;\theta\right)=\sum_{i=1}^{N}\sum_{j=1}^{J}p_{ij}^{*}\left(\ln\left(p_j\right)-\frac{T}{2}\ln\sigma_i^2-\frac{T}{2}\ln 2\pi-\sum_{t=1}^{T}\frac{\left(\varepsilon_{i,t}^{(j)}\right)^2}{2\sigma_i^2}\right). \tag{2.8}$$

[9] This group consists of Middle Eastern and Asian countries. We refer to this group as 'Asia'.

The first-order conditions for maximizing (2.8) [M-step] are given by

$$\frac{\partial L\left(g;\theta\right)}{\partial \beta_j} = \sum_{i=1}^{N} \frac{p_{ij}^*}{\sigma_i^2} \sum_{t=1}^{T} x_{i,t}\varepsilon_{i,t}^{(j)} = 0 \qquad \text{for } j = 1, ..., J, \tag{2.9}$$

$$\frac{\partial L\left(g;\theta\right)}{\partial \alpha_i} = \sum_{j=1}^{J} \frac{p_{ij}^*}{\sigma_i^2} \sum_{t=1}^{T} z_{i,t}\varepsilon_{i,t}^{(j)} = 0 \qquad \text{for } i = 1, ..., N, \tag{2.10}$$

$$\frac{\partial L\left(g;\theta\right)}{\partial \gamma} = \sum_{i=1}^{N} \sum_{j=1}^{J} \frac{p_{ij}^*}{\sigma_i^2} \sum_{t=1}^{T} w_{i,t}\varepsilon_{i,t}^{(j)} = 0 \tag{2.11}$$

$$\frac{\partial L\left(g;\theta\right)}{\partial \sigma_i^2} = \sum_{j=1}^{J} \frac{p_{ij}^*}{\sigma_i^2} \left( -\frac{T}{2\sigma_i^2} + \sum_{t=1}^{T} \frac{\left(\varepsilon_{i,t}^{(j)}\right)^2}{2\sigma_i^4} \right) = 0 \qquad \text{for } i = 1, ..., N. \tag{2.12}$$

The solution to these first-order conditions provides an update of the parameter estimates. The cluster membership probabilities are updated using

$$p_j = \frac{1}{N} \sum_{i=1}^{N} p_{ij}^*, \tag{2.13}$$

The E- and M-step are repeated until convergence is achieved. The resulting parameter values are equal to the ML estimates.

## 2.A.3 Posterior Cluster Membership Probabilities

Table 2.A.3.1: Posterior cluster membership probabilities for the model in Table 2.2

| Cluster 1 Countries | | | | | |
|---|---|---|---|---|---|
| Country | $p_{i1}^*$ | $p_{i2}^*$ | Country | $p_{i1}^*$ | $p_{i2}^*$ |
| Algeria | 0.93 | 0.07 | Jamaica | 0.99 | 0.01 |
| Barbados | 1.00 | 0.00 | Japan | 1.00 | 0.00 |
| Bolivia | 1.00 | 0.00 | Korea Republic of | 0.99 | 0.01 |
| Botswana | 0.98 | 0.02 | Lesotho | 0.85 | 0.15 |
| Brazil | 1.00 | 0.00 | Mauritius | 1.00 | 0.00 |
| Cameroon | 0.70 | 0.30 | Nicaragua | 1.00 | 0.00 |
| Colombia | 1.00 | 0.00 | Peru | 0.98 | 0.02 |
| Costa Rica | 0.91 | 0.09 | Rwanda | 0.88 | 0.12 |
| Ecuador | 0.85 | 0.15 | Senegal | 1.00 | 0.00 |
| Egypt | 1.00 | 0.00 | Sri Lanka | 0.99 | 0.01 |
| El Salvador | 1.00 | 0.00 | Swaziland | 1.00 | 0.00 |
| Gambia, The | 0.54 | 0.46 | Syria | 1.00 | 0.00 |
| Guatemala | 1.00 | 0.00 | Togo | 0.75 | 0.25 |
| Haiti | 0.55 | 0.45 | Trinidad &Tobago | 0.99 | 0.01 |
| Iran | 1.00 | 0.00 | Venezuela | 0.95 | 0.05 |
| **Cluster 2 Countries** | | | | | |
| Country | $p_{i1}^*$ | $p_{i2}^*$ | Country | $p_{i1}^*$ | $p_{i2}^*$ |
| Argentina | 0.00 | 1.00 | Nepal | 0.00 | 1.00 |
| Central African Republic | 0.00 | 1.00 | Niger | 0.09 | 0.91 |
| Chile | 0.03 | 0.97 | Pakistan | 0.02 | 0.98 |
| Dominican Republic | 0.01 | 0.99 | Papua New Guinea | 0.48 | 0.52 |
| Ghana | 0.03 | 0.97 | Paraguay | 0.00 | 1.00 |
| Honduras | 0.00 | 1.00 | Philippines | 0.00 | 1.00 |
| India | 0.00 | 1.00 | Sierra Leone | 0.00 | 1.00 |
| Indonesia | 0.00 | 1.00 | South Africa | 0.00 | 1.00 |
| Israel | 0.00 | 1.00 | Sudan | 0.04 | 0.96 |
| Jordan | 0.00 | 1.00 | Thailand | 0.00 | 1.00 |
| Kenya | 0.00 | 1.00 | Turkey | 0.00 | 1.00 |
| Liberia | 0.29 | 0.71 | Uganda | 0.00 | 1.00 |
| Malawi | 0.00 | 1.00 | Uruguay | 0.14 | 0.86 |
| Malaysia | 0.01 | 0.99 | Zimbabwe | 0.01 | 0.99 |
| Mexico | 0.00 | 1.00 | | | |

*Note*: The table presents posterior cluster membership probabilities for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. All regressors are allowed to have different marginal effects across clusters. Posterior cluster membership probabilities are given by (2.5) evaluated at the ML estimates.

Table 2.A.3.2: Posterior cluster membership probabilities for the model in Table 2.5

| Cluster 1 Countries | | | | | |
|---|---|---|---|---|---|
| Country | $p_{i1}^*$ | $p_{i2}^*$ | Country | $p_{i1}^*$ | $p_{i2}^*$ |
| Algeria | 0.96 | 0.04 | Lesotho | 0.94 | 0.06 |
| Barbados | 1.00 | 0.00 | Mauritius | 1.00 | 0.00 |
| Bolivia | 0.64 | 0.36 | Nepal | 0.64 | 0.36 |
| Botswana | 0.96 | 0.04 | Nicaragua | 1.00 | 0.00 |
| Brazil | 1.00 | 0.00 | Papua New Guinea | 0.61 | 0.39 |
| Cameroon | 0.83 | 0.17 | Peru | 0.98 | 0.02 |
| Colombia | 1.00 | 0.00 | Rwanda | 0.97 | 0.03 |
| Costa Rica | 0.69 | 0.31 | Senegal | 1.00 | 0.00 |
| Ecuador | 0.71 | 0.29 | Sri Lanka | 0.98 | 0.02 |
| Egypt | 1.00 | 0.00 | Swaziland | 1.00 | 0.00 |
| El Salvador | 1.00 | 0.00 | Syria | 1.00 | 0.00 |
| Guatemala | 1.00 | 0.00 | Togo | 0.95 | 0.05 |
| Iran | 1.00 | 0.00 | Trinidad &Tobago | 0.98 | 0.02 |
| Jamaica | 0.99 | 0.01 | Venezuela | 0.97 | 0.03 |
| Japan | 1.00 | 0.00 | Zimbabwe | 0.94 | 0.06 |
| Korea Republic of | 0.99 | 0.01 | | | |
| **Cluster 2 Countries** | | | | | |
| Country | $p_{i1}^*$ | $p_{i2}^*$ | Country | $p_{i1}^*$ | $p_{i2}^*$ |
| Argentina | 0.00 | 1.00 | Malawi | 0.00 | 1.00 |
| Central African Republic | 0.00 | 1.00 | Malaysia | 0.07 | 0.93 |
| Chile | 0.05 | 0.95 | Mexico | 0.00 | 1.00 |
| Dominican Republic | 0.02 | 0.98 | Niger | 0.00 | 1.00 |
| Gambia, The | 0.49 | 0.51 | Pakistan | 0.01 | 0.99 |
| Ghana | 0.06 | 0.94 | Paraguay | 0.00 | 1.00 |
| Haiti | 0.43 | 0.57 | Philippines | 0.00 | 1.00 |
| Honduras | 0.01 | 0.99 | Sierra Leone | 0.00 | 1.00 |
| India | 0.00 | 1.00 | South Africa | 0.00 | 1.00 |
| Indonesia | 0.00 | 1.00 | Sudan | 0.00 | 1.00 |
| Israel | 0.00 | 1.00 | Thailand | 0.00 | 1.00 |
| Jordan | 0.00 | 1.00 | Turkey | 0.00 | 1.00 |
| Kenya | 0.00 | 1.00 | Uganda | 0.00 | 1.00 |
| Liberia | 0.42 | 0.58 | Uruguay | 0.45 | 0.55 |

*Note*: The table presents posterior cluster membership probabilities for the finite mixture model (2.1) with $J = 2$ clusters, estimated for annual real GDP per capita growth rates for 59 countries over the period 1971-2000. Secondary school enrollment, population growth rate and inflation are assumed to have the same marginal effects across clusters, while the rest of the explanatory variables have different marginal effects across clusters. Posterior cluster membership probabilities are given by (2.5) evaluated at the ML estimates.

# Chapter 3

# Financial Development and Convergence Clubs

*Chapter 3 is based on Baştürk, Paap, and Van Dijk (2010).*

## 3.1 Introduction

Starting with neoclassical growth theory, it has been argued that the real per capita incomes of (subsets of) countries should converge in the long-run. Early empirical work studying the existence of income convergence, such as Barro (1991) and Mankiw and Romer (1992), investigates this issue using pure cross-sectional analysis. These studies examine the presence of convergence through the effect of initial income on the average real GDP per capita growth rates of countries over some long time period. Controlling for other variables, a negative coefficient on initial GDP implies that (relatively) poorer countries at the beginning of the sample period grow faster than richer countries. In return, such findings provide evidence for *conditional catch-up* given the rest of the covariates.

It is well documented that the pure cross-sectional analysis of convergence has severe pitfalls. The criticisms can be summarized by two main points. First, analyzing average GDP levels or growth rates over a long time period brings potential misspecification problems as this kind of data cannot uncover the time series dynamics of the GDP process (see e.g. Quah (1993) and Bernard and Durlauf (1995)). Second, these studies define convergence through the growth rates in GDP series. However, the main question in convergence is whether the poor countries will eventually catch up with the rich in real GDP per capita levels. The main focus should therefore be the gaps between GDP levels, instead of GDP *growth rates*, as convergence of the former is an implication of the catch-

up process (see Quah (1993); Friedman (1992); Evans (1998) and Bianchi and Menegatti (2007) among many others).

Following these criticisms, several studies adopt a (panel) time series approach to analyze convergence. In general, these studies focus on the GDP per capita gaps between countries over time using cointegration techniques to analyze whether income disparities between countries are persistent (see e.g. Evans (1998) and Pesaran (2007)). As opposed to the earlier cross-sectional studies, time-series tests for convergence tend to find no support for convergence among all countries.

It is argued that the findings of the time-series studies stem from the properties of the world income distribution. The cross-country distribution of real GDP per capita is far from a unimodal distribution, hence assuming a single long-run GDP level for all countries is unrealistic. Some empirical studies find that sub-groups of countries show similar GDP patterns in the long-run, but this result cannot be generalized to all countries (see Ben-David (1994); Quah (1996) and Durlauf and Johnson (1995b)). This evidence supports theories of *convergence clubs* (Baumol, 1986; Galor, 1996) in the sense that there is no global convergence, but rather groups of countries following similar GDP patterns. Furthermore, the time series in GDP levels show different forms of transitional behavior for countries. While some countries or economic regions are found to have similar GDP structures over time, other countries or regions show diverging GDP levels for certain periods of time, and catch-up in other time periods (Phillips and Sul, 2009).

More recent studies exogenously group countries, for instance based on regions, and test for convergence within these groups. Despite accounting for the time-series dynamics of the growth process, these studies cannot accommodate the possibility of changes in the composition of the convergence clubs over time, but can only test for the existence of convergence within the specified group of countries. For this reason, clustering based on mixture distributions and MCMC methods has recently drawn a lot of attention in the economic growth literature.

One of the main reasons for the popularity of mixture distributions is the possibility to analyze the distribution of countries over the poor and rich groups as well as the composition of the poor/rich groups over time. Paap and Van Dijk (1998); Bloom *et al.* (2003); Canova (2004); Paap *et al.* (2005) and Baştürk *et al.* (2008) use mixture distributions for this purpose. To our knowledge, neither the individual countries' movements between convergence clubs, nor the possible factors affecting these changes are analyzed in the existing studies.

The purpose of this chapter is hence to model the convergence process for a large set of countries in terms of GDP levels, accounting for changing intra-distributional dynamics.

Specifically, we address three questions regarding the unconditional convergence process[1]. The first question we address is whether there are sub-groups of countries that follow the same long-run path in real GDP per capita levels. With respect to this question, we estimate models with different specifications, and consider the number of distinct groups within the data. Furthermore, instead of using exogenous factors, such as regions, to define club memberships, we determine the club memberships endogenously, using real GDP data only.

Our second question concerns the composition of these clubs. Most of the convergence literature does not deal with possible changes in the composition of these clubs. Our methodology explicitly allows countries to switch between GDP clubs over time. Hence, the composition of the clubs is not fixed over time. Furthermore, we analyze whether macroeconomic and financial variables that affect the probability of switching to a different cluster can be identified.

Our third question is whether the composition of these GDP clubs can be explained by initial conditions and financial development[2]. Empirical studies on the effect of financial development on growth are document contradicting results, see Loayza and Ranciere (2006). Both the theoretical and empirical literature on the link between financial development and convergence clubs suggest a positive long-run effect of financial development on growth coexisting with a general negative effect in the short-run (Levine, 2004; Beck, 2008).

The empirical studies analyzing the effect of financial development on growth focus either in the long-run or in the short-run. One exception is Loayza and Ranciere (2006) accounting for both the short-run and long-run effects of financial development on real GDP per capita growth rates. They propose an error-correction model where there is a single long-run relationship between financial development indicators and growth. We follow their idea of incorporating possible short-run and long-run effects of financial development in the economic growth model. However, we do not assume a cointegrating relationship between financial development and growth. Instead, we define financial development indicators as factors that possibly affect the GDP club of a country in the long-run.

[1]Note that there is a distinction between the *conditional* and *unconditional* convergence in the literature. In this study, we adopt the exact definition of convergence, in the sense that long-run income levels converge within the endogenously determined clubs.

[2]Our model is general enough to specify different variables that affect the GDP levels in the short-run or in the long-run. However, following the discussion on the effects of financial development on growth, we especially focus on the conventional measures of financial development, namely measures of financial intermediary development and stock market development, on the formation of convergence clubs.

For the GDP club analysis, we propose a novel Markov Chain State Space Model that endogenously defines the groups of countries that show similar GDP structures. We model the common paths for the countries' real GDP levels and growth rates. We do not make stationarity assumptions for club memberships, but rather allow countries to switch between clubs over time. Our methodology provides a general analysis of convergence clubs. We do not specify an a priori group of countries that follows similar GDP structures, but rather extract the GDP behavior from the data.

In order to check whether financial development affects these club memberships in the short-run and the long-run, we extend the Markov switching specification allowing for covariates affecting the transition between GDP clubs, as well as defining covariates affecting the short-run fluctuations in GDP levels. The key point in this second model is the distinction between the short-run and long-run effects of financial development. We distinguish the short-run effects as factors affecting fluctuations around the common cluster levels. In the long-run however, financial development and the initial conditions are anticipated to explain the composition of the GDP clusters.

The models we propose are related to a wide range of studies focusing on convergence clubs, and studies employing methods for clustering the data in general. The long-run club formation we analyze is similar to the cointegration based methods, (Bernard and Durlauf, 1995; Pesaran, 2007). Our model generalizes these methods in the following way: we do not assume a single long-run relationship for the GDP series across countries, but rather allow for more than one convergence club for the included countries. Furthermore, we allow changes in the cluster memberships over time, indicating possible changes in the long-run GDP correlations of the included countries.

Specifically, a GDP club can have no member countries during a part of the sample period, indicating a merger between this club and one of the other clubs, depending on the changes on the club membership over time. Alternatively, the latent long-run GDP levels for some clusters can converge over time. Identifying the memberships for these clusters is harder, but the interpretation of the catching-up process holds: countries belonging to converging clusters have similar long-run GDP patterns. Note that the methodology we propose is more general compared to the beta and sigma convergence definitions in the literature, which analyze the existence of decreasing long-run trends in GDP and decreasing short-run fluctuations around the common long-run path, respectively.

Our methodology for clustering the data is related to studies proposing endogenous clustering techniques in order to cluster the GDP per capita of countries. In these specifications, one does not have to specify certain covariates defining the groups of countries. The classification rather depends on the data only. For example Paap and Van Dijk

(1998); Paap *et al.* (2005); Baştürk *et al.* (2008) use Markov Chain models and finite mixture models for this purpose. The models we propose are extensions of their work by clustering GDP levels directly, instead of GDP growth rates, as well as allowing for certain covariates to affect the changes in the club memberships and short-run GDP fluctuations.

In terms of the methodology, Frühwirth-Schnatter and Kaufmann (2008) and Hamilton and Owyang (2009) are the two papers closest to this paper. Both studies propose Markov Chain models to assess the subgroups of the data that show similar characteristics. They further allow for covariates to affect the group memberships. Our model builds on these models by modeling the time-dependent data characteristics. Incorporating the state space structure in the Markov Chain model, we estimate the common paths within the groups of data, while allowing for changes in the club compositions over time.

In order to estimate the model we pursue a Bayesian approach and use Gibbs Sampling (Geman and Geman, 1984a). We find that the club memberships are quite persistent, but still group compositions change substantially over time. In particular, several EU member countries and East Asian countries are found to belong to a higher GDP club in recent times compared to the beginning of the 1970s. Regarding the effects of financial development indicators on the GDP process, our results confirm the theoretical basis for different effects of financial development indicators in the short-run and the long-run. In the long-run, financial development is found to affect the countries' GDP level positively. In the short-run however, we find that the effect of financial intermediary development on GDP levels are in general negative.

The remainder of this paper is as follows: Section 5.2 introduces the models for GDP club formation. Section 3.3 presents the Bayesian estimation method and the Gibbs Sampling scheme. Section 3.4 applies the proposed models on the real GDP per capita data. Section 5.5 concludes.

## 3.2 Model Specification

In this section we propose our modeling approach which aims to identify clusters of countries sharing a common growth path in their GDP per capita. Note that in the economic growth context, the term 'convergence clubs' is used for groups of countries with a common growth path. In the mixture models literature, the term 'cluster' is more common. In this paper we use the latter terminology in describing the model specification.

Let $y_{i,t}$ denote the log real GDP per capita of country $i = 1, \ldots, N$ at time $t = 1, \ldots, T$. We assume that there are $J$ clusters of countries. Within each cluster countries have the same long-run growth path, while the growth paths may differ across clusters. We specify

the stochastic growth path of the $j$th cluster as a random walk[3] with a constant cluster-specific drift $\beta_j$ and variance $\sigma^2_{\nu,j}$:

$$\mu_{j,t} = \mu_{j,t-1} + \beta_j + \nu_{j,t}, \tag{3.1}$$

where $\nu_{j,t} \sim NID\left(0, \sigma^2_{\nu,j}\right)$ for $j = 1, \ldots, J$.

An important feature of our model is that the composition of the clusters need not be constant over time. Put differently, countries may switch between different clusters and hence may end up on a different long-run growth path. The latent variable $S_{i,t} \in \{1, \ldots, J\}$ is used to indicate which of the $J$ clusters country $i$ belongs to at time $t$. The real GDP per capita of country $i$ at time $t$ is described by

$$y_{i,t} = \left(\sum_{j=1}^{J} I[S_{i,t} = j]\mu_{j,t}\right) + x^s_{i,t}\psi^s + \varepsilon_{i,t}, \tag{3.2}$$

where $\varepsilon_{i,t} \sim NID\left(0, \sigma^2_i\right)$ and $I[.]$ is an indicator function taking the value of 1 if the argument is true, and 0 otherwise. The explanatory variables in $x^s_{i,t}$ together with the parameters $\psi^s$ describe the country-specific short-run fluctuations of the individual GDP series around the common long-run growth path. $x^l_{i,t}$ in (3.2) may, for instance, include the degree of financial development or other macroeconomic and financial variables.

To complete the model, we have to specify the properties of the cluster membership, as represented by the latent variable $S_{i,t}$. In this paper we propose two different specifications. In the first specification we consider a first-order discrete Markov process for $S_{i,t}$ (see Hamilton (1994, Ch.22)). Let $p_{kj}$ denote the probability that a country belonging to cluster $k$ in period $t-1$ belongs to cluster $j$ in period $t$, that is, $p_{kj} = \Pr\left(S_{i,t} = j \mid S_{i,t-1} = k\right)$ for $j, k \in \{1, \ldots, J\}$. In case of $J$ clusters, these transition probabilities are collected in the matrix $P$,

$$P = \begin{bmatrix} p_{11} & & p_{1J} \\ & \ddots & \\ p_{J1} & & p_{JJ} \end{bmatrix}, \tag{3.3}$$

for all $i, t$. By definition, the Markov switching probabilities are such that $p_{kj} \in [0, 1]$ for all $k, j$ and $\sum_j p_{kj} = 1$ (see Hamilton (1994, p. 262)).

The first-order Markov specification for $S_{i,t}$ has the advantage that the clusters of countries sharing the same growth path are determined completely endogenously, only based on the GDP data, without any a priori grouping based on characteristics such as

[3]Note that modeling annual log real GDP per capita series as a random walk process is the standard approach in the literature as most cross-sectional series for the annual log real GDP per capita are found to have one unit root only.

geographic location. On the other hand, it may be restrictive as transition probabilities are the same for all countries and constant over time. In the second specification we therefore relax both assumptions by relating cluster membership probabilities to certain explanatory variables. If we assume a clear ordering in the growth paths with respect to the explanatory variables, we can use an ordered probit model to describe segment membership:

$$\begin{aligned} S_{i,t} &= j \text{ iff } \left(\gamma_{j-1} < s^*_{i,t} \leq \gamma_j\right) \text{ for } j = 1, \ldots, J \\ s^*_{i,t} &= x^l_{i,t}\psi^l + \zeta_{i,t}, \end{aligned} \tag{3.4}$$

where $\gamma_j$ for $j = 0, \ldots, J$ are threshold parameters with $\gamma_0 = -\infty$ and $\gamma_J = \infty$, and where $\zeta_{i,t} \sim \text{NID}(0, 1)$ with the variance of $\zeta_{i,t}$ fixed at 1 for identification. The parameter vector $\psi^l$ describes the effect of the $x^l_{i,t}$ variable on the cluster membership and hence the long-run level and the growth rate for each observation. Positive values of these coefficients imply that the probability of a country to belong to a higher GDP cluster increases with the covariates. In our empirical analysis below, we include initial conditions and financial development indicators in the $x^l_{i,t}$ vector.

It is useful to note that there are several ways to relate the club membership to explanatory variables. Our choice for an ordered probit specification is motivated by the properties of the GDP data. First, a Markov process with time varying transition probabilities might seem a natural extension of the first specification above. Such a specification however requires a rather large number of observations to accurately identify and estimate the model parameters. Although we consider a large cross-section of countries, the number of time periods in our sample is quite restricted. Second, the ordered probit model provides a natural ranking in the GDP clubs. Other methods, such as a multinomial logit model for cluster memberships, does not provide such a ranking for the cluster levels, and solving the label switching problem in these models can be quite cumbersome (Frühwirth-Schnatter, 2006; Geweke, 2007)[4].

## 3.3 Estimation and Inference

We opt for a Bayesian approach to do inference in our proposed model. Specifically, posterior results for the model parameters and related statistics are obtained using Gibbs Sampling (Geman and Geman, 1984a) together with data augmentation (Tanner and Wong, 1987).

[4]The label switching problem may occur since the model we propose does not define a ranking of groups such as clusters with low/high growth rates.

To implement the Gibbs sampler, we first consider the complete data likelihood function of our model, which is given by

$$f(\mathcal{S};\vartheta)\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{t=1}^{T}\phi(\mu_{j,t};\mu_{j,t-1}+\beta_j,\sigma^2_{\nu,j})^{I[S_{i,t}=j]} \\ \phi(y_{i,t};\textstyle\sum_{j=1}^{J} I[S_{i,t}=j]\mu_{j,t}+x^s_{i,t}\psi^s,\sigma^2_i), \quad (3.5)$$

where $\phi(\cdot;\mu,\sigma^2)$ is the density function of a normal distribution with mean $\mu$ and variance $\sigma^2$ and $f(\mathcal{S};\vartheta)$ denotes the likelihood contribution of the model for the cluster membership variables summarized in $\mathcal{S}$ with parameter vector $\vartheta$. For the Markov switching specification as given in (3.3) we have

$$f(S,\vartheta)\propto\prod_{j=1}^{J}\prod_{k=1}^{J}p_{kj}^{\mathcal{N}_{kj}}, \quad (3.6)$$

where $\mathcal{N}_{kj}$ denotes the number of transitions from cluster $k$ to $j$ and $\vartheta = P$. For the ordered probit specification (3.4) we have

$$f(S;\vartheta)\propto\prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{j=1}^{J}\left(I[\gamma_{j-1}<s^*_{i,t}<\gamma_j]\phi(s^*_{i,t};x^l_{i,t}\psi^l,1)\right)^{I[S_{i,t}=j]} \quad (3.7)$$

with $\vartheta=\{\gamma_1,\ldots,\gamma_{J-1},\psi^{(l)}\}$.

Regarding the priors, we consider conjugate or flat priors. More precisely, we use inverted Gamma priors for the variance parameters $\{\sigma^2_i\}_{i=1}^N$ and $\{\sigma^2_{\nu,j}\}_{j=1}^J$, flat priors for the parameters $\psi^l$, $\psi^s$, the cluster-specific drifts $\{\beta_j\}_{j=1}^J$ and the initial conditions $\{\mu_{0,j}\}_{j=1}^J$, Dirichlet priors for the transition probabilities in $P$ and flat priors for the ordered probit parameters $\gamma$, taking into account the ordering in the parameters (i.e. $\gamma_k\leq\gamma_s$ for $k<s$).

For both specifications of $S_{i,t}$ our model is basically a (Markov) mixture State Space model. To obtain posterior results we can use the results of Carter and Kohn (1994) together with Frühwirth-Schnatter and Kaufmann (2008), except for updating the latent state variables (i.e. the cluster-specific growth paths $\mu_{j,t}$) and the cluster probabilities for the ordered probit model specification. To save notation we summarize the data by $\mathcal{Y}=\{\{y_{it}\}_{t=1}^T\}_{i=1}^N$. The sampling scheme is given by

(a) Given the cluster memberships and common growth paths, draw the model parameters:

- Draw $\{\sigma^2_{\nu,j}\}_{j=1}^J$ from $p(\sigma^2_{\nu,j}\mid\mathcal{Y},\mathcal{S},\{\mu_{j,t}\}_{t=0}^T,\psi^s,\vartheta,\{\sigma^2_i\}_{i=1}^N,\{\beta_j\}_{j=1}^J)$ for $j=1,\ldots,J$ which are inverted Gamma distribution

- Draw $\{\sigma_i^2\}_{i=1}^N$ from $p(\sigma_i^2 \mid \mathcal{Y}, \mathcal{S}, \{\mu_{j,t}\}_{t=0}^T, \psi^s, \vartheta, \{\sigma_{\nu,j}^2\}_{j=1}^J, \{\beta_j\}_{j=1}^J)$ for $i = 1, \ldots, N$ which are inverted Gamma distributions.
- Draw $\psi^s$ from $p(\psi^s \mid \mathcal{Y}, \mathcal{S}, \{\mu_{j,t}\}_{t=0}^T, \vartheta, \{\sigma_{\nu,j}^2\}_{j=1}^J, \{\sigma_i^2\}_{i=1}^N, \{\beta_j\}_{j=1}^J)$ which is a normal distribution.

(b) Given the model parameters and cluster memberships, draw the latent state variables according to the underlying model:

- $\{\mu_{j,t}\}_{t=0}^T, \beta_j$ from $p(\{\mu_{j,t}\}_{t=0}^T, \beta_j \mid \mathcal{Y}, \mathcal{S}, \psi^s, \vartheta, \{\sigma_{\nu,j}^2\}_{j=1}^J, \{\sigma_i^2\}_{i=1}^N)$ for $j = 1, \ldots, J$ using the simulation smoother of Carter and Kohn (1994).

(c) Given the model parameters and the latent state variables, draw the cluster membership variables $\mathcal{S}$:

- Draw $\mathcal{S}$ from $p(\mathcal{S} \mid \mathcal{Y}, \{\mu_{j,t}\}_{t=0}^T, \psi^l, \vartheta, \{\sigma_{\nu,j}^2\}_{j=1}^J, \{\sigma_i^2\}_{i=1}^N, \{\beta_j\}_{j=1}^J)$. In case of a first-order Markov process model for the cluster memberships we can use the simulation smoother of (Albert and Chib, 1993a). For the ordered probit specification we can simply simulate the individual elements of $\mathcal{S}$ from multinomial distributions.

(d) The final steps of the Gibbs Sampling scheme concerns the simulation of the $\vartheta$ parameters.

- Draw $\vartheta$ from $p(\vartheta \mid \mathcal{Y}, \mathcal{S}, \{\{\mu_{j,t}\}_{t=0}^T\}_{j=1}^J, \psi^s, \{\sigma_{\nu,j}^2\}_{j=1}^J, \{\sigma_i^2\}_{i=1}^N)$. In case we opt for a first order Markov mixture model, we can draw $P$ from a Dirichlet distribution, see Koop (2003, p. 256)). For the ordered probit model specification on the other hand, we need to draw the $\psi^l$ parameter, the latent variables $\mathcal{S}^* = \{\{s_{i,t}^\star\}_{t=1}^T\}_{i=1}^N$ and the probit thresholds $\gamma_1, \ldots, \gamma_{J-1}$ in (3.4), which can be done following Albert and Chib (1993b) and Koop (2003, p. 218).

The abovementioned Gibbs sampler is conditional on the number of clusters $J$. Following other frequentist and Bayesian studies in this area, we rely on information criteria for the choice of the number of clusters in the data. In the Bayesian context, most studies use the *Bayesian information criteria* (BIC) (Schwarz, 1978). It is however documented that in the missing data models or latent class models, the penalty for model complexity in the BIC criteria might not be satisfactory. Spiegelhalter *et al.* (2002) show this result in particular for hierarchical models, and propose *deviance information criteria* (DIC). For mixture models, Celeux *et al.* (2006) proposes extensions of the DIC. The main difference in interpretation between the BIC and DIC is the penalty for model complexity.

Whereas the former explicitly accounts for the number of effective model parameters, DIC on the other hand considers the dispersion of the log-likelihood draws. Celeux *et al.* (2006) propose 8 different DIC specifications. We use the conditional DIC (*DIC7*) since this specification implicitly accounts for the latent variables as additional parameters. The BIC and DIC criteria are defined as:

$$BIC_J = -2\tilde{l}_J(\theta) + \ln(N \times T) \times \kappa_J \tag{3.8}$$

$$DIC_J = -4\bar{l}_J(\theta) + 2l_J(\bar{\theta}), \tag{3.9}$$

where $\theta = \{\psi^s, \{\sigma^2_{\nu,j}\}_{j=1}^J, \{\sigma_i^2\}_{i=1}^N, \vartheta, \mathcal{S}, \{\mu_{j,t}\}_{t=0,j=1}^{t=T,j=J}, \{\beta_j\}_{j=1}^J\}$ contains the model parameters and the latent variables, $\kappa_J$ is the number of parameters for the model with $J$ clusters, and $l_J(\theta)$ is the complete data (log) likelihood for the model with $J$ clusters evaluated at $\theta$. Furthermore, $l_J(\bar{\theta})$ is the log-likelihood function evaluated in the posterior mean of $\theta$, while $\bar{l}_J(\theta)$ and $\tilde{l}_J(\theta)$ are the mean and the maximum value of the log-likelihood over the draws.

## 3.4 Empirical Results

In this section we apply the convergence club model of Section 5.2 with the Markov switching specification (3.3) and the ordered probit model specification (3.4) to annual log real Gross Domestic Product per capita levels. We apply the model to two different data sets. In the first application we take a large cross-section of 163 countries but we do not consider covariates for the club memberships or short-run fluctuations. The reason for this decision is that our covariates of interest are available for a limited number of countries over a sufficiently long period of time. In the second application we consider a smaller set of 33 countries for which financial development indicators are available, in order to examine whether these are informative for the process of convergence club formation and short-run fluctuations in GDP.

In Section 3.4.1 we discuss the data sets in detail. Empirical results for the first and second applications are given in Section 3.4.2 and Section 3.4.3.

### 3.4.1 Data

The data for annual real GDP per capita are taken from the Penn World Tables (PWT), version 6.3, Heston *et al.* (2009). We consider the natural logarithm GDP, as percentage changes in GDP are more intuitive than the absolute changes.

In the first application we consider a balanced panel of 163 countries for the period 1970–2007. This data set is used to assess convergence club formation and changes in

the composition of convergence clubs over time endogenously, without specifying certain covariates that affect the cluster membership probabilities or short-run GDP fluctuations. The countries in this data set are listed in Table 3.1.

Table 3.1: Markov switching specification (3.3): Included countries

| |
|---|
| Afghanistan, Albania, Algeria, Angola, Antigua and Barbuda, Argentina, Australia, Austria, Bahamas, Bangladesh, Barbados, Belgium, Belize, Benin, Bermuda, Bhutan, Bolivia, Botswana, Brazil, Brunei Darussalam, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Central African Republic, Chad, Chile, China, Channel Islands, Colombia, Comoros, Congo Dem. Rep., Congo Rep., Costa Rica, Côte d'Ivoire, Cuba, Cyprus, Denmark, Djibouti, Dominica, Dominican Republic, Ecuador, Egypt Arab Rep., El Salvador, Equatorial Guinea, Ethiopia, Fiji, Finland, France, Gabon, The Gambia, Germany, Ghana, Greece, Grenada, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hong Kong (China), Hungary, Iceland, India, Indonesia, Iran Islamic Rep., Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kenya, Kiribati, Korea Rep., Kuwait, Lao PDR, Lebanon, Lesotho, Liberia, Libya, Luxembourg, Macao (China), Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Marshall Islands, Mauritania, Mauritius, Mexico, Micronesia Fed. Sts., Mongolia, Morocco, Mozambique, Namibia, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, Norway, Oman, Pakistan, Palau, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Romania, Rwanda, Samoa, São Tomé and Principe, Saudi Arabia, Senegal, Seychelles, Sierra Leone, Singapore, Solomon Islands, Somalia, South Africa, Spain, Sri Lanka, St. Kitts and Nevis, St. Lucia, St. Vincent and the Grenadines, Sudan, Suriname, Swaziland, Sweden, Switzerland, Syrian Arab Republic, Taiwan (China), Tanzania, Thailand, Togo, Tonga, Trinidad and Tobago, Tunisia, Turkey, Uganda, United Arab Emirates, United Kingdom, United States, Uruguay, Vanuatu, Venezuela RB, Vietnam, Zambia, Zimbabwe. |

*Note*: The data consists of 163 countries for the period 1970–2007.

For the second application the data consists of annual real GDP per capita levels as well as financial development indicators. For this dataset, financial development indicators are taken as covariates possibly determining convergence club formation, and short-run GDP fluctuations of the individual series around the club-level.

The financial development indicators we consider can be classified into two main categories. The first category is labeled financial intermediary development, measured by Deposit Money Bank Assets as a percentage of GDP (*Bank assets*) and Commercial bank assets as a percentage of total assets (*Commercial/Central bank*). The second category is

labeled stock market development, measured by stock market turnover (*turnover*). For more detailed descriptions of these financial development indicators, see Beck and Levine (2004); Rioja and Valev (2004); Aghion *et al.* (2005); Loayza and Ranciere (2006); Méon and Weill (2008). The data for financial development are taken from theBeck *et al.* (2000) database of financial development indicators, revised in May 2009.

The limited availability of the financial development indicators reduces both the time-series and the cross-sectional dimension of the data. For a balanced dataset including financial development indicators, our sample comprises 33 countries for the period 1989–2006. The countries included in this data set are listed in Table 3.2.

Table 3.2: Ordered probit model extension (3.4): Included countries

| |
|---|
| Argentina, Australia, Canada, Côte d'Ivoire, Chile, Denmark, Egypt, Finland, Greece, India, Indonesia, Israel, Italy, Jamaica, Japan, Jordan, Korea Rep., Malaysia, Morocco, New Zealand, Nigeria, Pakistan, Philippines, Portugal, Spain, Sri Lanka, Thailand, Trinidad and Tobago, Tunisia, Turkey, United Kingdom, United States. |

*Note*: The data consists of 33 countries for the period 1989–2006.

As discussed in the introduction, financial development may have both short-run and long-run effects on GDP, albeit in opposite directions. In order to improve the identification of these short- and long-run effects, we consider different transformations of the financial development indicators. Specifically, we include lagged 5-year moving averages of the financial development indicators in the vector $x_t^l$, affecting the cluster probabilities in the ordered probit specification in (3.4). For the short-run variables, on the other hand, we use the one-year lagged financial development indicator in deviation from its 5-year moving average. Finally, in order to control for the effect of initial conditions on club memberships, we include the starting level of GDP per capita as an additional variable in $x_t^l$.

Note that the sample size substantially decreases when we include the financial development indicators, due to missing values in these variables. We could keep the sample size larger if we adopted a method to handle missing data, such as the Expectation Maximization algorithm. These methods however assume that the data are missing at random, which may not be appropriate for the financial development indicators, which have missing values for consecutive time periods rather than at random time points. Dealing with missing data in this case means that we have to assume a statistical model for the financial indictor variables. Due to the large number of missing observations, the results may depend strongly on the proposed method for dealing with missing data. Therefore, we do

not pursue this approach and we choose to limit the sample size according to data availability. Both the cross-sectional and the time-series dimensions of the data are important for parameter estimation. Extracting the growth paths relies on the time series dimension of the data. On the other hand, the cross-sectional dimension of the data is necessary for a comparative analysis of the convergence clubs. For this reason, the choice for the number of countries and the time period in the above datasets is based on achieving a reasonable number of observations in both dimensions.

### 3.4.2 Endogenous formation of convergence clubs

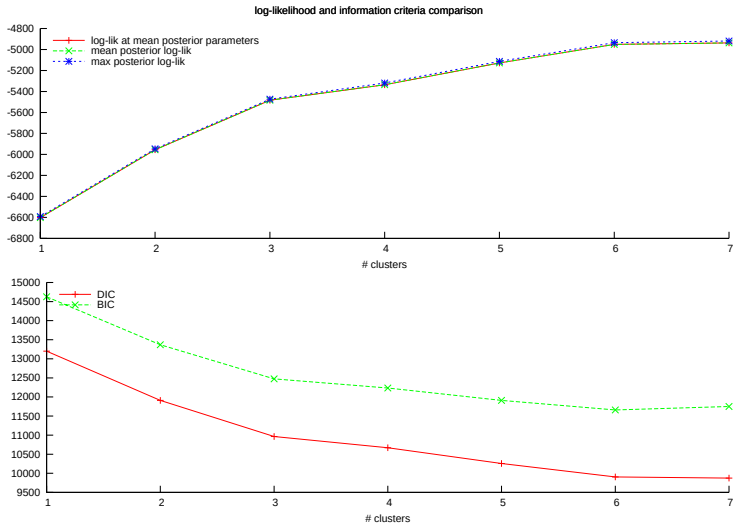
We first apply the model in (3.2) to annual log real GDP per capita for 163 countries for the period 1970-2007. The purpose of this analysis is to analyze the formation of convergence clubs and the dynamics of their composition endogenously. For this reason we use the Markov switching specification in (3.3) for the club membership variable $S_{i,t}$, hence not specifying factors that may affect the club membership probabilities. Furthermore, we do not consider the effects of covariates for the short-run fluctuations in GDP, that is, we impose $\psi^s = 0$ in (3.2).

The proper priors for error variances and transition probabilities are defined as follows: Error variances for the latent GDP levels in each club $\{\sigma^2_{\nu,j}\}_{j=1}^J$, and country-specific error variances $\{\sigma_i^2\}_{i=1}^N$ have inverted Gamma priors. For the former, we divide the GDP series into $J$ groups, where observations belonging to each group is determined by $J$ quantiles of the GDP distribution in each period. For each group, we then set the mean of the inverted Gamma density equal to the estimated variance of within-group growth rates. The scale parameter is fixed at 5 to allow for a relatively large prior variance on the error variances. For $\{\sigma_i^2\}_{i=1}^N$, we use an inverted Gamma density with mean equal to the estimated variances of GDP within each country, and scale parameter set equal to 50. The Dirichlet priors for the transition probabilities $P$ are defined as $\text{Dir}(x)$, where $x = 2 \times \imath$ and $\imath$ is the $1 \times J$ vector of ones [5].

An important aspect of the analysis is to determine the number of convergence clubs in the data. The common practice in previous literature is to employ two to four GDP clubs (see e.g. Hansen (2000); Canova (2004); Paap *et al.* (2005)). Here we estimate models with two to seven clubs, as well as a single club model (which implies a pooled regression for all included countries) and use the information criteria (BIC and DIC) summarized in Section 3.3 to determine the appropriate number of convergence clubs.

[5]The results are insensitive to small changes in the prior specifications. Furthermore, the posterior results for the variance terms are quite different from the prior means.

Figure 3.1: Markov switching specification (3.3): Posterior log-likelihood summary and Information Criteria Comparisons




*Note*: The figure shows posterior results for the first-order Markov switching specifications with 1 to 7 clubs, for the dataset with 163 countries. The top figure shows the posterior conditional log likelihood summaries: log-likelihood values at mean posterior parameters, mean posterior log-likelihood values, and maximum posterior log-likelihood values. The bottom graph shows BIC and DIC values.

The BIC and DIC comparisons for the different models are given in Figure 3.1, together with posterior log-likelihood summaries. Notice that posterior log-likelihood values at mean posterior parameters, mean posterior log-likelihood values, and maximum posterior log-likelihood values are almost the same in all cases. Both information criteria decrease with the number of clubs. This stems from the fact that the model we propose has a relatively small number of parameters compared to the number of unobserved variables, i.e. the long-run growth paths $\{\mu_{j,t}\}_{t=0}^{T}$ for $j = 1, \ldots, J$ and the club membership variable $S_{i,t}$. These latent variables, which bring additional uncertainty to the estimates, are not explicitly accounted for in both information criteria. In the literature, it is shown that DIC performs relatively better in mixture models, as it accounts for the model complexity through the deviance in the posterior log-likelihood values. For the mixture model with latent variables we employ here, we cannot confirm this finding [6].

[6]Other criteria for model comparison, such as AIC, AIC3, and CAIC using posterior mode values of the log-likelihood lead to similar results. In particular, AIC3 and CAIC indicate the presence of five convergence clubs. However, the criteria values for the models with three and four clubs are quite close. Furthermore, estimating a model with five clubs leads to three clubs with very similar GDP growth paths.

Comparing BIC and DIC values, we confirm that the DIC penalty for model complexity, which depends on the posterior likelihood dispersion, is relatively higher. Despite this result, both information criteria do not provide sufficient penalty for model complexity for this model. Therefore we consider the differences in information criteria and posterior log-likelihood values for different numbers of clubs.

Figure 3.1 shows that the declines in the value of the information criteria become much smaller when more than three clubs are included in the model. The same result holds for the posterior log-likelihood summaries reported in Figure 3.1. We therefore choose the model with three clubs as our base model.

Table 3.3 shows posterior means for the transition probabilities in the model with three convergence clubs. The diagonal elements of the transition matrix are quite close to unity, indicating strong persistence in club membership. This is not a surprising result in the sense that shifts in long-run GDP growth paths by definition are not expected to occur frequently. Furthermore, there are no sudden shifts of countries from the low GDP club to the high GDP club, or vice versa, as the transition probabilities between these clubs are almost equal to zero.

Table 3.3: Posterior mean of the transition probability matrix

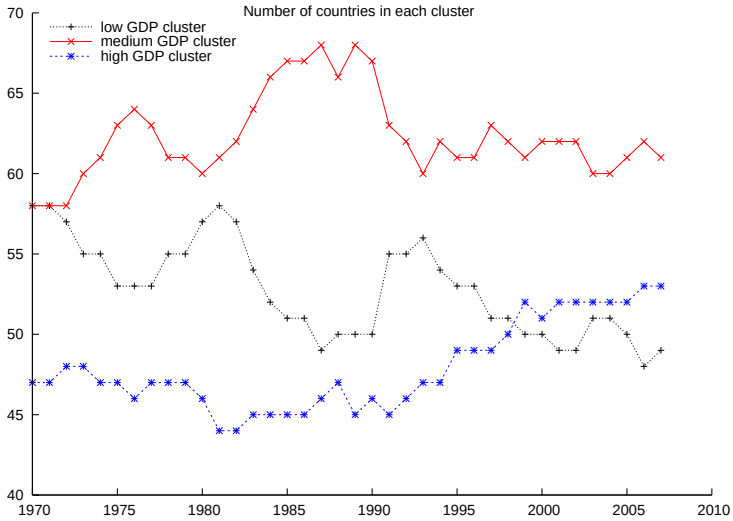
| | Low GDP | Medium GDP | High GDP |
|---|---|---|---|
| Low GDP | 0.98 | 0.01 | 0.00 |
| Medium GDP | 0.02 | 0.98 | 0.01 |
| High GDP | 0.00 | 0.01 | 0.99 |

*Note*: The table presents posterior mean of $P$ in (3.3) for the dataset with 163 countries (1970–2007) without covariates. Clubs *high GDP*, *medium GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period.

We emphasize, however, that the observed persistence in the transition probability matrix does not imply that countries never change from one convergence club to another. Instead, the near-identity transition probability matrix implies occasional changes in convergence club memberships. For example, it may be the case that a small number of countries that belong to the low GDP club during the beginning of the sample period switch at some point in time to the medium GDP club for the rest of the sample period. Although this results in a low off-diagonal transition probability, it is important to pinpoint this change in the club membership as it implies a structural change for these specific countries. Therefore we provide a more detailed analysis of cluster compositions for this model.

Figure 3.2 shows the number of countries in each club over the sample period, where we assign a country to a club based on the posterior mode value of $S_{i,t}$. We find that the composition of clubs clearly changes over time, despite the persistence of the Markov process. Two patterns stand out from the graph. First, until 1990 the number of countries in the high GDP club remains fairly constant, but a substantial number of countries switch between the low and medium GDP clubs. For example, the number of countries in the medium GDP club ranges between 58 in 1970-2 and 68 in 1987. Second, after 1990 the number of countries in the lowest GDP club steadily declines from 56 to 48, while the number of countries in the high GDP club increases from 45 to 52.

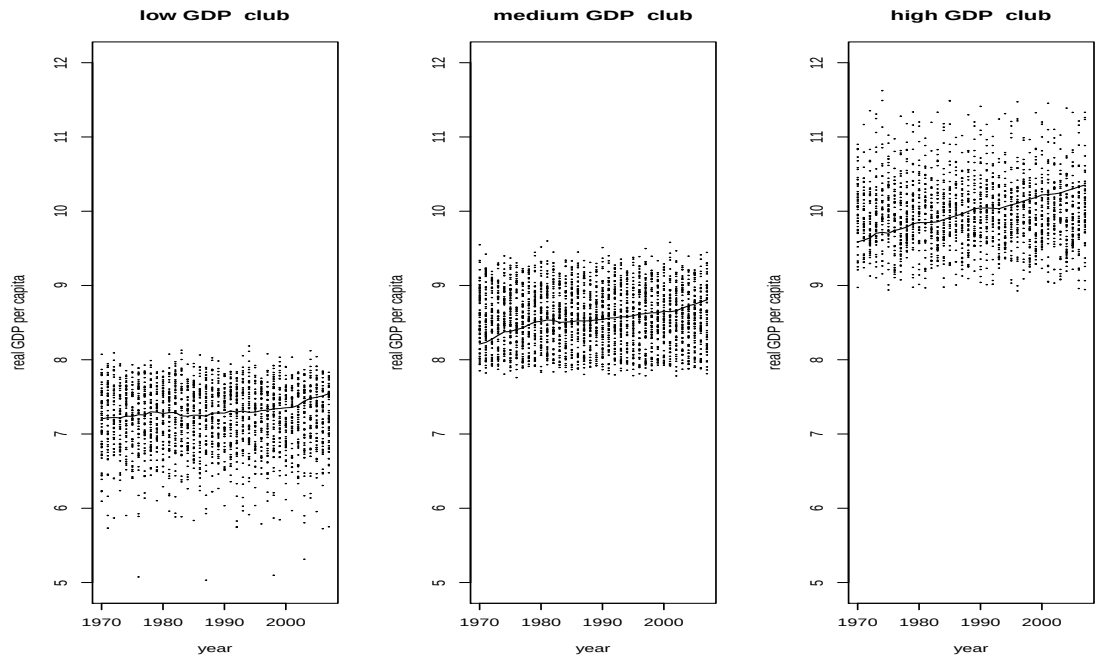
Figure 3.2: Markov switching specification (3.3): Number of countries in each club over time




*Note*: The figure shows the number of countries in each club over the sample period for the three club model, for the dataset with 163 countries. Clubs *high GDP*, *medium GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period. Club membership is determined by the posterior mode of the $S_{i,t}$ variable with a Markov switching specification.

The number of countries in the medium GDP clubs remains approximately constant during the second half of the sample period. Given that direct transitions from the low to high GDP club (almost) do not occur, this implies substantial changes in the composition of the medium GDP club, with some countries entering this group from the low GDP club and other countries exiting to the high GDP club. Also note that this result is in line with the arguments on changing intra-distributional dynamics, noted by Quah (1996), for instance. Quah (1996) argues that the world income distribution is subject to changes

Figure 3.3: Markov switching specification (3.3): Club Memberships




*Note*: The figure shows the mean posterior club levels (lines), and mean posterior club memberships (symbols), for the dataset with 163 countries. Clubs *high GDP*, *medium GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period. Club membership is determined by the posterior mode of the $S_{i,t}$ variable with a Markov switching specification.

with the middle income class disappearing. Our results also indicate changes in the cross-sectional GDP distribution over time, but the changes are rather caused by the decrease in the number of countries belonging to the *low GDP* club rather than a disappearing middle income class.

Club-specific GDP paths and observations belonging to each club are shown in Figure 3.3. It can be seen that GDP levels and trends are quite different across clubs. We show these differences in detail in Table 3.4, reporting the initial GDP levels, growth rates and variances of GDP for each club.

Table 3.4 shows that initial GDP levels are clearly different across clubs, both in terms of the posterior mean and the reported percentiles. Mean annual growth rates for the *high GDP* and *medium GDP* clubs are both around 2%. The *low GDP* club on the other hand has a much lower growth rate. These results do not indicate convergence between clubs. First, the low GDP club starts with a relatively low GDP level, and the growth rate in this club is much lower than the other clubs. Hence we find that the low GDP club rather

diverges from the rest of the convergence clubs. Second, although growth rates in the medium and high GDP club are very close, these clubs do not converge to a common level as a result of the substantial difference in initial conditions. Table 3.4 also shows variations in GDP values in each club. These values are quite similar across samples, hence the division of convergence clubs is not related to the club-specific GDP fluctuations.

We next report club memberships for each country. Table 3.5 reports the countries that stay in the same club over the sample period. Table 3.6 on the other hand reports the countries that change clubs at least once during the sample period. It can be seen that this is the case for around one third of the countries.

According to Tables 3.5 and 3.6, most sub-Saharan African countries in the sample are in the low GDP club throughout the entire sample period. Latin American countries are in general in the medium GDP club. Oil producing countries, are in the high GDP club throughout the sample period, with the exception of Oman, which temporarily switches to the medium GDP club in the year 1973.

Table 3.7 depicts a selection of countries that change clubs during the sample period together with the year in which the change took place. In particular, the model shows that several Asian countries (*Asian tigers*) switched to a higher GDP club compared to their initial club membership. Furthermore, some EU member countries, such as Cyprus, Hungary and Malta switched to the high GDP club.

Table 3.4: Markov switching specification (3.3): Initial GDP levels, growth rates and error variances for each club

| | *Club specification* | | |
|---|---|---|---|
| | low GDP | medium GDP | high GDP |
| GDP level $\mu_{j,t}$ in 1970 | 7.21 | 8.21 | 9.59 |
| | (7.19, 7.22) | (8.20, 8.22) | (9.57, 9.60) |
| Growth rate $\beta_j$ | 0.006 | 0.021 | 0.023 |
| | (0.002, 0.012) | (0.014, 0.040) | (0.020, 0.036) |
| Error variances $\sigma^2_{\nu,j}$ | 0.039 | 0.038 | 0.038 |
| | (0.027, 0.059) | (0.027, 0.058) | (0.025, 0.055) |

*Note*: The table shows posterior means for initial GDP levels ($\mu_{j,1970}$), growth rates ($\beta_j$) and error variances ($\sigma^2_{\nu,j}$) for each club, for the Markov switching specification. 2.5% and 97.5% percentiles of the posterior densities are reported in parentheses. Clubs *high GDP*, *medium GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period.

Table 3.5: Markov switching specification (3.3): Countries that do not change clubs over time

| **Low GDP club:** |
|---|
| Afghanistan, Bangladesh, Benin, Burkina Faso, Burundi, Cambodia, Central African Rep., Chad, Comoros, Congo Dem. Rep., Ethiopia, Gambia, The Ghana, Guinea-Bissau, Haiti, Kenya, Lao PDR, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Nepal, Niger, Nigeria, Pakistan, Papua New Guinea, Rwanda, Senegal, Solomon Islands, Somalia, Sudan, Syria, Tanzania, Togo, Uganda. |
| **Medium GDP club:** |
| Algeria, Belize, Bolivia, Brazil, Bulgaria, Colombia, Costa Rica, Cuba, Dominican Rep., Ecuador, El Salvador, Fiji, Guatemala, Honduras, Jordan, Marshall Islands, Mexico, Morocco, Namibia, Panama, Paraguay, Peru, Philippines, Poland, Romania, Samoa, São Tomé and Principe, South Africa, St. Lucia, Swaziland, Tonga, Tunisia, Turkey, Uruguay, Vanuatu. |
| **High GDP club:** |
| Australia, Austria, Bahamas, Barbados, Belgium, Bermuda, Brunei, Darussalam, Canada, Denmark, Finland, France, Germany, Greece, Hong Kong, Iceland, Ireland, Israel, Italy, Japan, Kuwait, Libya, Luxembourg, Macao, Netherlands, New Zealand, Norway, Palau, Portugal, Puerto Rico, Qatar, Saudi Arabia, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, United States. |

*Note*: The table presents the countries that stay in the same convergence club over the whole sample period (1970-2007). Clubs *high GDP*, *medium GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period. Club membership is determined by the posterior mode of the $S_{i,t}$ variable with a Markov switching specification.

Table 3.6: Markov switching specification (3.3): Countries that change clubs over the sample period

| |
|---|
| Albania, Angola, Antigua and Barbuda, Argentina, Bhutan, Botswana, Cameroon, Cape Verde, Chile, China, Channel Islands, Congo Rep., Côte d'Ivoire, Cyprus, Djibouti, Dominica, Egypt Arab Rep., Equatorial Guinea, Gabon, Grenada, Guinea, Guyana, Hungary, India, Indonesia, Iran Islamic Rep., Iraq, Jamaica, Kiribati, Korea Rep., Lebanon, Malaysia, Maldives, Malta, Mauritius, Micronesia Fed. Sts., Mongolia, Nicaragua, Oman, Seychelles, Sierra Leone, Singapore, Sri Lanka, St. Kitts and Nevis, St. Vincent and the Grenadines, Suriname, Taiwan, China, Thailand, Trinidad and Tobago, Venezuela, Vietnam, Zambia, Zimbabwe. |

*Note*: The table presents the countries that change clubs at least once during the sample period (1970-2007). Club membership is determined by the posterior mode of the $S_{i,t}$ variable with a Markov switching specification.

Table 3.7: Markov switching specification (3.3): Selected countries and the time periods they change clubs

| Previous club \ New club | Low GDP | Medium GDP | High GDP |
|---|---|---|---|
| Low GDP | | Bhutan (2001)<br>Botswana (1976)<br>China (1995)<br>Egypt (1983)<br>India (2006)<br>Indonesia (1987)<br>Thailand (1977)<br>Vietnam (2006) | |
| Medium GDP | Zambia (1976)<br>Zimbabwe (2003) | | Chile (1995)<br>Cyprus (1982)<br>Hungary[a] (1999)<br>Korea, Rep (1990)<br>Malaysia (1995)<br>Malta (1986)<br>Singapore (1982)<br>Taiwan (1987) |
| High GDP | | Iran (1978)<br>Lebanon[a] (1988)<br>Venezuela (1989) | |

*Note*: The table summarizes selected countries' club changes over time, for the Markov switching specification, dataset with 163 countries for the period between 1970–2007. The year of change is indicated in parentheses. Clubs *high GDP*, *medium GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period. Club membership is determined by the posterior mode of the $S_{i,t}$ variable with a Markov switching specification.

[a] denotes the countries that change clubs more than once. For these countries we report the year after which the country stays in the same club.

In sum, we conclude that there are three separate clubs of countries in terms of the common GDP paths. The changing number of countries in each club, together with the differences in mean club levels, suggests that it is interesting to analyze the GDP clubs with a model which explains changes in club memberships over time. In Section 3.4.3 we analyze the effects of financial development indicators on the formation of GDP clubs and the dynamics of their composition.

### 3.4.3 Effects of financial intermediary development and stock market development on convergence clubs

The purpose of this section is to analyze whether initial conditions and financial development indicators have explanatory power for the long-run GDP growth paths (in terms of club membership probabilities) as well as the short-run deviations from these. The dataset concerns a balanced panel of 33 countries for the period 1989-2006, and includes log real GDP levels, financial intermediary development and stock market development indicators as explained in Section 3.4.1. Priors for the error variances and transition probabilities in (3.1) and (3.2) are defined as in Section 3.4.2[7].

We consider the ordered probit specification (3.4), using financial development indicators as long-run and short-run factors affecting GDP club membership probabilities and short-run deviations from the club levels, respectively. We focus on the two club model for a number of reasons. First, most countries included in the smaller dataset belong to either the medium or high GDP club in the analysis in Section 3.4.2. Second, for more than three clubs, we found that at least one of the clubs systematically disappeared, with no observations belonging to the club. Third, the results for three clubs were quite sensitive to the included countries. Finally, we examined the robustness of the results by repeating the estimation 33 times, where one of the countries was removed from the dataset in turn. This analysis shows that the estimation results of the two-club model are quite robust. For the three-club model on the other hand, both GDP levels within clubs and the marginal effects of the covariates vary substantially depending on the composition of the panel. Therefore, in the remainder of this section, we focus on the two-club model.

Table 3.8 shows club membersips for the two club model. Similar to the model employed in Section 3.4.2, club memberships seem to be persistent as most countries do not change clubs over time. Still we find that 15% of the countries change clubs at some point during the sample period. Note that the number of years in this dataset is quite small compared to that of Section 3.4.2, which explains the smaller percentage of countries changing clubs over time[8].

Club-specific GDP paths and the observations belonging to each club are shown in Figure 3.4. GDP levels are quite different across clubs. Specifically, the high GDP club has a relatively high initial GDP level and the club levels do not seem to converge at the

[7]The results are not sensitive to small changes in the prior specifications. Furthermore, the posterior results for the variance terms are quite different from the prior means.

[8]We also applied the Markov switching model to this dataset. Club memberships in that case are quite similar to the results reported here.

end of the sample period. This difference is shown in detail in Table 3.9, where we report initial GDP levels, GDP growth rates and the error variances for each club.

Table 3.9 shows that the mean GDP growth rates are the same across clubs. Together with this finding, the difference in initial GDP levels implies a persisting difference in the

Table 3.8: Ordered probit model extension (3.4): Club memberships with covariates (financial intermediary and stock market development indicators)

| |
|---|
| *Countries that are always in the low GDP club:* |
| Côte d'Ivoire, Egypt, India, Indonesia, Jamaica, Jordan, Morocco, Nigeria, Pakistan, Philippines, Sri Lanka, Thailand, Tunisia, Turkey. |
| *Countries that are always in the high GDP club:* |
| Australia, Canada, Denmark, Finland, Greece, Italy, Israel, Japan, Korea Rep., New Zealand, Portugal, Spain, United Kingdom, United States. |
| *Countries that change clubs over time:* |
| Argentina, Chile, Malaysia, Trinidad and Tobago, Venezuela RB. |

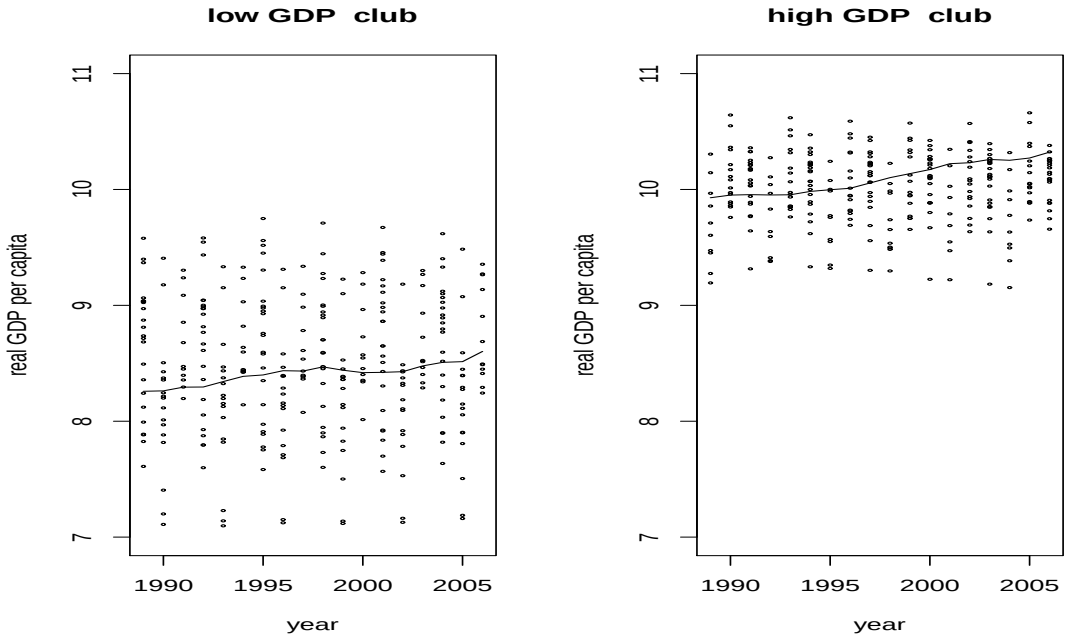
*Note*: The table presents the posterior results for the ordered probit model, sample with 33 countries. We report countries that stay in the same club over the sample period together with the respective posterior clubs, and the countries that change clubs over time. Clubs *high GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period. Club membership is determined by the posterior mode of the $S_{i,t}$ variable with an ordered probit specification.

Table 3.9: Ordered probit model extension (3.4): Initial GDP levels, growth rates and error variances for each club

| | *Club specification* | |
|---|---|---|
| | low GDP | high GDP |
| GDP level $\mu_{j,t}$ in 1989 | 8.26 | 9.93 |
| | (8.22, 8.30) | (9.93, 9.97) |
| Growth rate $\beta_j$ | 0.018 | 0.018 |
| | (0.001, 0.031) | (0.000, 0.029) |
| Error variances $\sigma^2_{\nu,j}$ | 0.104 | 0.103 |
| | (0.060, 0.172) | (0.058, 0.175) |

*Note*: The table shows posterior means for initial GDP levels ($\mu_{j,1989}$), growth rates ($\beta_j$) and error variances ($\sigma^2_{\nu,j}$) for each club for each club, for the Ordered probit model extension. 2.25% and 97.5% percentiles of the posterior densities are reported in parentheses. Clubs *high GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period.

Figure 3.4: Ordered probit model extension (3.4): Club memberships with covariates (financial intermediary and stock market development indicators)




*Note*: The figures show the mean posterior club levels (lines), and mean posterior club memberships (symbols) for the dataset with financial development indicators. Clubs *high GDP*, and *low GDP* are labeled according to posterior mean of GDP levels ($\mu_{j,t}$) in the last period. Club membership is determined by the posterior mode of the $S_{i,t}$ variable with an ordered probit specification.

GDP levels across clubs. Hence for this dataset, we conclude that there is no indication of convergence (or divergence) across clubs. The GDP clubs are rather defined in terms of two separate paths with similar growth rates, but different levels. Variations of GDP around the club-level (given by $\sigma^2_{\nu,j}$) are also similar across clubs. Hence the convergence clubs do not seem to be related to differences in GDP fluctuations.

Note that the results in terms of club levels and growth rates are in line with the results in Section 3.4.2: Most countries included in this smaller dataset were found to belong to the *medium* or *high* GDP clubs in the previous analysis. Table 3.9 shows that the club-specific GDP growth rates in the smaller dataset are around 2%, similar to the average GDP growth rate within the medium and high GDP clubs found before.

We now turn to the key aspect of this model, namely the effects of initial conditions and financial development indicators on GDP in the short-run and the long-run. Table 3.10 reports the posterior mean and mode values for the short- and long-run coefficients $\psi^s$ and $\psi^l$, together with the posterior probability that the respective coefficient is positive.

We first discuss the long-run effects of the covariates reported in the top panel of Table 3.10. Recall that a positive coefficient indicates that the probability of belonging to the club with a higher average GDP level is increasing with the respective covariate. According to the posterior mean and mode values reported in Table 3.10, all factors we consider have such a positive effect on the GDP level in the long-run. Notice that initial

Table 3.10: Ordered probit model extension (3.4): Posterior results for the effects of financial development indicators

Long-run effects of financial development indicators

| | mean | median | post. prob.[a] |
|---|---|---|---|
| initial GDP | 4.07 | 4.09 | 1.00 |
| *Financial intermediary development:* | | | |
| Bank assets | 0.47 | 0.47 | 0.94 |
| Commercial/Central Bank | 2.18 | 2.14 | 0.99 |
| *Stock market development:* | | | |
| turnover | 0.67 | 0.66 | 1.00 |

Short-run effects of financial development indicators

| | mean | median | post. prob.[a] |
|---|---|---|---|
| *Financial intermediary development:* | | | |
| Bank assets | -0.36 | -0.36 | 0.16 |
| Commercial/Central Bank | -1.27 | -1.31 | 0.17 |
| *Stock market development:* | | | |
| turnover | -0.09 | -0.09 | 0.00 |

*Note*: The table summarizes the posterior results for the effects of financial development indicators: coefficients for the financial intermediary development variables and the stock market development variable.

[a] denotes the posterior probability that the coefficient is positive.

condition is the only covariate that is constant over time. Hence the substantial effect of this covariate explains the reported persistence in club memberships.

The significance of these long-run effects is shown in Table 3.10 by the posterior probability that the respective coefficient is positive. These posterior probabilities exceed 0.9 for all variables, indicating that the long-run effects are quite significant.

Next, we consider the short-run effects of financial development indicators, reported in the bottom panel of Table 3.10. A positive coefficient in this case means that the fluctuations of countries' GDP around the long-run path is affected positively by the respective factor. According to posterior mean and mode values, all factors we analyze have a negative effect on GDP in the short-run. This result is opposite to the long-run effects, and supports the evidence documented by Loayza and Ranciere (2006).

The significance of the short-run effects differs across variables. The posterior probability of a positive coefficient for stock market development in Table 3.10 is zero, indicating that this variable has a clear negative effect in the short-run. On the other hand, the posterior probabilities of positive short-run effects for the financial development indicators are around 15%, corresponding to posterior probabilities around 85% for negative short-run effects. Hence the negative effects of financial development indicators in the short-run are less clear. Therefore we conclude that financial development has a deteriorating effect on the countries' GDP levels mainly through changes in stock market development.

## 3.5 Conclusion

In this paper we develop a statistical approach to model the GDP convergence process for a large set of countries, and to assess the short-run and long-run effects of financial development on the GDP process. We introduce a novel Markov Chain State Space Model that allows for changes in club memberships over time. We further extend this model allowing for certain covariates to affect the club memberships, as well as the short-run fluctuations around the long-run GDP levels.

In an empirical application to a large cross-section of countries we find that the club memberships are quite persistent over time, but nevertheless the composition of the clubs changes substantially during the time period 1970-2007. This result points out the importance of taking the dynamic properties of the GDP data into account. In terms of the factors affecting club memberships, we first note that initial conditions, measured by initial real GDP per capita, are important determinants of club membership (for a subset of these countries). Furthermore, we find that an increase in the level of financial development can move a country to a higher GDP club. In terms of the short-run effects of

financial development, we find a deteriorating effect of financial development specifically through stock market development.

As a final point, we note that the model we propose does not explicitly deal with changing number of clubs over time. In future work, we intend to extend our model allowing for changing number of clubs over time.

# Chapter 4

# A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood

*Chapter 4 is based on Ardia, Baştürk, Hoogerheide, and Van Dijk (2010).*

## 4.1 Introduction

This paper provides a comparative study on the efficiency of some commonly used Monte Carlo estimators of marginal likelihood in the context of highly non-elliptical posterior distributions. As the key ingredient in Bayes factors, the marginal likelihood lies at the heart of model selection and model discrimination in Bayesian statistics, see e.g., Kass and Raftery (1995). In several cases of scientific analysis, e.g., in non-linear regression models, mixture models, and instrumental variables models, one deals with target distributions that may have very non-elliptical contours and that are not members of a known class of distributions.

We focus is on the situation in which one uses either Importance Sampling (IS; due to Hammersley and Handscomb (1964), introduced in econometrics and statistics by Kloek and Van Dijk (1978) ), or the independence chain Metropolis-Hastings algorithm (MH; Metropolis *et al.* (1953), Hastings (1970)) for posterior simulation. That is, our analysis is especially relevant for those cases where the model structure implies that Gibbs sampling (Geman and Geman, 1984b) is not feasible; e.g., non-linear models like the example model of Ritter and Tanner (1992) that we will consider in Section 4. Obviously, the Griddy-Gibbs sampler of Ritter and Tanner (1992) is still feasible in such cases, but we

discard this approach due to the computational efforts that it requires. For the Griddy-Gibbs sampler the computing time required for obtaining results with a high precision is typically enormously larger than for the IS and MH approaches.

The marginal likelihood is given by

$$p(y) = \int_{\theta\in\Theta} k(\theta \mid y)d\theta = \int_{\theta\in\Theta} p(y \mid \theta)p(\theta)d\theta, \tag{4.1}$$

where $\theta$ denotes the set of parameters of interest; $\Theta$ is the parameter space; $k(\theta \mid y) = p(y \mid \theta)p(\theta)$ is the kernel function of the joint posterior $p(\theta \mid y)$; $p(y \mid \theta)$ is the likelihood function of $\theta$ for the vector of observations $y = (y_1 \cdots y_T)'$; $p(\theta)$ is the exact prior density of $\theta$. This marginal likelihood (sometimes also referred to as model likelihood; see e.g., Frühwirth-Schnatter (2001)) is equal to the normalizing constant of the joint posterior density. The estimation of $p(y)$ can be a difficult task in practice, especially for complex statistical models.

The aim of this paper is to investigate the effect that strategic choices may have on the results when estimating a marginal likelihood. We argue that these choices are important for the following issues:

(i) the sensitivity to the choice of the particular sampling procedure (either IS or MH);

(ii) the sensitivity to the choice of the candidate distribution (e.g., a Student-$t$ distribution or a mixture of Student-$t$ distributions);

(iii) the impact of aiming at the posterior density kernel or aiming at a 'warped' version of it;

(iv) the reliability of different types of numerical standard errors (NSE's) as signals for the uncertainty on the respective estimators.

The analysis of the robustness and efficiency of these estimators in the context of non-elliptical posteriors has not been much investigated so far. Frühwirth-Schnatter (2004) provides an excellent survey but it is restricted to the special case of mixture models. Our results demonstrate the importance of a robust and flexible estimation strategy which explores the full joint posterior. A poor choice of the importance density can lead to a substantial loss of efficiency, where the numerical standard error can be highly unreliable. On the other hand, given an appropriately chosen candidate density, the straightforward IS approach provides a computationally efficient marginal likelihood estimator along with a reliable numerical standard error. The approach of Hoogerheide *et al.* (2007b) that constructs an adaptive mixture of Student-$t$ distributions (AdMit) is particularly useful for automatically obtaining an appropriate candidate density.

This article proceeds as follows. Section 4.2 provides a summary of some commonly used Monte Carlo estimators of the marginal likelihood. Section 4.3 gives a brief overview of the AdMit approach. In Section 4.4 we investigate the robustness and efficiency of these estimators in the case of a three-dimensional highly non-elliptical example distribution, a posterior distribution in a non-linear regression model. In Section 4.5 we consider the reliability of numerical standard errors. In Section 4.6 we analyze the performance in a mixture GARCH model. Section 4.7 concludes.

## 4.2 Some Monte Carlo methods for marginal likelihood estimation

We first summarize some of the most commonly used Monte Carlo estimators of marginal likelihood. For more details, see Ardia *et al.* (2009c). We extend the overview of Frühwirth-Schnatter (2004) on Monte Carlo estimators of marginal likelihood by including the approach of Chib and Jeliazkov (2001), and addressing some more details on implementation, advantages and drawbacks of alternative methods. We especially pay attention to the case of the one-block independence chain MH approach. Further review papers that deal with a comparative review of marginal likelihood estimation methods are Han and Carlin (2001) and Miazhynskaia and Dorffner (2006).

**Importance sampling (IS)**

The IS estimator (Hammersley and Handscomb, 1964; Kloek and Van Dijk, 1978; Van Dijk and Kloek, 1980; Geweke, 1989) is given by

$$\hat{p}_{\text{IS}}(y) = \frac{1}{L} \sum_{l=1}^{L} \frac{k(\theta^{[l]} \,|\, y)}{q(\theta^{[l]})}, \tag{4.2}$$

where $\{\theta^{[l]}\}_{l=1}^{L}$ are i.i.d. draws from the exact importance density $q$ which should approximate the joint posterior density $p(\theta \,|\, y)$. The IS approach of marginal likelihood estimation is a *globally oriented* method that aims at directly evaluating the integral $\int_{\theta \in \Theta} k(\theta \,|\, y) d\theta$ over the whole parameter space $\Theta$. An importance density which *globally* matches the joint posterior closely will lead to efficient estimation. For this purpose, the tails of $q$ should not be thinner than the tails of the posterior. That is, $q$ should 'wrap' the posterior density in the sense that all areas of the parameter space $\Theta$ that contain substantial posterior probability mass must be 'wrapped' with a reasonable amount of candidate probability mass.

**Reciprocal importance sampling (RIS)**

The RIS estimator (Gelfand and Dey, 1994) is given by

$$\hat{p}_{\mathrm{RIS}}(y) = \left[ \frac{1}{M} \sum_{m=1}^{M} \frac{q_{\mathrm{aux}}(\theta^{[m]})}{k(\theta^{[m]} \,|\, y)} \right]^{-1}, \tag{4.3}$$

where $\{\theta^{[m]}\}_{m=1}^{M}$ are (correlated) posterior draws from an MCMC sampler. $q_{\mathrm{aux}}$ is an exact auxiliary density from which we do not require draws. That is, even if the MCMC draws $\{\theta^{[m]}\}_{m=1}^{M}$ are simulated using a candidate density, then this candidate density should generally not be $q_{\mathrm{aux}}$. The RIS approach makes use of the fact that *for each* $\theta \in \Theta$ there holds $p(y) = k(\theta \,|\, y)/p(\theta \,|\, y)$. High efficiency is most likely to result if $q_{\mathrm{aux}}$ roughly matches the posterior density. However, the RIS estimator is still consistent if $q_{\mathrm{aux}}$ only covers a small part of the parameter space $\Theta$. For stability of the estimator, the tails of $q_{\mathrm{aux}}(\theta)$ should not be fatter than those of the posterior in order to keep the ratio $q_{\mathrm{aux}}(\theta)/k(\theta \,|\, y)$ bounded. Van Dijk and Kloek (1980), Hop and Van Dijk (1992) and Gelfand and Dey (1994) propose a multivariate Gaussian or Student-$t$ density whose mean vector and covariance matrix are estimated from the joint posterior sample. Geweke (1999) proposes the use of a multivariate Gaussian density, truncated to a subspace of $\Theta$.

An advantage of the RIS estimator is that the auxiliary density $q_{\mathrm{aux}}$ does not have to cover the whole posterior. Still, we do require that the MCMC draws $\{\theta^{[m]}\}_{m=1}^{M}$ be representative of the whole posterior distribution, otherwise the RIS estimator is no longer consistent.

A special case of (4.3) is the harmonic mean estimator by Newton and Raftery (1994), in which the prior $p(\theta)$ is used as the auxiliary density. However, it is well-known that this estimator is unstable. Therefore, we do not investigate the version of the harmonic mean.

**(Optimal) bridge sampling (BS)**

The BS estimator (Meng and Wong, 1996) is obtained as the limit of the sequence

$$\hat{p}_{\mathrm{BS}}^{(t)}(y) = \hat{p}_{\mathrm{BS}}^{(t-1)}(y) \cdot \frac{\frac{1}{L} \sum_{l=1}^{L} \frac{\hat{p}(\theta^{[l]} \,|\, y)}{Lq(\theta^{[l]}) + M\hat{p}(\theta^{[l]} \,|\, y)}}{\frac{1}{M} \sum_{m=1}^{M} \frac{q(\theta^{[m]})}{Lq(\theta^{[m]}) + M\hat{p}(\theta^{[m]} \,|\, y)}}, \tag{4.4}$$

where $\hat{p}(\theta \,|\, y) = k(\theta \,|\, y)/\hat{p}_{\mathrm{BS}}^{(t-1)}(y)$ and the initial value $p_{\mathrm{BS}}^{(0)}(y)$ is set to (4.2), for instance. The $\{\theta^{[m]}\}_{m=1}^{M}$ are (correlated) posterior draws from an MCMC sampler and $\{\theta^{[l]}\}_{l=1}^{L}$ are i.i.d. draws from the importance density $q$. Usually, we set $M = L$. Convergence of the bridge sampling technique requires few steps in practice (i.e., typically less than ten iterations). Moreover, these steps do not require many additional computational efforts: no extra draws or evaluations of candidate or target densities are needed. The BS estimator

provides (asymptotically) the optimal combination of draws $\{\theta^{[m]}\}_{m=1}^{M}$ and $\{\theta^{[l]}\}_{l=1}^{L}$ for the estimation of a (ratio of) normalizing constant(s). That is, the BS estimator gives the optimal *bridge* between the posterior kernel and the candidate density $q$. The original BS estimator in (4.4) is optimal if the draws $\{\theta^{[m]}\}_{m=1}^{M}$ are i.i.d. We refer to this estimator as the BS1 estimator. A simple correction for correlated draws is proposed by Meng and Schilling (2002). This correction means that one substitutes $M$ by an 'effective number of draws' $\tilde{M}$, defined as $\tilde{M} = M(1-\rho_1)/(1+\rho_1)$ with $\rho_1$ the first order serial correlation of the likelihood evaluations of the $\{\theta^{[m]}\}_{m=1}^{M}$. We refer to this estimator as the BS2 estimator.

In general, an advantage of the BS estimator is that its variance depends on a ratio that is bounded regardless of the tail behavior of the importance density $q$, which renders the estimator robust. A disadvantage is that we require both a set of draws from the posterior and a set of independent candidate draws. Further, it requires some implementation cost. It has been investigated by Frühwirth-Schnatter (2004) in the context of mixture models, where it has shown a good performance.

The optimal bridge sampling estimator is a special case of the general bridge sampling (GBS) estimator

$$\hat{p}_{\mathrm{GBS}}(y) = \frac{\frac{1}{L}\sum_{l=1}^{L} \alpha(\theta^{[l]}) k(\theta^{[l]} \mid y)}{\frac{1}{M}\sum_{m=1}^{M} \alpha(\theta^{[m]}) q(\theta^{[m]})}. \tag{4.5}$$

The IS and RIS estimators are also members of this class of GBS estimators: these correspond to the choices of $\alpha_{\mathrm{IS}}(\theta) = 1/q(\theta)$ and $\alpha_{\mathrm{RIS}}(\theta) = 1/k(\theta \mid y)$, respectively. The BS1 estimator corresponds to the choice

$$\alpha_{\mathrm{BS1}}(\theta) \propto \frac{1}{L\, q(\theta) + M\, p(\theta \mid y)},$$

that asymptotically minimizes the relative error of the GBS estimator $\hat{p}_{\mathrm{GBS}}(y)$ if the posterior draws $\{\theta^{[m]}\}_{m=1}^{M}$ are independent.

**Chib and Jeliazkov (2001) (CJ)**

The CJ estimator for marginal likelihood estimation on the basis of MH draws is given by

$$\hat{p}_{\mathrm{CJ}}(y) = \frac{k(\theta^* \mid y)}{\hat{p}(\theta^* \mid y)}, \tag{4.6}$$

where $\theta^*$ is a certain point in the parameter space $\Theta$ with $p(\theta^* \mid y) > 0$. In the case of the independence chain MH algorithm, the estimated density $\hat{p}(\theta^* \mid y)$ of the CJ estimator is given by

$$\hat{p}(\theta^* \mid y) = q(\theta^*) \frac{\frac{1}{M}\sum_{m=1}^{M} \alpha_{\mathrm{MH}}(\theta^{[m]}, \theta^*)}{\frac{1}{L}\sum_{l=1}^{L} \alpha_{\mathrm{MH}}(\theta^*, \theta^{[l]})}, \tag{4.7}$$

with $\alpha_{\mathrm{MH}}(\theta, \theta')$ the probability that a transition from $\theta$ to $\theta'$ is accepted in the MH algorithm,

$$\alpha_{\mathrm{MH}}(\theta, \theta') = \min \left\{ 1, \frac{k(\theta' \mid y)}{k(\theta \mid y)} \frac{q(\theta)}{q(\theta')} \right\} .$$

The CJ approach can be applied for each $\theta^* \in \Theta$ with $p(\theta^* \mid y) > 0$. However, for efficiency, the point $\theta^*$ must be taken to be a high-density point in $\Theta$, typically the posterior mode. In the case of a highly non-elliptical posterior distribution it may be a bad strategy to use the (estimated) posterior mean, as this may have a low (or even zero) posterior density value.

The CJ estimator is another member of the class of GBS estimators, corresponding to the choice of

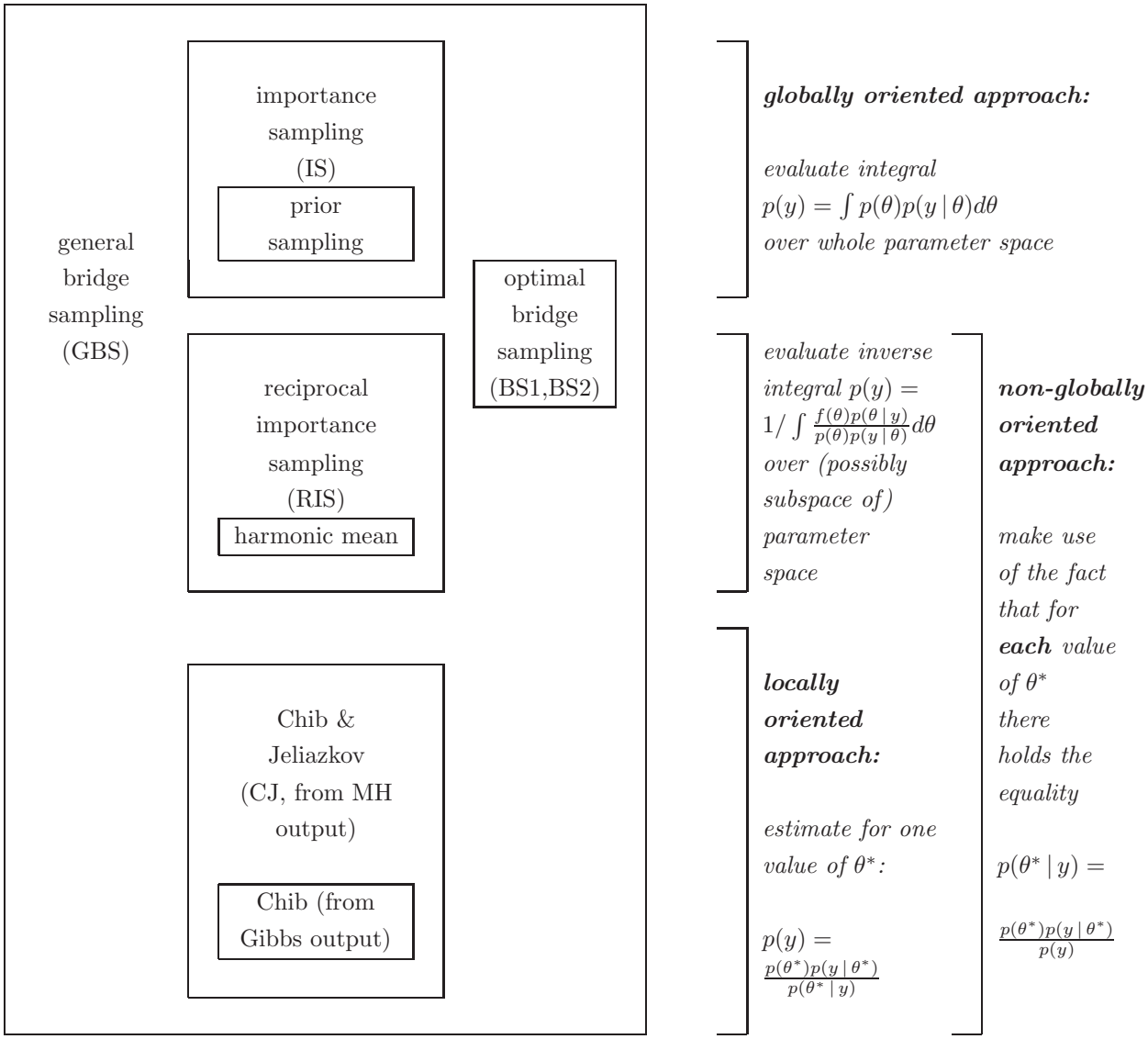
$$\alpha_{\mathrm{CJ},\theta^*}(\theta) = \min \left\{ \frac{q(\theta^*)}{q(\theta)}, \frac{k(\theta^* \mid y)}{k(\theta \mid y)} \right\} .$$

See Meng and Schilling (2002) and Mira and Nicholls (2004) who show that also other variations proposed by Chib and Jeliazkov (2001) are individual cases of bridge sampling. This suggests that the CJ approach should always be dominated by the optimal BS method. However, BS1 is only optimal: (i) asymptotically; and (ii) if the posterior draws were i.i.d.. For the BS2 estimator, the optimality is also asymptotical and the 'effective number of draws' may provide a crude correction. Therefore, it still makes sense to compare the performance of the CJ and BS methods.

Of the approaches that we consider, the CJ method is the *most local* method: we only estimate the posterior density in one point $\theta^*$. This is in sharp contrast with the IS approach where the whole posterior is 'wrapped' by a fat-tailed candidate. In between we have the RIS method, where (possibly a subspace of) the parameter space is covered by a thin-tailed auxiliary density. A graphical overview of these methods is given by Figure 4.1.

The Gibbs sampler is a special case of the MH approach, so that the method of Chib (1995) that estimates the marginal likelihood from Gibbs draws, is a special case of the CJ method. In the case of IS we can in principle use the prior as the importance density. However, we do not consider this option in this paper, as this approach is typically very inefficient; see Van Dijk (1999). In general, the prior has much higher variance than the posterior, so that the IS estimate would then be based on only a few IS weights (i.e., likelihood evaluations), with most likelihood values being close to zero.

Figure 4.1: Classification of some well-known methods for estimating marginal likelihoods




*Note*: All estimators are members of the class of general bridge sampling estimator.

### Warping

The methods above can be used in combination with another technique: warping the target posterior (see Meng and Schilling (2002)). If we assume that the parameter space of $\theta$ is $\Theta = \mathbb{R}^d$, then

$$p(y) = \int_{\theta \in \Theta} k(\theta \mid y)d\theta = \int_{\theta \in \Theta} \frac{1}{2}\left[k(\theta \mid y) + k(-\theta + 2\theta_0 \mid y)\right]d\theta\,. \tag{4.8}$$

This implies that application of the aforementioned methods to the *warped* posterior kernel

$$\tilde{k}(\theta \mid y) = \frac{1}{2}\left[k(\theta \mid y) + k(-\theta + 2\theta_0 \mid y)\right] , \tag{4.9}$$

rather than to the posterior kernel $k(\theta \mid y)$, also yields an estimator of the marginal likelihood. The *warped* posterior kernel $\tilde{k}(\theta \mid y)$ is point symmetric around $\theta_0$, where one typically chooses $\theta_0$ as the (estimated) posterior mean. This gain in symmetry may substantially improve the approximation of the target density by the candidate density, typically a symmetric density (e.g., Gaussian and Student-$t$). This may yield a substantial increase in efficiency. However, a disadvantage is that for each candidate draw we now require two evaluations of the posterior density kernel instead of one. We refer to the transformation in (4.9) as the Warp1 transformation.

In the two terms of the Warp 1 transformation in (4.9) we either take the original parameter vector $\theta$ or the 'mirror image' of all elements. A further gain in symmetry is obtained by taking an average over all $2^d$ combinations where individual elements of $\theta$ may be 'mirrored'. Obviously, a disadvantage is that for increasing values of the dimension $d$, the number of posterior kernel evaluations per candidate draw increases exponentially. We refer to this transformation as the Warp2 transformation.

Meng and Schilling (2002) use the name Warp-III for both these Warp1 and Warp2 transformations: Warp-I and Warp-II correspond to adapting the location and variance of the target density to the candidate. We always use candidate distributions of which the location and variance are adapted to the target, so that we only explicitly make use of the Warp-III type transformation that eliminates asymmetries via mixtures of the target.

Table 4.1 provides an overview of the number of candidate draws and function evaluations that are required by different methods. The candidate distributions that we will consider are Student-$t$ distributions and mixtures of Student-$t$ distributions. The auxiliary densities (of RIS) will be truncated Gaussian. Evaluations of these densities and the simulation of pseudo-random draws from these distributions is done easily and quickly. Therefore, the computational efforts mainly depend on the number of posterior kernel evaluations. For a *fair* comparison between methods, we apply these in such a way that the numbers of posterior kernel evaluations are equal. The IS and RIS estimators are members of the general bridge sampling (GBS) class of which the BS2 estimator is (approximately, asymptotically) optimal. However, this result holds for $L$ and $M$ taken equal in IS, RIS and BS. In this paper the equal numbers of posterior kernel evaluations imply that we take $L_{\mathrm{IS}}$ and $M_{\mathrm{RIS}}$ twice as large as $L_{\mathrm{BS}} = M_{\mathrm{BS}}$, so that IS and RIS could very well outperform BS.

We focus on the cases of IS and the independence chain MH algorithm. So, we compare the following strategies:

(IS) use all candidate draws in the IS estimator (4.2);

(RIS, CJ) transform all candidate draws to a sequence of MH draws (plus a burn-in) and use these in the RIS estimator (4.3) or the CJ estimator (4.6);

(BS) transform 50% of the candidate draws to a sequence of MH draws (plus a burn-in) and combine these with the other 50% of the candidate draws in the BS1 estimator (4.4) – with $M$ substituted by the 'effective number of draws' $\tilde{M}$ for the BS2 estimator.

In Sections 4, 5 and 6 the methods will be applied to several target distributions. In the next section we briefly review the method of Hoogerheide *et al.* (2007b) that uses an adaptive mixture of Student-$t$ distributions (AdMit) as the importance or candidate distribution.

Table 4.1: Computations required by different marginal likelihood estimation approaches, in case we make use of IS or the independence chain MH algorithm.

| | number of posterior kernel evaluations | number of candidate draws | number of candidate density evaluations | number of auxiliary density evaluations |
|---|---|---|---|---|
| IS | $L$ | $L$ | $L$ | - |
| RIS | $M$ | $M$ | $M$ | $M$ |
| BS | $L+M$ | $L+M$ | $L+M$ | - |
| CJ | $L+M$ | $L+M$ | $L+M$ | - |
| Warp1 IS | $2L$ | $L$ | $L$ | - |
| Warp1 BS | $2(L+M)$ | $L+M$ | $L+M$ | - |
| Warp2 IS | $2^dL$ | $L$ | $L$ | - |
| Warp2 BS | $2^d(L+M)$ | $L+M$ | $L+M$ | - |

*Note:* $L$ is the number of candidate draws that are not used in the MH algorithm. $M$ is the number of independence chain MH draws from the posterior. Warp1 and Warp2 refer to the Warp transformations of Meng and Schilling (2002) where one aims at a mixture of 2 or $2^d$ 'mirror images' of the posterior density that is typically more symmetric than the posterior itself. Further explanations are given in Section 4.2.

## 4.3 The Adaptive Mixture of Student-$t$ method

The Adaptive Mixture of Student-$t$ (AdMit) approach (Hoogerheide *et al.*, 2007b) consists of two steps. First, it constructs a mixture of Student-$t$ distributions which approximates a target distribution of interest. The fitting procedure relies only on a kernel of the target density, so that the normalizing constant is not required. In a second step, this approximation is used as an importance function in IS (or as a candidate density in the independence chain MH algorithm) to estimate characteristics of the target density. The estimation procedure is fully automatic and thus avoids the difficult task, especially for non-experts, of tuning a sampling algorithm. In a standard case of IS the candidate density is unimodal. Then a multimodal target distribution may lead to some draws having huge importance weights or some modes may even be completely missed. Thus, an important problem is the choice of the importance density, especially when little is known a priori about the shape of the target density. The importance density should be close to the target density, and it is especially important that the tails of the candidate should not be thinner than those of the target. Hoogerheide *et al.* (2007b) mention several reasons why mixtures of Student-$t$ distributions are natural candidate densities. First, they can provide an accurate approximation to a wide variety of target densities, with substantial skewness and high kurtosis. Furthermore, they can deal with multi-modality and with non-elliptical shapes due to asymptotes. Second, this approximation can be constructed in a quick, iterative procedure and a mixture of Student-$t$ distributions is easy to sample from. Third, the Student-$t$ distribution has fatter tails than the Gaussian distribution; especially if one specifies Student-$t$ distributions with few degrees of freedom, the risk is small that the tails of the candidate are thinner than those of the target distribution. Finally, Zeevi and Meir (1997) showed that under certain conditions any density function may be approximated to arbitrary accuracy by a convex combination of basis densities; the mixture of Student-$t$ distributions falls within their framework.

The AdMit approach determines the number of mixture components, the mixing probabilities, the modes and scale matrices of the components in such a way that the mixture density approximates the target density $p(\theta \mid y)$ of which we only know a kernel function $k(\theta \mid y)$. The AdMit strategy consists of the following steps:

(0) Initialization: computation of the mode and scale matrix of the first component (typically the posterior mode and minus the inverse Hessian of the log-posterior evaluated at the mode), and drawing a sample from this Student-$t$ distribution;

(1) Iterate on the number of components: add a new component that covers a part of the space of $\theta$ where the previous mixture density was relatively small, as compared

to $k(\theta \mid y)$. The new component is based on the ratio of the target density kernel $k(\theta \mid y)$ and the previous mixture of Student-$t$ densities. It is located where this ratio is *relatively* high, which does not depend on the normalizing constant of the target density;

(2) Optimization of the mixing probabilities: the mixing probabilities are chosen such that the coefficient of variation, i.e., the standard deviation divided by the mean, of the IS weights is minimized. This coefficient of variation does not depend on the normalizing constant of the target density;

(3) Drawing a sample from the new mixture;

(4) Evaluation of IS weights: if the coefficient of variation of the IS weights has converged, then stop. Otherwise, go to step (1).

For more details on the AdMit procedure we refer to Hoogerheide *et al.* (2007b), Ardia *et al.* (2009a) and Ardia *et al.* (2009b). The package `AdMit` (Ardia *et al.*, 2008), an R implementation (R Development Core Team, 2008), is available from the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/package=AdMit`.

The AdMit approach has been successfully applied to the simulation of posterior draws from non-elliptical posterior distributions, where the reason for non-elliptical shapes is typically *local non-identification* of certain parameters. Examples are the IV model with weak instruments, or mixture models where one component has weight close to zero. This paper provides the first analysis of the AdMit method's performance in the case of marginal likelihood estimation.

## 4.4 Application 1: non-linear regression model

In this section we apply our methods in order to estimate the marginal likelihood in a non-linear regression model. We consider the biochemical oxygen demand (BOD) data from Marske (1967) that are analyzed by Bates and Watts (1988) and Ritter and Tanner (1992).

We consider the non-linear model of Bates and Watts (1988)

$$y_i = \theta_1(1 - \exp(-\theta_2\, x_i)) + \varepsilon_i\,, \tag{4.10}$$

with independent errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, where $y_i$ is the BOD at time $x_i$ $(i = 1, \ldots, 6)$.

Following Ritter and Tanner (1992), we specify a flat prior. However, we use a flat prior on a bounded domain: $(\theta_1, \theta_2, \sigma) \in [-20, 50] \times [-2, 6] \times [0, 20]$. Ritter and Tanner

(1992) do not restrict the interval of $\sigma$; for the identification of a marginal likelihood we make this choice in order to have a proper prior. Obviously, the marginal likelihood will crucially depend on the prior specification. We consider the model and data from Ritter and Tanner (1992) in order to compare the efficiency of alternative estimation methods and illustrate the results in the case of a well-known, three-dimensional highly non-elliptical posterior distribution for a very small data set. In Section 4.6 we will consider a marginal likelihood and posterior distribution for a large data set.

Especially if one specifies uninformative priors, one can argue that one should not use the marginal likelihood but instead use the predictive likelihood (see e.g., Eklund and Karlsson (2007)) for model choice or combination. The predictive likelihood may be computed as the ratio of two marginal likelihoods, the marginal likelihood for the whole data set divided by the marginal likelihood for a subset of the data, the so-called *training sample*. Therefore, the efficient computation of marginal likelihoods is also important when one bases model choice or combination on the predictive likelihood. Our focus on the marginal likelihood must not be interpreted as a statement that we consider it appropriate to base model choice directly on the marginal likelihood in case of a noninformative prior. We only restrict our focus to comparing the efficiency of different estimation methods, so that a comparison between marginal and predictive likelihood, and between different choices for the training sample, is outside the scope of this paper. We think that the search for *the good approach* for model choice or combination in case of no or little prior knowledge is a highly interesting, but separate topic.

The top-left panel of Figure 4.2 gives an illustration of the shapes of the posterior distribution of $\theta = (\theta_1, \theta_2, \sigma)'$; it shows a Highest Posterior Density (HPD) credible set. Note the bimodality and the curved shapes of the larger mode. The sets $\{\theta : \theta_1 > 0, \theta_2 > 0\}$ and $\{\theta : \theta_1 < 0, \theta_2 < 0\}$ correspond to concave and convex increasing functions (through the origin) in (4.10), respectively. The smaller mode reflects the small posterior probability of a convex function.

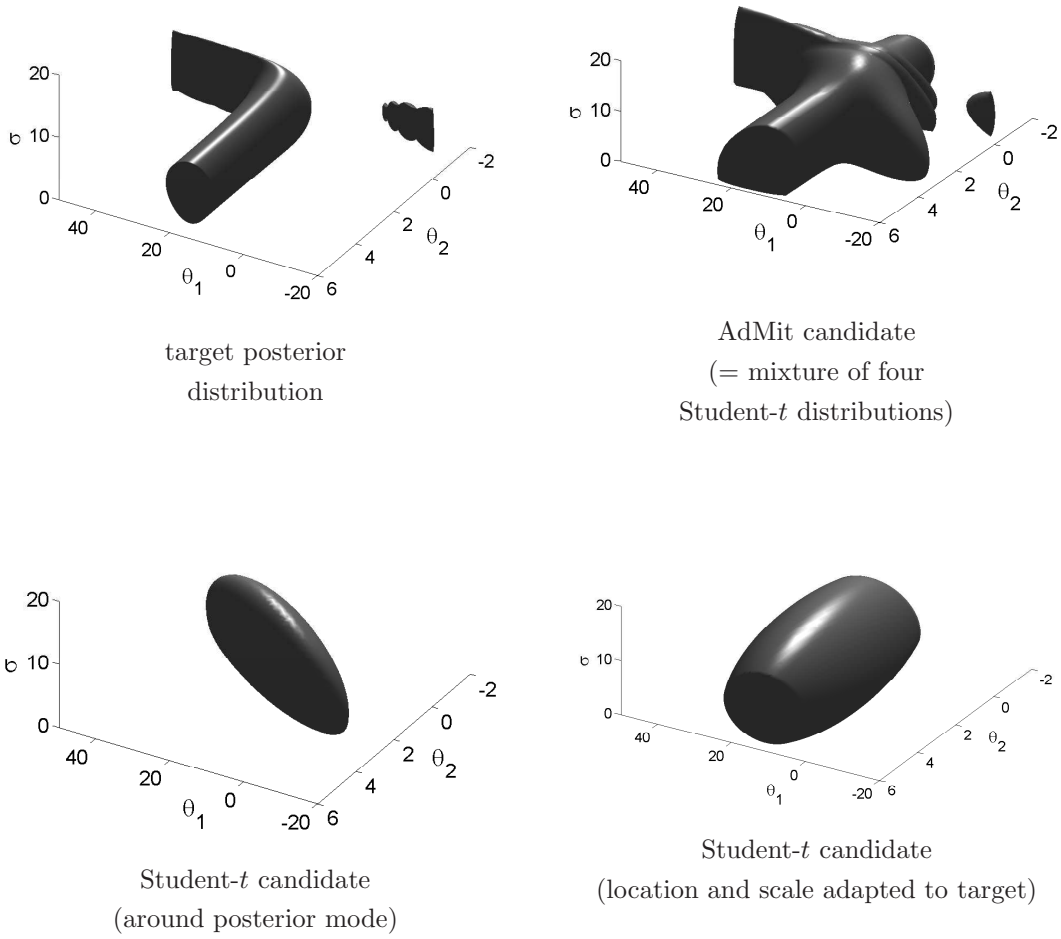
For the IS and independence chain MH algorithms we consider three candidate distributions:

1. the mixture of Student-$t$ distributions resulting from the AdMit procedure;

2. an 'adaptive' Student-$t$ distribution where the mode and scale have been iteratively updated by several IS steps (starting with the posterior mode and iteratively using the estimated posterior mean and covariance as the mode and scale in the next iteration);

3. a so-called 'naive' Student-$t$ distribution around the posterior mode.

Figure 4.2: Non-linear regression model (4.10): Highest Posterior Density credible region and Highest Candidate Density regions for approximating densities




target posterior distribution

AdMit candidate (= mixture of four Student-$t$ distributions)

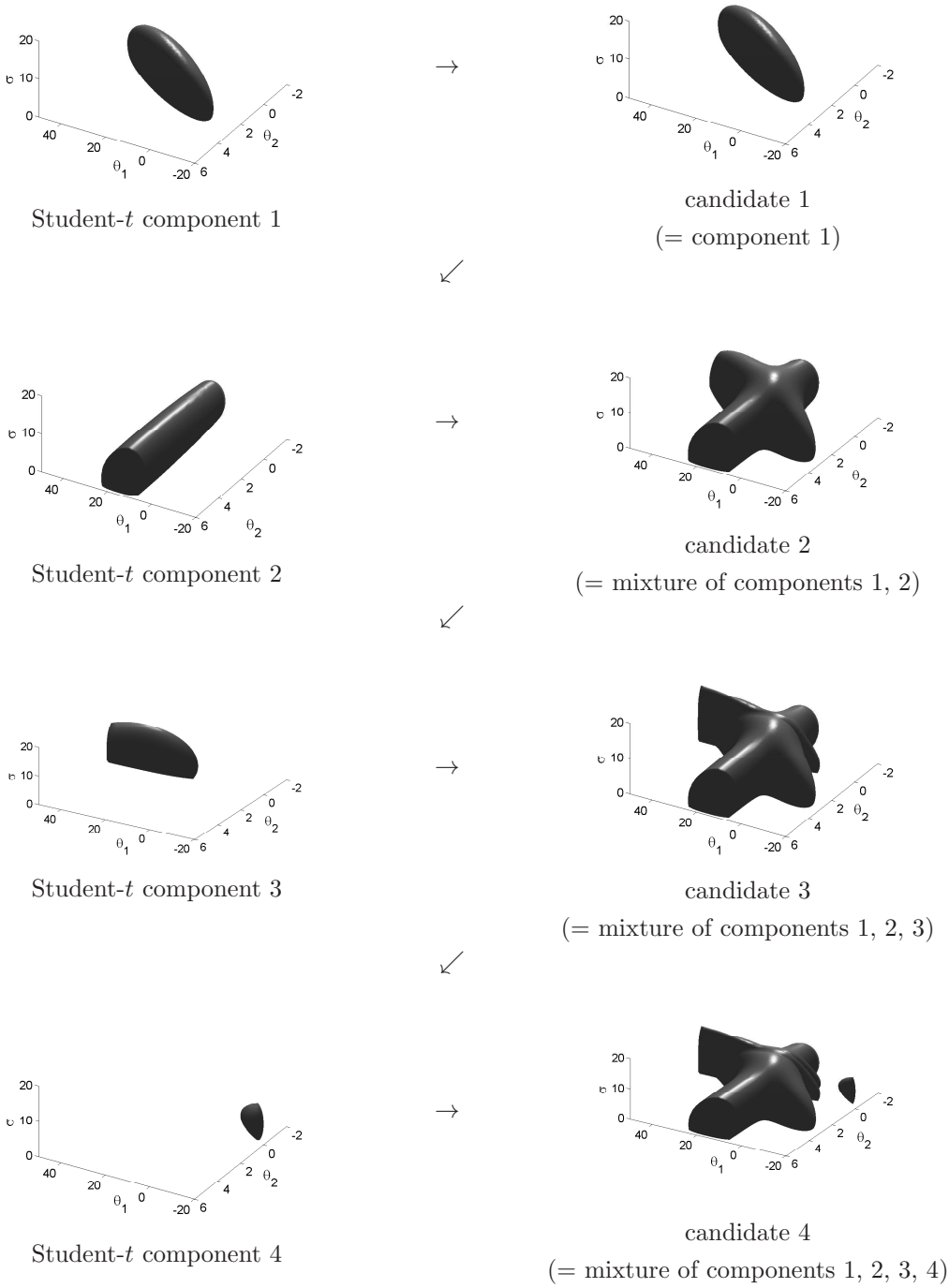
Student-$t$ candidate (around posterior mode)

Student-$t$ candidate (location and scale adapted to target)

*Note*: The figure shows 'Highest Posterior Density credible region' of $\theta = (\theta_1, \theta_2, \sigma)'$ (top left) and 'Highest Candidate Density regions' for mixture of Student-$t$ (AdMit, top right), 'naive' Student-$t$ (bottom left) and adaptive Student-$t$ (bottom right) candidate distributions.

In order to minimize the risk that the candidate 'misses' parts of the posterior, we specify very fat-tailed candidates: we choose one degree of freedom (i.e., Cauchy tails). Figure 4.2 shows the shapes of the three candidate distributions. Notice that the AdMit candidate nicely 'wraps' the relevant areas of the parameter space with candidate probability mass. Figure 4.3 illustrates how the AdMit approach has constructed this 'wrapping' distribution.

Figure 4.3: Non-linear regression model (4.10): Step-by-step AdMit approximation to $\theta = (\theta_1, \theta_2, \sigma)'$




*Note*: The figure shows the Highest Posterior Density regions for the AdMit algorithm for each component (left panel), and for the achieved Student-$t$ mixture at each step (right panel).

We will now use these three candidate distributions in combination with the marginal likelihood estimators of Section 2. For the IS estimator we generate $L = 100000$ candidate draws. For the RIS and CJ estimators we take $M = 100000$ independence chain MH draws; we use a burn-in of 1000 draws, so that we actually generate 101000 draws. The reason for not including the burn-in in the 100000 draws is that a burn-in of fewer than 1000 draws may suffice. For the BS estimators we use $L = 50000$ candidate draws and $M = 50000$ MH draws, again not counting a burn-in of 1000 draws.

For the RIS estimator we use a truncated Gaussian auxiliary density around the posterior mode. For the CJ estimator we choose $\theta^*$ as the posterior mode.

For each estimator, we repeat the simulation 500 times. Simulation results are reported in Table 4.2. Boxplots of the 500 marginal likelihood estimates are given in Figure 4.4. The real value of the marginal likelihood is (rounded to two digits) $12.79 \cdot 10^{-10}$. This real value is computed by deterministic integration which is still feasible (but already quite time-consuming) in this three-dimensional example.

Table 4.2: Non-linear regression model (4.10): Estimation of the marginal likelihood (ML) based on 100000 draws from AdMit mixture of four Student-$t$ distributions, adaptive Student-$t$ or naive Student-$t$ distribution.

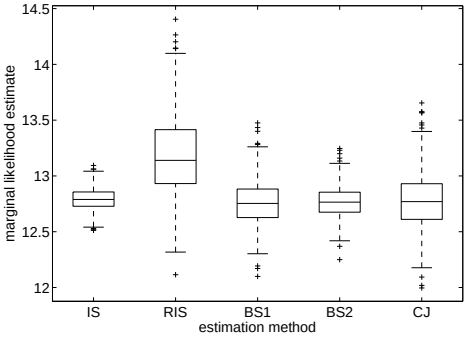
| $10^{10}\cdot$ ML | AdMit | | | adaptive | | | naive | | |
|---|---|---|---|---|---|---|---|---|---|
| | mean | st.dev. | (RMSE) | mean | st.dev. | (RMSE) | mean | st.dev. | (RMSE) |
| IS | 12.7906 | 0.0962 | (0.10) | 12.7899 | 0.1791 | (0.18) | 12.7317 | 1.0945 | (1.10) |
| RIS | 13.1803 | 0.3435 | (0.52) | 12.8792 | 0.9456 | (0.95) | 12.8846 | 2.5144 | (2.52) |
| BS1 | 12.7621 | 0.1984 | (0.20) | 12.8348 | 0.4238 | (0.43) | 13.0995 | 4.3776 | (4.39) |
| BS2 | 12.7636 | 0.1405 | (0.14) | 12.7890 | 0.2739 | (0.27) | 13.0877 | 4.2780 | (4.29) |
| CJ | 12.7816 | 0.2568 | (0.26) | 12.7814 | 0.2841 | (0.28) | 13.1030 | 4.4004 | (4.41) |

*Note:* Mean, standard deviation and root mean squared error of 500 estimates of $10^{10}\cdot$ ML from 500 simulation runs. True value is ML$= 12.79 \cdot 10^{-10}$.

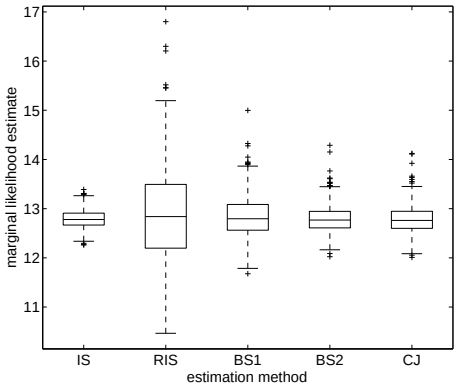
First, notice the very inefficient estimators that make use of the naive Student-$t$ candidate distribution. Even though this naive Student-$t$ distribution is chosen very fat-tailed (one degree of freedom), the resulting estimators have much higher variance than the estimators based on the AdMit and adaptive candidates. The boxplots show that the naive Student-$t$ candidate may result in extreme outliers for all marginal likelihood estimators. This stresses the importance of wisely specifying an appropriate candidate distribution.

Second, the AdMit candidate clearly outperforms the adaptive Student-$t$ candidate: iteratively adding Student-$t$ distributions to the mixture candidate distribution leads to

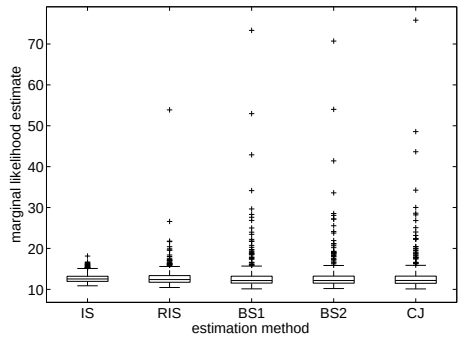
Figure 4.4: Non-linear regression model (4.10): Estimates of $10^{10}\cdot$ marginal likelihood under different candidate distributions




candidate draws from AdMit distribution




candidate draws from adaptive Student-$t$ distribution




candidate draws from naive Student-$t$ distribution

*Note*: The figure shows ML estimates using candidate distributions from the AdMit approach, adaptive Student-$t$ or naive Student-$t$ distribution. The estimates are based on 500 simulation runs, and 100000 draws for each simulation run.

far more precise estimators than merely iteratively adapting the location and scale of the Student-$t$ candidate.

Third, the IS estimator is the best, whereas the RIS estimator is clearly the worst. For the RIS estimator the bias is substantial, which results in a RMSE that is much larger than the standard deviation for AdMit-RIS. The BS2, BS1 and CJ are typically ranked second to fourth, although in case of the adaptive candidate the CJ estimator outperforms the BS1 estimator. In that case, the difference between the 'i.i.d. optimal' BS1 estimator and the 'serial correlation corrected' BS2 estimator is substantial, reflecting the high serial correlation in the MH chain.

In this example, the winner is clearly the AdMit-IS estimator, the IS estimator based on the AdMit candidate. It outperforms the alternative estimators (including the BS estimators) that make use of the same number of candidate draws and function evaluations.

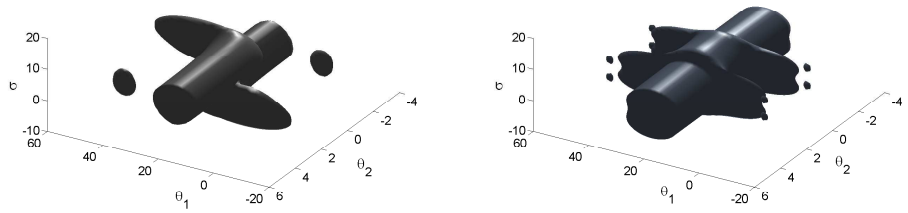
Simulating draws from a mixture of Student-$t$ distributions takes hardly more time than generating draws from a Student-$t$ distribution. The AdMit approach does require the evaluation of multiple Student-$t$ densities, in our case four, instead of one; but the little extra computing time required for this is typically very small compared to the time required for evaluation of the posterior density kernel. Further, the 'victory' of the IS estimator over alternative estimators is actually slightly larger than that represented by the tables: the burn-in of the MCMC draws is neglected and the implementation of the IS estimation of the marginal likelihood and its numerical standard error are relatively straightforward.

In this example, one comparison is still to be made: the comparison with methods aimed at the 'warped' target density. Figure 4.5 shows the shapes of the warped posterior kernels. These are more symmetric than the posterior kernel itself; especially the Warp2 distribution looks 'closer to' a Student-$t$ distribution than the original posterior distribution. This illustrates the elimination of asymmetries by using mixtures of the posterior distribution. Table 4.3 shows the results of IS, BS1 and BS2 (the three best performing algorithms) for Warp1 and Warp2 transformations in combination with an adaptive Student-$t$ candidate. The rows with 100000 posterior kernel evaluations correspond to IS with 50000 and 12500 draws (BS with 25000+25000 and 6250+6250 draws) for Warp1 and Warp2, respectively. The Warp1-IS results are comparable to the regular IS results with an adaptive Student-$t$ candidate. The Warp1-BS estimators are somewhat better than the 'unwarped' BS estimators. The Warp2 results are worse than their 'unwarped' counterparts; the obvious reason is that the number of candidate draws is now much smaller in order to keep the number of posterior kernel evaluations equal to 100000.

Even if we use the same number of *candidate draws*, thereby requiring two or eight times more posterior kernel evaluations in the Warp1 and Warp2 approach, the resulting estimators do not outperform the AdMit-IS estimator. This confirms that the AdMit-IS estimator is clearly the winner. In this example, warping may provide a slight improvement, but appropriately *wrapping* the posterior yields a much larger gain in computational efficiency than the *warping* method.

Until now we have considered the standard deviations of the estimators, when the simulation process is repeated 500 times. In practice, we usually do not compute standard deviations in such a time consuming way. Instead, we estimate the standard deviation by a numerical standard error based on a single simulation run. In the next section we consider

Figure 4.5: Non-linear regression model (4.10): the effect of warping of posterior density kernel




Warp1 (mixture of 2 posterior transformations)

Warp2 (mixture of $8 = 2^3$ posterior transformations)

*Note*: The figure shows a mixture of the posterior density and its 'mirror images'achieved by Warp 1 and Warp 2 transformations.

the reliability of numerical standard errors. The importance of thoroughly evaluating the accuracy of marginal likelihood estimators is also stressed by Frühwirth-Schnatter and Wagner (2008).

Table 4.3: Non-linear regression model (4.10): Marginal likelihood estimation making use of Warp1 or Warp2 transformations in combination with an adaptive Student-$t$ candidate distribution

| st.dev. $10^{10}\cdot$ ML | IS | BS1 | BS2 |
|---|---|---|---|
| Warp1 (100000 posterior kernel evaluations) | 0.1750 | 0.3535 | 0.2250 |
| Warp2 (100000 posterior kernel evaluations) | 0.3097 | 0.5813 | 0.4054 |
| Warp1 (100000 candidate draws) | 0.1250 | 0.2575 | 0.1623 |
| Warp2 (100000 candidate draws) | 0.1182 | 0.2131 | 0.1522 |

*Note*: The table reports standard deviation of 500 estimates of $10^{10}\cdot$ ML from 500 simulation runs.

## 4.5 Numerical standard errors

For the IS estimator, the computation of a numerical standard error (NSE) is particularly straightforward. One divides the standard deviation of the terms $k(\theta^{[l]} \mid y) \,/\, q(\theta^{[l]})$ $(l = 1, \ldots, L)$ by $\sqrt{L}$. However, for the RIS, BS1, BS2 and CJ estimators we make use of the usual *delta rule*. Moreover, the latter four estimators make use of correlated MCMC draws where we need to take into account serial correlation. In this section we will consider three

methods for computing the standard error of a sample mean of such correlated series; that is an estimate of the standard deviation of

$$\hat{g} = \frac{1}{M} \sum_{m=1}^{M} g(\theta^{[m]}) \,, \tag{4.11}$$

where $\{\theta^{[m]}\}_{m=1}^{M}$ is a series of MCMC draws.

The first estimate of the variance $\text{var}(\hat{g})$ that we consider, is the estimate of Newey and West (1987)

$$\widehat{\text{var}}_{\text{NW}}(\hat{g}) = \frac{1}{M} \left[ \hat{\gamma}_0 + 2 \sum_{i=1}^{b} \left( 1 - \frac{i}{b+1} \right) \hat{\gamma}_i \right] , \tag{4.12}$$

where $b$ is a constant that should represent the lag at which the autocorrelation tapers off, $\hat{\gamma}_0$ is the sample variance of the series $\{g(\theta^{[m]})\}_{m=1}^{M}$, and $\hat{\gamma}_i$ is its $i$-th order sample autocovariance. This Newey-West (NW) estimate is used by Chib (1995) and Chib and Jeliazkov (2001), who set $b$ equal to 10 and 40, respectively. We choose a bandwidth of $b = 40$.

The second and third estimate we consider are from Geyer (1992): the initial positive sequence estimator and the initial monotone sequence estimator. These are specialized for reversible Markov chains such as the series of MH draws. Theorem 3.1 of Geyer (1992) states the following. For a stationary, irreducible, reversible Markov chain with autocovariance $\gamma_i$ let $\Gamma_t = \gamma_{2t} + \gamma_{2t+1}$ be the sums of adjacent pairs of autocovariances. Then $\Gamma_t$ is a strictly positive, strictly decreasing, strictly convex function of $t$.

The initial positive sequence estimator (IPSE) estimator is now given by

$$\widehat{\text{var}}_{\text{IPSE}}(\hat{g}) = \frac{1}{M} \left[ \hat{\gamma}_0 + 2 \sum_{t=0}^{2h+1} \hat{\gamma}_t \right] = \frac{1}{M} \left[ -\hat{\gamma}_0 + 2 \sum_{t=0}^{h} \hat{\Gamma}_t \right] , \tag{4.13}$$

where $\hat{\Gamma}_t = \hat{\gamma}_{2t} + \hat{\gamma}_{2t+1}$ and where $h$ is chosen to be the largest integer such that $\hat{\Gamma}_t > 0$ for $t = 1, \ldots, h$.

In the initial monotone sequence estimator (IMSE) the value of $h$ is chosen to be the largest integer such that $\hat{\Gamma}_{t-1} > \hat{\Gamma}_t$ *and* such that $\hat{\Gamma}_t > 0$ for $t = 1, \ldots, h$. Therefore, the resulting estimates satisfy: $\widehat{\text{var}}_{\text{IMSE}}(\hat{g}) \leq \widehat{\text{var}}_{\text{IPSE}}(\hat{g})$. For derivations of NSE's for normalizing constants we refer to Chen *et al.* (2000).

We now inspect the NSE in the example from the previous section. Figure 4.6 shows boxplots, comparing the numerical standard errors to the standard deviations for the three candidate distributions. For the naive Student-$t$ candidate distribution the NSE is often unreliable: huge underestimation of the uncertainty in the marginal likelihood estimator

often occurs. For the adaptive Student-$t$ candidate distribution the NSE is more reliable than in the naive case. However, for all estimators a substantial underestimation of the uncertainty may still occur. The NSE based on the IPSE should be preferred over the NSE from the IMSE and NW formula. For the AdMit candidate distribution the NSE is more reliable than for the other candidates. Especially for the AdMit-IS estimator, the 'winner' of Section 4, the NSE seems reliable. For the BS1, BS2 and CJ estimators, the NSE from the IPSE should again be preferred over the NSE from the IMSE or NW approach. Only for the RIS estimator, which anyway performs poorly in this example, the IMSE provides a NSE that yields a huge overestimation of the uncertainty.

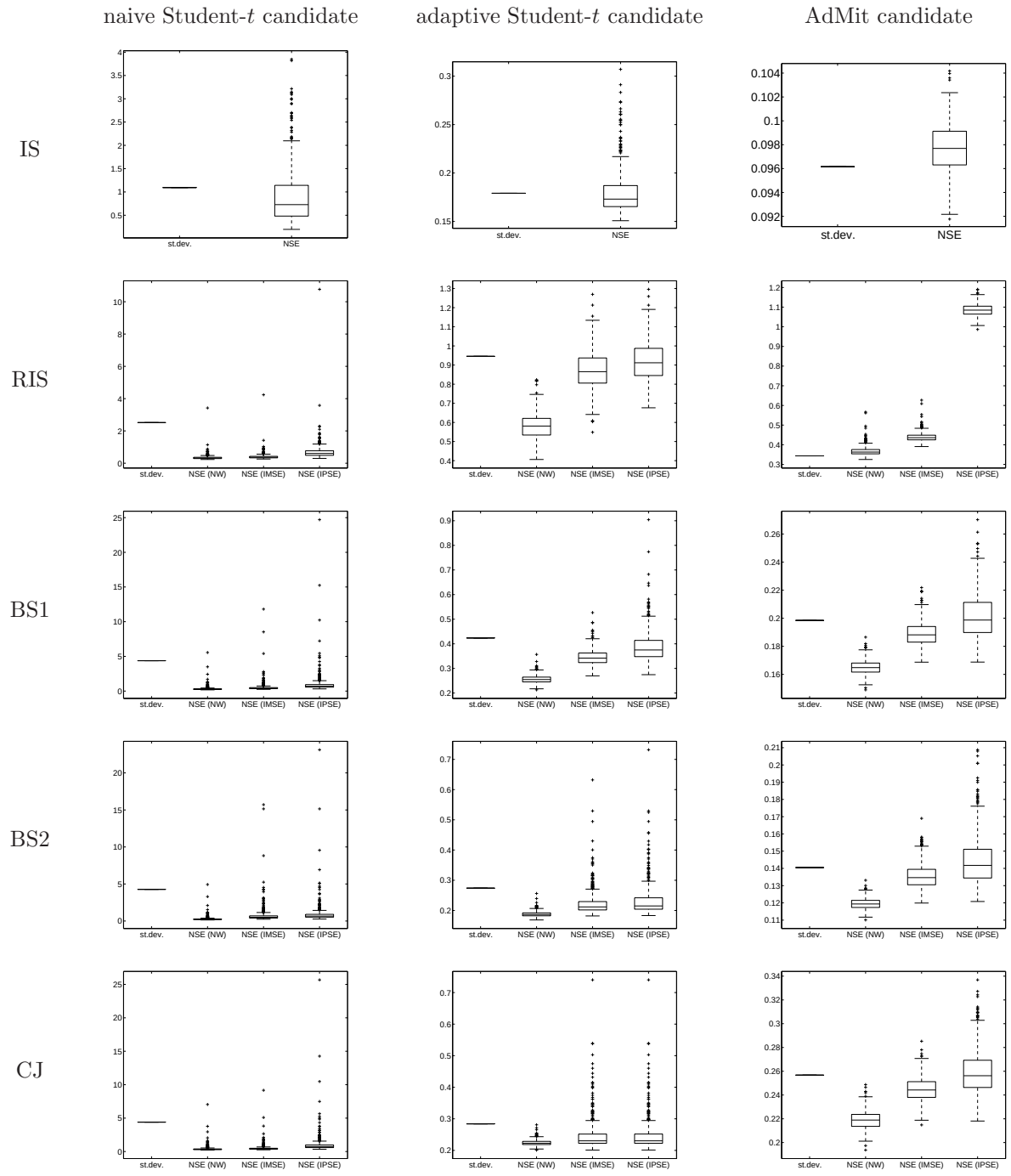
Another way of assessing the performance of the numerical standard errors is to inspect the coverage rate of estimated 90% intervals

$$\left(\hat{p}(y) - 1.645 \cdot \mathrm{NSE}_{\hat{p}(y)},\ \hat{p}(y) + 1.645 \cdot \mathrm{NSE}_{\hat{p}(y)}\right).$$

Table 4.4 gives these coverage rates. In (approximately) 90% of the simulations, the interval should include the true value $p(y)$, whereas the situations with too low or too high intervals should both occur in (about) 5% of the simulations. For the naive candidate distribution, significant deviations from the correct rates can be found for the intervals of all estimators. For the adaptive Student-$t$ candidate, the coverage rates are incorrect for all but the IS estimator. This confirms the unreliable character of the NSE for the naive or adaptive candidate distributions. For the AdMit-IS estimator the coverage rates are correct, whereas for the BS1, BS2 and CJ estimators using AdMit draws only the IPSE and IMSE provide (approximately) correct rates.

We conclude that also in terms of the reliability of the NSE and confidence intervals the AdMit-IS approach performs best. For other AdMit estimators (BS1, BS2 and CJ) the initial monotone sequence estimator of Geyer (1992) provides a useful NSE. For the adaptive (and naive) candidate we find that all three types of NSEs may be (highly) unreliable. The reason for the failure of the NSE based on the Newey-West formula is partly that the 'bandwidth' $b = 40$ is too small a value. Still, also the IPSE and IMSE that automatically adapt the 'bandwidth' to the autocorrelation in the given series of MCMC draws (slightly) fail in case of the naive (and adaptive) candidate distribution. Therefore, the fixed value of $b = 40$ is arguably not always the only reason for its failure.

Figure 4.6: Non-linear regression model (4.10): Boxplots of 500 numerical standard errors (NSE's) for estimates of $10^{10}\cdot$ marginal likelihood based on 100000 candidate draws using different candidate distributions




*Note*: The standard deviation of the 500 marginal likelihood estimates is shown as the horizontal line in the first column. NSE's are computed using the delta rule, where NW, IMSE, IPSE refer to the approach of Newey and West (1987), the initial monotone sequence estimator and the initial positive sequence estimator (Geyer, 1992) for taking into account the serial correlation in the MH draws.

Table 4.4: Non-linear regression model (4.10): Coverage rate of estimated 90% interval for $p(y)$ based on different NSE's (in 500 repetitions)

| | 90% interval from Newey-West NSE | | | 90% interval from IMSE NSE | | | 90% interval from IPSE NSE | | |
|---|---|---|---|---|---|---|---|---|---|
| | too low | ok | too high | too low | ok | too high | too low | ok | too high |
| AdMit candidate | | | | | | | | | |
| IS | 0.056 | 0.902 | 0.042 | 0.056 | 0.902 | 0.042 | 0.056 | 0.902 | 0.042 |
| RIS | 0.002 | 0.730 | 0.268 | 0.002 | 0.836 | 0.162 | 0.000 | 1.000 | 0.000 |
| BS1 | 0.106 | 0.824 | 0.070 | 0.068 | 0.886 | 0.046 | 0.052 | 0.912 | 0.036 |
| BS2 | 0.102 | 0.844 | 0.054 | 0.082 | 0.884 | 0.034 | 0.072 | 0.900 | 0.028 |
| CJ | 0.092 | 0.834 | 0.074 | 0.058 | 0.880 | 0.062 | 0.038 | 0.908 | 0.054 |
| adaptive Student-$t$ candidate | | | | | | | | | |
| IS | 0.052 | 0.902 | 0.046 | 0.052 | 0.902 | 0.046 | 0.052 | 0.902 | 0.046 |
| RIS | 0.440 | 0.312 | 0.248 | 0.412 | 0.360 | 0.228 | 0.338 | 0.532 | 0.130 |
| BS1 | 0.128 | 0.728 | 0.144 | 0.080 | 0.846 | 0.074 | 0.068 | 0.872 | 0.060 |
| BS2 | 0.118 | 0.772 | 0.110 | 0.082 | 0.864 | 0.054 | 0.080 | 0.874 | 0.046 |
| CJ | 0.092 | 0.834 | 0.074 | 0.086 | 0.866 | 0.048 | 0.086 | 0.866 | 0.048 |
| naive Student-$t$ candidate | | | | | | | | | |
| IS | 0.258 | 0.740 | 0.002 | 0.258 | 0.740 | 0.002 | 0.258 | 0.740 | 0.002 |
| RIS | 0.440 | 0.312 | 0.248 | 0.412 | 0.360 | 0.228 | 0.338 | 0.532 | 0.130 |
| BS1 | 0.548 | 0.220 | 0.232 | 0.490 | 0.316 | 0.194 | 0.354 | 0.546 | 0.100 |
| BS2 | 0.578 | 0.172 | 0.250 | 0.450 | 0.416 | 0.134 | 0.368 | 0.536 | 0.096 |
| CJ | 0.518 | 0.266 | 0.216 | 0.484 | 0.314 | 0.202 | 0.342 | 0.564 | 0.094 |

*Note*: Estimated coverage rates are based on different NSE's in 500 repetitions. For the IS estimators there is no serial correlation in the series of draws, so that only one (straightforward) NSE formula is used.

## 4.6 Application 2: mixture GARCH model

In this section we apply our methods in order to estimate the marginal likelihood in a two-component Gaussian mixture GARCH(1,1) model, a model with six parameters. Ausín and Galeano (2007) propose a Griddy-Gibbs sampler for Bayesian estimation of this model, and note that MH algorithms could improve the efficiency of this method despite the drawback of the effort required to calibrate the candidate distribution. We show that given an appropriately tuned candidate density, straightforward IS provides an efficient method for parameter estimation. We extend the study of Ausín and Galeano (2007) by providing an efficient estimation method for the marginal likelihood. The data

consist of 1859 daily closing prices of the SMI, for the period 1/Jul/1991 - 14/Aug/1998. Daily nominal log-returns are expressed in percentages.

A two-component Gaussian mixture GARCH(1, 1) model for the series $y_t$ is defined by

$$\begin{aligned} y_t &= \mu + \sqrt{h_t}\,\varepsilon_t\,, \\ h_t &= \omega + \alpha\,(y_{t-1}-\mu)^2 + \beta h_{t-1}\,, \\ \varepsilon_t &\sim \begin{cases} N(0,\sigma^2) & \text{with probability } \rho \quad , \\ N(0,\sigma^2/\lambda) & \text{with probability } 1-\rho \quad , \end{cases} \end{aligned} \tag{4.14}$$

where $h_t$ is the conditional variance of $y_t$ given previous information $I_{t-1} = \{y_{t-1}, y_{t-2}, \dots\}$, $0 < \lambda < 1$ and $\sigma^2 = 1/(\rho + (1-\rho)/\lambda)$, so that $\mathrm{var}(\varepsilon_t) = 1$. Similar to Ausín and Galeano (2007), we assume that the initial variance $h_0$ is a known constant, $\varepsilon_t \sim$ mixture Gaussian $(\lambda, \rho)$, and the following parameter restrictions hold: $\omega > 0$, $\alpha \geq 0$, $0.5 \leq \rho < 1$, $\beta \geq 0$ and $\alpha + \beta < 1$. Notice that these parameter restrictions ensure positivity of $h_t$, and that there is a higher probability that an observation falls into the state with smaller variance. Following Ausín and Galeano (2007), we specify a flat prior. However, we truncate the domain for $\mu$ and $\omega$ to finite (wide) intervals to have a proper (non-informative) prior: $(\rho, \lambda, \mu, \alpha, \beta, \omega) \in [0.5, 1] \times [0, 1] \times [-1, 1] \times [0, 1] \times [0, 1] \times (0, 1]$ with $\alpha + \beta < 1$.

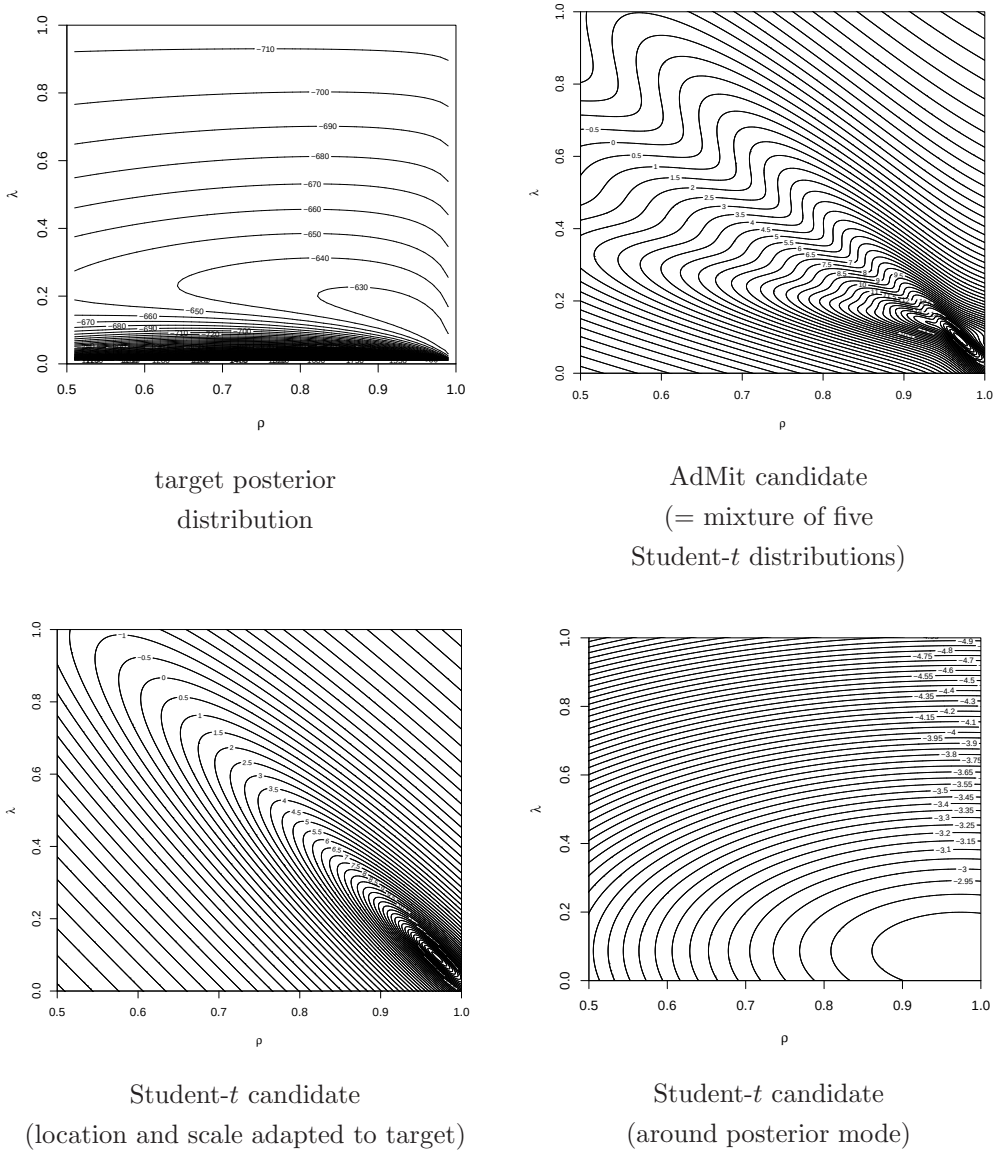
Again, we stress that one can argue that one should not directly use the resulting marginal likelihood for model choice or combination in this case of a noninformative prior. As an alternative, the predictive likelihood (see e.g., Eklund and Karlsson (2007)) can be computed as the marginal likelihood divided by the marginal likelihood for a subset of the data, the so-called training sample.

For the IS and independence chain MH algorithms, we consider three candidate distributions based on Student-$t$ densities with Cauchy tails: the mixture of Student-$t$ distributions resulting from the AdMit procedure, an 'adaptive' Student-$t$ distribution where the mode and scale have been iteratively updated by several IS steps and a 'naive' Student-$t$ distribution around the posterior mode.

Figure 4.7 shows the shapes of the three candidate distributions, together with the conditional posterior density of $(\lambda, \rho)$; parameters $(\omega, \beta, \alpha, \mu)$ are fixed at their posterior mean values. Figure 4.7 illustrates that the AdMit candidate outperforms both adaptive and naive Student-$t$ candidates: the relevant subdomain of the posterior density is wrapped by the AdMit candidate. In particular, the naive Student-$t$ candidate is quite inadequate for wrapping the posterior.

In order to illustrate the local non-identification in the model and the corresponding irregularity in the posterior density, we consider the posterior density of $(1/\lambda, \rho)$. The posterior density for $(1/\lambda, \rho)$ is shown in Figure 4.8, where the other parameters are fixed

Figure 4.7: Mixture GARCH(1,1) model (4.14): Contour plots for (the logarithm of) the conditional posterior density of $(\rho, \lambda)$




target posterior distribution

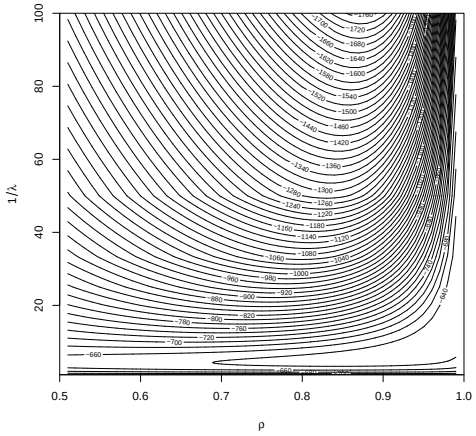
AdMit candidate (= mixture of five Student-$t$ distributions)

Student-$t$ candidate (location and scale adapted to target)

Student-$t$ candidate (around posterior mode)

*Note*: The figures show the conditional posterior density (top left) and candidate density contours for mixture of Student-$t$ (AdMit, top right), adaptive Student-$t$ (bottom left), 'naive' Student-$t$ around the posterior mode (bottom right). Parameters $(\mu, \omega, \alpha, \beta)$ are fixed at the posterior mean values.

Figure 4.8: Mixture GARCH(1,1) model (4.14): Contour plot for (the logarithm of) the conditional posterior density of $(\rho, 1/\lambda)$ given $(\mu, \omega, \alpha, \beta)$ equal to the posterior mean.




*Note*: Parameters $(\mu, \omega, \alpha, \beta)$ are fixed at the posterior mean values.

at posterior means. For $\rho \to 1$, $1/\lambda$ becomes unidentified since the corresponding large variance regime disappears from the model. For $\lambda \to 1$, the conditional variances in both states are identical, hence the mixing probability $\rho$ cannot be identified. This explains why the shapes of the posterior density are far from elliptical, and wisely choosing a candidate that can approximate this non-elliptical shape can provide considerable efficiency gains.

We will now use these three candidate distributions, using 100000 draws for IS. Parameter estimates and NSE's for all cases, together with Ausín and Galeano (2007) estimates, are reported in Table 4.5. Ausín and Galeano (2007) consider log-returns instead of log-returns in percentages, hence their parameter estimates for $\omega$ and $\mu$ have been multiplied by 10000 and 100, respectively. Estimates under the adaptive and AdMit approaches are similar to Ausín and Galeano (2007). Further, we find two main results. First, the naive Student-$t$ density has the worst performance among the candidate densities we compare. The estimates are substantially biased for 100000 draws. This shows that IS fails to provide accurate results using a poor candidate density. Hence, in the rest of our analysis, we compare the performances of only the adaptive Student-$t$ and AdMit candidates. Second, the AdMit candidate clearly outperforms the adaptive Student-$t$ candidate: NSE's obtained using the AdMit candidate are much smaller than those obtained using the adaptive Student-$t$ candidate. Results from the independence chain MH algorithm were similar to the IS results. Only for MH the NSE's were somewhat larger than for IS. According to NSE's, AdMit-IS has the best performance.

We now analyze the sensitivity of the marginal likelihood estimators in Section 4.2 to the choices of the candidate distribution. We consider the IS, BS1, BS2 and CJ estimators. We discard the RIS estimator, since this already gave particularly bad results in Section 4.4. Table 4.6 reports marginal likelihood estimates and the respective NSE's. For the BS1, BS2 and CJ estimators, NSE's are calculated by IPSE, as it was shown to produce the most accurate estimator in Section 5. First, the AdMit candidate yields more precise estimates than the adaptive Student-$t$ candidate. This points out the importance of wisely specifying an appropriate candidate both for IS and independence chain MH algorithms. In particular for the IS estimator, this gain in efficiency is substantial: NSE's for IS estimates decrease at least 50% when the posterior density is wrapped nicely by the AdMit candidate.

We now consider marginal likelihood estimates resulting from aiming at a 'warped' version of the posterior kernel using the Warp1 method. We do not consider the Warp2 method because of its computational cost: for this six-dimensional posterior each draw would require $2^6 = 64$ posterior kernel evaluations. For the adaptive Student-$t$ candidate, the NSE's are reported in the top panel of Table 4.6. Considering BS1, BS2 and CJ estimators, warping the posterior density leads to smaller NSE's only when the number of candidate draws is kept constant. Hence this gain in efficiency is related both to warping the posterior and the increased number of posterior kernel evaluations. For the IS estimator on the other hand, the Warp1 transformation does provide efficiency gains, even when the number of kernel evaluations is the same as for the 'unwarped' counterpart. Finally, the IS estimate using the adaptive Student-$t$ density in combination with the Warp1 method is less precise than the 'unwarped' IS estimate using the AdMit candidate, which is shown in the bottom panel of Table 4.6. Hence the gain from 'warping'

Table 4.5: Mixture GARCH(1,1) model (4.14): Estimated posterior means and corresponding NSE's, obtained by IS using different candidate densities.

| | Griddy-Gibbs estimates of Ausín and Galeano (2007) | IS estimates | | | | | |
|---|---|---|---|---|---|---|---|
| | | AdMit | | adaptive | | naive | |
| | mean | mean | NSE ·100 | mean | NSE ·100 | mean | NSE ·100 |
| $\omega$ | 0.11 | 0.08 | 0.02 | 0.07 | 0.08 | 0.33 | 2.20 |
| $\lambda$ | 0.13 | 0.12 | 0.03 | 0.12 | 0.15 | 0.27 | 2.20 |
| $\beta$ | 0.74 | 0.80 | 0.03 | 0.80 | 0.12 | 0.45 | 2.05 |
| $\alpha$ | 0.15 | 0.13 | 0.02 | 0.13 | 0.06 | 0.20 | 1.68 |
| $\rho$ | 0.92 | 0.95 | 0.02 | 0.95 | 0.16 | 0.56 | 0.13 |
| $\mu$ | 0.11 | 0.11 | 0.01 | 0.11 | 0.03 | 0.08 | 0.45 |

the posterior is smaller than that of 'wrapping' the posterior with a mixture of Student-$t$ distributions.

Table 4.6: Mixture GARCH(1,1) model (4.14): marginal likelihood estimates and corresponding NSE's for IS and independence chain MH based on adaptive Student-$t$ and AdMit candidates, and NSE's for IS and MH in combination with the Warp1 method

| adaptive | | | | | |
|---|---|---|---|---|---|
| | | IS | BS1 | BS2 | CJ |
| $10^{280}\cdot$ estimate | 'unwarped' | 3.96 | 3.95 | 3.90 | 3.96 |
| $10^{282}\cdot$ NSE | 'unwarped' | 11.12 | 11.88 | 12.07 | 10.95 |
| | Warp1$^*$ | 4.97 | 12.19 | 12.28 | 14.11 |
| | Warp1$^\dagger$ | 4.20 | 8.77 | 8.52 | 10.11 |
| AdMit | | | | | |
| | | IS | BS1 | BS2 | CJ |
| $10^{280}\cdot$ estimate | 'unwarped' | 4.06 | 4.04 | 4.02 | 4.06 |
| $10^{282}\cdot$ NSE | 'unwarped' | 2.95 | 4.49 | 3.70 | 5.49 |
| | Warp1$^*$ | 2.48 | 7.03 | 5.90 | 8.62 |
| | Warp1$^\dagger$ | 2.46 | 4.92 | 4.13 | 6.05 |

*Note*: Warp1$^*$ refers to $10^5$ posterior kernel evaluations (i.e. $0.5 \cdot 10^5$ candidate draws), whereas Warp1$^\dagger$ refers to $10^5$ candidate draws (i.e. $2 \cdot 10^5$ posterior kernel evaluations).

Until now, we have considered 'warping' and 'wrapping' the posterior kernel separately. A natural extension is to combine these methods. Hence we analyze the changes in NSE's when the posterior kernel is both 'warped' and 'wrapped'. The bottom panel in Table 4.6 shows the NSE's estimated by IS, BS1, BS2 and CJ algorithms making use of the AdMit candidate and the Warp1 method, together with the 'unwarped' counterparts. Warp1 transformation does not lead to efficiency gains in BS1, BS2 and CJ estimators. The Warp1 transformation leads to a small decrease in NSE for the IS algorithm, given the same number of posterior kernel evaluations. Therefore we conclude that 'warping' and 'wrapping' the posterior at the same time increases the efficiency of the IS algorithm, but most of this efficiency gain stems from constructing an appropriate candidate density.

We make a final comparison for the NSE's for the BS1, BS2 and CJ estimators, making use of the Newey-West estimator, IPSE and IMSE of Section 4.5. Table 4.7 reports NW, IPSE and IMSE standard errors under adaptive and AdMit candidates, without or with a Warp1 transformation ($10^5$ candidate draws). For the NW estimator, we choose a bandwidth of 40. NSE's using all estimators are still larger than those of IS using the AdMit candidate. Hence the victory of AdMit-IS is not related to the choice of the NSE

estimators. Furthermore, NW estimates are quite different from IMSE and IPSE values. Notice that this result is in line with Section 4.5 where we show that the NW estimator is less reliable than the IPSE and IMSE.

Table 4.7: Mixture GARCH(1,1) model (4.14): numerical standard errors

| adaptive | | | | adaptive, Warp1 combination | | | |
|---|---|---|---|---|---|---|---|
| $10^{282} \cdot$ NSE | | | | $10^{282} \cdot$ NSE | | | |
| | BS1 | BS2 | CJ | | BS1 | BS2 | CJ |
| NW | 7.20 | 6.07 | 7.30 | NW | 4.39 | 3.65 | 4.23 |
| IPSE | 11.88 | 12.07 | 10.95 | IPSE | 8.77 | 8.52 | 10.11 |
| IMSE | 11.88 | 12.07 | 10.95 | IMSE | 8.77 | 8.52 | 10.11 |
| AdMit | | | | AdMit, Warp1 combination | | | |
| $10^{282} \cdot$ NSE | | | | $10^{282} \cdot$ NSE | | | |
| | BS1 | BS2 | CJ | | BS1 | BS2 | CJ |
| NW | 5.08 | 3.83 | 5.43 | NW | 4.13 | 3.88 | 4.23 |
| IPSE | 4.49 | 3.70 | 5.49 | IPSE | 4.92 | 4.13 | 6.05 |
| IMSE | 4.32 | 3.70 | 5.49 | IMSE | 3.65 | 4.13 | 4.38 |

*Note*: The table reports numerical standard errors based on the Newey-West, IPSE and IMSE methods for independence chain MH sampler using adaptive and AdMit candidates, without or with Warp1 transformation ($10^5$ candidate draws)

## 4.7 Conclusion

We have considered two very different model structures (for data sets with different sample sizes), a non-linear regression model (for a very small data set) and a mixture GARCH model (for a large data set), with clearly different non-elliptical posterior shapes. Still, we obtain roughly the same findings. Given a suitable candidate distribution, which can be obtained by the AdMit method, the IS algorithm delivers a computationally efficient marginal likelihood estimator (and a reliable, easily computed numerical standard error), which outperforms the RIS, BS1, BS2 and CJ estimators. Warping the posterior density can lead to a further gain in efficiency, but it is more important that the posterior kernel is appropriately wrapped by the (AdMit) candidate distribution than that is warped. Moreover, warping requires evaluations of the warped posterior density kernel which are only used for marginal likelihood estimation. For the straightforward IS estimator of the marginal likelihood only computations are required that are typically already performed

for parameter estimation or forecasting; no *extra* computations are required for marginal likelihood estimation.

If one uses a marginal likelihood estimator on the basis of serially correlated MCMC draws, the IPSE of Geyer (1992) performs best among the considered methods for computing numerical standard errors.

One can argue that one should not directly use the marginal likelihood for model choice or model combination when one uses a noninformative prior. An alternative is the predictive likelihood (see e.g., Eklund and Karlsson (2007)), which may be computed as the marginal likelihood for the whole data set divided by the marginal likelihood for a subset of the data, the so-called training sample. For the training sample, the posterior shapes will typically be even more non-elliptical than for the whole data set. Therefore the use of an appropriate (AdMit) candidate distribution may be even more important if one bases model choice or combination on predictive likelihoods than if one compares marginal likelihoods.

In further research, we intend to consider different empirical applications. We have considered models with three and six parameters. We intend to investigate whether AdMit-IS remains the best method for models with many more parameters. We will further compare the performance of different types of bridge sampling estimators with the approach of Chib (1995) in cases of non-elliptical posteriors where the Gibbs sampler is applicable, such as the change-point models considered by Bauwens and Rombouts (2010). We will also consider the quality of the estimators when these are applied in combination with the radial-based transformation of Bauwens *et al.* (2004). Another possibility is to consider the path sampling method of Gelman and Meng (1998), which extends the bridge sampling approach. Finally, we show the IS estimator provides gains in efficiency compared to the BS1, BS2, CJ and RIS estimators given a certain candidate density, ignoring the computing time required for constructing the AdMit draws. Typically, the amount of time required to construct the AdMit candidate is much smaller than the time required to evaluate the posterior kernel, but a formal comparison of the computing time for constructing the AdMit candidate is left for future research.

# Chapter 5

# Measuring Returns to Education: Bayesian Analysis Using Weak or Possibly Endogenous Instrumental Variables

*Chapter 5 is based on Baştürk, Hoogerheide, and Van Dijk (2010).*

## 5.1 Introduction

A simple regression of earned income on years of education in order to measure the education-income effect neglects important issues. Although higher education levels are expected to increase an individual's earnings, the income-education relationship can be subject to omitted variables such as individuals' intellectual capabilities, measurement errors in reported earnings, or simultaneity as individuals can determine the amount of education they receive judging the possible monetary returns (Angrist *et al.*, 1996).

A more formal way to state the issue is the following: The analysis of causal effects of economic variables brings difficulties when the explanatory variable is itself not exogenous. This *endogeneity* problem might arise as a result of the abovementioned phenomena of omitted variables, measurement errors, or simultaneity between the variables of interest (Goldberger, 1972). The standard treatment of such relationships is the use of Instru-

mental Variable (IV) models[1]. IV models rely on a set of instruments (proxies) that can be used to make inference about the endogenous variable (see e.g. Sargan (1958); Bowden and Turkington (1990)).

One of the drawbacks of IV models is the loss of precision, stemming from the use of only an approximation to the explanatory variable. This issue is relatively more severe in case of *weak instruments*, where the instrument has a small amount of information regarding the endogenous variable. Furthermore, the *validity*, i.e. the exogeneity of instruments are often questioned. If there are more instruments than (possibly) endogenous variables, this validity assumption can be tested formally (Sargan, 1958), however in most cases researchers use informal economic reasoning. Finally, similar to other models, the linearity of the relationships can be questioned. Hence the use of IV models naturally brings the necessity to judge the proposed instruments', or generally, the model's performance.

The main purpose of this chapter is to define a general approach to assess alternative models' performance in analyzing the income-education relationship, and in general in IV models. We analyze the income-education relationship in two datasets, which provide different instruments for education levels. Note that the problems associated with IV models often depend on the available instrument, hence we briefly discuss possible problems for the two datasets we consider.

We first consider Angrist and Krueger (1991) data on income and education, which includes quarter of birth as instruments[2]. This dataset has been analyzed in several studies, providing a common result that these instruments are quite weak in explaining education, at least in parts of the data (Bound *et al.*, 1995; Hoogerheide and Van Dijk, 2006). Second, we analyze the data from the German Socio-Economic Panel Survey (SOEP), where the available instrument is individual's father's education. Unlike instruments based on quarter of birth, father's education (and other family background variables), are shown to be quite strong instruments (Parker and Van Praag, 2006). Despite this advantage, several studies show that family background variables may have a direct effect on individuals' earnings, and hence the *exact* exogeneity of these instruments is questioned (Trostel

[1]We note that there is a difference in terminology between IV *models* and IV *estimation technique*. We prefer to use the term 'IV model' to denote the multi-equation model, an incomplete simultaneous equations model, in which the regressor in the key equation of interest can be endogenous, and this problem is treated using instrumental variables.

[2]Angrist and Krueger (1991) quarter of birth as instruments based on schooling laws on compulsory education, which allow individuals to end their education only at a certain age.

*et al.*, 2002; Psacharopoulos and Patrinos, 2004)[3]. Given the distinct nature of these data on income and education, we define the alternative models accordingly.

For assessing model performance, we rely on Bayesian methods, as they provide general probabilistic tools to explicitly account for parameter and model uncertainty. Regarding the latter, Bayesian treatment of model uncertainty relies on evaluating posterior model probabilities and the degree to which alternative models are suitable to the data (Clyde and George, 2004).

In the existence of model uncertainty, it is often very difficult or impossible to determine the true model structure, that is which model is correct (for an extensive discussion on model uncertainty and adequacy, see Geweke (2010)). For example, in the income-education analysis, one can define two models, one treating education as an endogenous explanatory variable (i.e. the IV model), and the second model treating education as an exogenous explanatory variable (i.e. the simple linear regression model). To choose one of these competing models can be problematic. In such situations, Bayesian Model Averaging (BMA) provides a theoretical motivation (Bates and Granger, 1969), and has been applied mainly in the forecasting literature by success (Min and Zellner, 1993; Eklund and Karlsson, 2007; Geweke and Amisano, 2010). The advantage of BMA over selecting one of the alternative models is to account for the uncertainty inherent in the model selection process by averaging over many different competing models. It appears more plausible and effective to weight the evidence of the alternative model structures and to take a weighted average of the two or more structures to predict the effect of the explanatory variable on the dependent variable.

Analyzing model performance (hence the implementation of BMA) brings challenges when one does not have strong prior information about the data (Bartlett, 1957). Following recent studies analyzing the two datasets we consider in Bayesian context (Hoogerheide and Van Dijk, 2006; Hoogerheide *et al.*, 2010), we adopt uninformative priors for the income-education analysis and show that the problems associated with uninformative priors in the IV models can be avoided using a *predictive likelihood* approach, recently studied by Eklund and Karlsson (2007).

Predictive likelihoods can be interpreted as follows: the posterior results for a subsample of the data, the *training sample*, is treated as the prior for the analysis of the remaining observations, the *hold-out sample*. Hence the effect of the originally noninformative prior is eliminated (see Laud and Ibrahim (1995) for a discussion on predictive likelihood methods in the Bayesian context).

[3]For the Angrist and Krueger (1991) data, endogeneity of instruments is arguably not a big problem as the instruments, quarter of birth dummies, are not expected to have a direct effect on income.

We apply this technique to simulated datasets with differing degrees of endogeneity problem and instrument strength, and two different datasets on income-education relationship. We show that this method can be used to weight the evidence of different models and to assess some of the highly debated issues in IV models, such as the trade-off between choosing the correct model and precision, and differing effects of endogenous variable across subsamples of data.

Our empirical results show that the income-education relationship in the US states and divisions, using quarter of birth of individuals, is subject to significant heterogeneity in terms of the degree of endogeneity in education levels, as well as the strength of instruments. For these data, we conclude that the precision losses from employing an IV model can be quite severe. We rather propose a model averaging approach to infer returns to education, for which the model weights take into account the degree of endogeneity and the instrument strength.

Second, for the income-education relationship for the German SOEP data, we extend Hoogerheide *et al.* (2010) by considering different marginal effects of education on income across male and female respondents, and analyze whether returns to education are constant during the time period considered. These extensions are motivated by several studies suggesting lower returns to educations for men, which in turn mitigate the gender wage discrimination (see Psacharopoulos and Patrinos (2004) among others), and changing effects of education on earnings, for example as a result of technological improvements (see e.g. Acemoglu (1998)). We show that employing predictive likelihoods for model comparison provides a useful tool to judge whether the extensions we propose improve over the benchmark model with constant returns to education across individuals and over time. We find gender discrimination in earnings decreases with increasing education levels and returns to education are higher for the recent time period. We further show that the findings are relatively robust to possible endogeneity of the instruments.

In sum, the methodological contribution of this study is the following: We show that the evidence for alternative IV models can be assessed using predictive likelihoods instead of marginal likelihoods, and BMA using predictive likelihoods provides a tool for efficient measurement of the marginal effects of regressors. On the other hand, the empirical contributions of this study are as follows: First, for the US data on the income-education relationship, we address the issues of weak instruments and degree of endogeneity, and show that for this dataset, the use of an IV model alone is not supported. A weighted average using an IV model and a simple regression model without instruments provides a more appropriate method to analyze the effects of education on the income. Second, for the SOEP data on income-education relationship, we address the issue of possible

endogeneity of the instruments, together with possible heterogenous effects of education on income.

The remainder of this chapter is organized as follows: Section 5.2 presents the IV model, and the predictive likelihood approach. Section 5.3 applies the proposed method to assess the degree of endogeneity and instrument strength to simulated datasets with differing degrees of endogeneity and instrument strength, and to the Angrist and Krueger (1991) data accordingly. Section 5.4 applies the predictive likelihood approach to the SOEP data and considers possible heterogeneities in the effects of education on income, allowing for possible endogeneity of instruments. Section 5.5 concludes.

## 5.2 Standard IV model and the predictive likelihood approach

The IV model with one explanatory endogenous variable and $p$ instruments is defined by (Bowden and Turkington, 1990):

$$y_1 = y_2\beta + \epsilon, \tag{5.1}$$

$$y_2 = x\Pi + \nu, \tag{5.2}$$

where $y_1$ is the $N \times 1$ vector of the dependent variable, $y_2$ is the $N \times 1$ vector of the endogenous explanatory variable, $x$ is the $N \times p$ matrix of instruments and all variables are demeaned i.e. $x$ does not include a constant term. For $i \in 1, \ldots, N$, $\begin{pmatrix} \epsilon_i & \nu_i \end{pmatrix}' \sim NID\left(0, \Sigma\right)$ where $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$ is a positive definite symmetric (PDS) matrix, and the disturbances are uncorrelated across observations $i$.

The matrix notation for the model is given by:

$$Y = x\Pi B + \upsilon, \tag{5.3}$$

where $Y = \begin{bmatrix} y_1 & y_2 \end{bmatrix}$ is the $N \times 2$ matrix of endogenous variables, $x$ is the $N \times p$ matrix of instruments, $\Pi$ is $p \times 1$ vector of parameters, $B = \begin{bmatrix} \beta & 1 \end{bmatrix}$, where $\beta$ is a scalar, $\upsilon = \begin{bmatrix} \epsilon & \nu \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix}$ is the $N \times 2$ matrix of error terms with $\text{vec}\left(\upsilon\right) \sim N\left(0, \Omega \otimes I_N\right)$, where $\Omega = \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix}' \Sigma \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix}$.

The likelihood function for the model in (5.3) is:

$$p\left(Y \mid x, B, \Pi, \Omega\right) \propto \left|\Omega\right|^{-\frac{N}{2}} \exp\left(-\frac{1}{2} tr\left(\Omega^{-1}\left(Y - x\Pi B\right)'\left(Y - x\Pi B\right)\right)\right). \tag{5.4}$$

Next, we briefly discuss some of the important issues in estimating the IV model.

*Instruments' validity:* The instruments $x$ are said to be valid only if the condition $\text{cov}(x_i, \epsilon_i) = 0$ is satisfied. Possible endogeneity of instruments has been recently analyzed by Conley *et al.* (2008), allowing for a certain amount of variation in this condition.

*Instrument strength:* In the IV model, strength of the instruments is reflected by parameter $\Pi$. In the extreme case when $\Pi = 0$, the instruments have no explanatory power for the possibly endogenous explanatory variable $y_2$. This case is referred to as the case of *irrelevant* instruments.

*The degree of endogeneity in the data:* Assuming valid instruments, one can immediately see that the endogeneity problem in the model simply arises from the correlation between the structural errors:

$$\text{cov}(y_{2i}, \epsilon_i) = \text{cov}(x_i\Pi + \nu_i, \epsilon_i) \tag{5.5}$$

$$= \text{cov}(\nu_i, \epsilon_i). \tag{5.6}$$

Therefore the instruments are only necessary if the correlation between the structural errors, $\rho \equiv \text{corr}(\epsilon_i, \nu_i) = \sigma_{12}/(\sigma_{11}\sigma_{22})^{1/2}$ is different from zero. It is important to note that when there is no endogeneity issue, standard estimation methods using a simple linear model instead of the IV model suffices.

*The existence and the precision of the posterior of $\beta$ and $\Pi$:* Under flat priors, the joint posterior density kernel of $(\Pi, \beta)$ has a ridge at $\Pi = 0$ and the marginal posterior density kernel of $\Pi$ has an asymptote at $\Pi = 0$. Further, in the case of exact identification (with as many instruments as explanatory endogenous variables), the posterior of $\Pi$ and $\beta$ is not a proper density (Dréze, 1976, 1977; Kleibergen and Van Dijk, 1994a).

Kleibergen and Van Dijk (1998); Hoogerheide *et al.* (2007a) show that Jeffreys' prior eliminates some of these undesired properties. However, the precision in the posterior of $\beta$ under Jeffreys priors also depends on the instrument strength. Kleibergen and Van Dijk (1998) show that the marginal posterior density of $\beta$ under Jeffreys prior is a ratio of two t-densities, where the denominator depends on the explanatory power of the instruments. Intuitively, the posterior density is wider when one has weak instruments. In case of no endogeneity, this decrease in precision can be avoided by using a model without instruments, hence assessing the degree of endogeneity is important especially when instruments are rather weak in explaining the endogenous explanatory variable.

We next illustrate predictive model probabilities and how they can be employed to assess alternative IV model structures.

**Predictive model probabilities and the Savage Dickey Density Ratio:**
Define two models $M_0$ and $M_1$, where $M_0$ is a nested model compared to $M_1$. Further, the nested model corresponds to $M_1$ with a parameter restriction: $\eta = 0$.

The posterior odds ratio, $K_{01}$ for $M_0$ is the product of the Bayes factor and the prior odds ratio:

$$K_{01} = \frac{p\left(Y \mid M_0\right)}{p\left(Y \mid M_1\right)} \times \frac{p\left(M_0\right)}{p\left(M_1\right)}, \tag{5.7}$$

where $Y$ is the observed data, and the prior model probabilities $\left(p\left(M_1\right), p\left(M_0\right)\right) \in (0,1) \times (0,1)$ and $p\left(M_1\right) + p\left(M_0\right) = 1$.

It is often difficult to compute $K_{01}$ since the marginal likelihoods are given by the following integrals:

$$p\left(Y \mid M_0\right) = \int_{\theta_{-\eta}} \ell\left(\eta = 0, \theta_{-\eta}\right) p_0\left(\theta_{-\eta}\right) d\left(\theta_{-\eta}\right) \tag{5.8}$$

$$p\left(Y \mid M_1\right) = \int_{\theta_{-\eta},\eta} \ell\left(\eta, \theta_{-\eta}\right) p\left(\eta, \theta_{-\eta}\right) d\left(\eta\right) d\left(\theta_{-\eta}\right), \tag{5.9}$$

where $\ell\left(\theta\right)$ is the likelihood function given parameters $\theta$ and $\theta_{-\eta}$ are the model parameters apart from $\eta$.

In order to calculate model probabilities, we make use of the Savage-Dickey Density Ratio (SDDR). Dickey (1971) shows that the Bayes factor can be calculated using a single model if the alternative models are nested and the prior densities satisfy the condition that the prior for $\theta_{-\eta}$ in the restricted model $M_0$ equals the conditional prior for $\theta_{-\eta}$ given $\eta = 0$ in the model $M_1$, i.e. $p_1\left(\theta_{-\eta} \mid \eta = 0\right) = p_0\left(\theta_{-\eta}\right)$. In this case, (5.7) becomes:

$$K_{01} = \frac{p(\eta = 0 \mid Y, M_1)}{p(\eta = 0 \mid M_1)} \times \frac{p\left(M_0\right)}{p\left(M_1\right)}, \tag{5.10}$$

where $p(\eta \mid Y) = \int p(\eta, \theta_{-\eta} \mid Y) d\theta_{-\eta}$ and $p(\eta) = \int p(\eta, \theta_{-\eta}) d\theta_{-\eta}$[4].

One important consideration in model comparison is the effect of relatively uninformative priors. Choosing a prior $p(\eta, \theta_{-\eta})$ diffuse enough compared to $p(\theta_{-\eta})$, posterior odds ratio in (5.7) becomes larger independent of the data. Hence if we consider non-informative priors, the most restrictive model will typically be favored. This phenomenon is called Bartlett's paradox (Bartlett, 1957). Specifically, the prior $p\left(\eta \mid \theta_{-\eta}\right)$ must be proper for the Bayes factor to be well defined.

If one does not have informative priors for the model at hand, model probabilities can be computed using *predictive likelihoods* instead of the conventional *marginal likelihoods*. Eklund and Karlsson (2007) show that the sensitivity of model probabilities to the prior

[4]As a generalization, Verdinelli and Wasserman (1995) show that $K_{01}$ is equal to the Savage-Dickey density ratio in (5.10) times a correction factor when the prior condition fails.

choice can be handled using predictive likelihoods and summarize alternative ways to calculate the predictive likelihood.

A predictive likelihood for the model $M_1$ is computed by splitting the dataset $Y = (y_1, \ldots, y_N)$ into a training sample $y^* = (y_1, \ldots, y_m)$ and a hold-out sample $\tilde{y} = (y_{m+1}, \ldots, y_N)$. The predictive likelihood is given by:

$$p(\tilde{y} \mid y^*, M_1) = \int p(\tilde{y} \mid \theta_1, y^*, M_1) p(\theta_1 \mid y^*, M_1) d\theta_1, \tag{5.11}$$

where $\theta_1$ are the model parameters for model $M_1$. Notice that equation (5.11) corresponds to the marginal likelihood for the hold-out sample $\tilde{y}$ times the exact posterior density after observing $y^*$ as the prior. Therefore, model probabilities using this predictive measure does not provide the exact posterior model probability given the data, but rather a *predictive* model probability[5]. The exact posterior density $p(\theta_1 \mid y^*, M_1)$ is obtained by Bayes' rule:

$$p(\theta_1 \mid y^*, M_1) = \frac{p(y^* \mid \theta_1, M_1) p(\theta_1 \mid M_1)}{p(y^* \mid M_1)} \tag{5.12}$$

$$= \frac{p(y^* \mid \theta_1, M_1) p(\theta_1 \mid M_1)}{\int p(y^* \mid \theta_1, M_1) p(\theta_1 \mid M_1) d\theta_1}. \tag{5.13}$$

Substituting (5.13) into (5.11) leads to:

$$p(\tilde{y} \mid y^*, M_1) = \frac{\int p(\tilde{y} \mid \theta_1, y^*, M_1) p(y^* \mid \theta_1, M_1) p(\theta_1 \mid M_1) d\theta_1}{\int p(y^* \mid \theta_1, M_1) p(\theta_1 \mid M_1) d\theta_1}, \tag{5.14}$$

$$= \frac{\int p(y \mid \theta_1, M_1) p(\theta_1 \mid M_1) d\theta_1}{\int p(y^* \mid \theta_1, M_1) p(\theta_1 \mid M_1) d\theta_1}. \tag{5.15}$$

In case of predictive likelihoods, model probabilities are again calculated from the posterior odds ratio:

$$\frac{p(M_0 \mid y)}{p(M_1 \mid y)} = \frac{p(\tilde{y} \mid y^*, M_0)}{p(\tilde{y} \mid y^*, M_1)} \frac{p(M_0)}{p(M_1)}. \tag{5.16}$$

Combining the predictive likelihood formula in (5.16) and SDDR in (5.10), posterior odds ratio becomes:

$$K_{01} = \frac{p\left(M_0 \mid \tilde{y}, y^*\right)}{p\left(M_1 \mid \tilde{y}, y^*\right)} \tag{5.17}$$

$$= \frac{p(\eta = 0 \mid \tilde{y}, y^*, M_1)}{p(\eta = 0 \mid y^*, M_1)} \times \frac{p\left(M_0\right)}{p\left(M_1\right)}, \tag{5.18}$$

where $p(\eta \mid \tilde{y}, y^*) = \int p(\eta, \theta_{-\eta} \mid \tilde{y}, y^*) d\theta_{-\eta}$ and $p(\eta \mid y^*) = \int p(\eta, \theta_{-\eta} \mid y^*) d\theta_{-\eta}$ are the *exact* marginal posterior densities using the full data, and the training sample, respectively.

[5]Equivalently, the term *predictive model weights* is used in the literature, see e.g. Eklund and Karlsson (2007).

A final point concerning the calculation of predictive likelihoods is the size of the training sample. More stable results are achieved as the training sample size decreases, but the training sample should be large enough to provide a proper density given the originally diffuse/uninformative prior of parameters. Different training sample sizes have been proposed in the literature (see Gelfand and Dey (1994) for an overview of the forms of predictive likelihood under different training sample choices). More recently, Eklund and Karlsson (2007) suggests that the training sample should consist of around 20% of the data. Our choice of training sample size is based on this finding, however, we experiment with variations in the sample size considering training samples consisting of 5%, 10% and 25% of the observations.

**Model averaging:**
Given the posterior odds ratio, it is possible to weight the evidence of alternative models using Bayesian Model Averaging (BMA). We consider the effect of model uncertainty on the estimation of the parameter $\beta$, as this parameter is the main focus in most cases. The information about $\beta$ is summarized by the following posterior (Koop, 2003, ch. 11):

$$p(\beta \mid Y) = p(\beta \mid Y, M_0)\, p(M_0 \mid Y) + p(\beta \mid Y, M_1)\, p(M_1 \mid Y). \tag{5.19}$$

Furthermore, functions of parameters, i.e. $g(\beta)$ in the IV model are estimated by:

$$E[g(\beta \mid Y)] = E[g(\beta \mid Y, M_0)]p(M_0 \mid Y) + E[g(\beta \mid Y, M_1)]p(M_1 \mid Y). \tag{5.20}$$

Hence both models under consideration should be estimated, and the inference on parameters is simply the weighted average of the results in both models. The weights in averaging the results are the predictive model probabilities.

## 5.3 Assessing the degrees of endogeneity and instrument strength in IV models

In this section we apply the predictive likelihood approach to weight the evidence of two alternative models: The first model under consideration is the IV model, that takes into account the endogeneity problem in the data at the expense of possible efficiency losses. The second model we consider is the *nested* model which assumes that there is no endogeneity problem in the data. This nested model corresponds the IV model in (5.1) and (5.2), with a parameter restriction $\rho = \sigma_{12}/\left(\sigma_{11}\sigma_{22}\right)^{1/2} = 0$.

Regarding the predictive model probabilities, we make use of the methods in Section 5.2, where calculation of predictive model probabilities is based on estimating the

*general model*, which is the IV model in this case. Estimation of the *nested* model without endogeneity problem is straightforward: one can simply use the linear model in (5.1) to infer $p(\beta \mid Y, M_0)$.

The applications in this section focus on simulated data with different degrees of endogeneity and instrument strength, and Angrist and Krueger (1991) data on the income-education relationship. In the latter dataset, Hoogerheide *et al.* (2007a) show that this dataset suffers from weak instruments, and that some of the problems associated with weak instruments can be avoided using Jeffreys prior. Therefore in this section we consider the IV model under the Jeffreys prior.

For the IV model in (5.1) and (5.2), Jeffreys prior is given by:

$$p(\beta, \Pi, \Sigma) \propto \sigma_{11}^{(p-1)/2} \left|\Pi' X' X \Pi\right| \left|\Sigma\right|^{-\frac{p+3}{2}}. \tag{5.21}$$

The joint posterior is:

$$\begin{aligned} p(\beta, \Pi, \Sigma \mid Y, x) &\propto p(\Pi, \beta, \Sigma)\, p(Y \mid x, \beta, \Pi, \Sigma) && (5.22) \\ &\propto \sigma_{11}^{(p-1)/2} \left|\Pi' X' X \Pi\right| \left|\Sigma\right|^{-\frac{p+3}{2}} \left|\Sigma\right|^{-\frac{N}{2}} \\ &\quad \times \exp\left(-\frac{1}{2} tr\left(\Sigma^{-1} (Y - X\Pi B)' (Y - X\Pi B)\right)\right). && (5.23) \end{aligned}$$

Notice that conditional on a set of coefficients, the posterior density in (5.23) is rather standard, and it is straightforward to make inference relying on this density. However, the last term in (5.23) includes the product of parameters $\beta$ and $\Pi$, and this issue is not easy to treat (Kleibergen and Van Dijk, 1994b, 1998; Hoogerheide *et al.*, 2007c). Therefore we explain the proposed posterior simulator in detail.

**Posterior simulator:**

For the parameter restriction we consider, $\rho = 0$, the computation of the posterior odds ratio in (5.18) boils down to estimating the marginal posterior densities for the whole sample and the training sample, $p(\rho = 0 \mid y)$ and $p(\rho = 0 \mid y^*)$, respectively. However, to our knowledge, the marginal posterior density for the variance-covariance matrix for the IV model with Jeffreys' prior is not derived in the literature. Under flat priors, the conditional density for $\Pi$ given $\beta$ and the marginal density of $\Pi$ under Jeffreys prior are derived in Kleibergen and Van Dijk (1994b) and Kleibergen and Van Dijk (1998). Under flat priors, conditional density for $\beta$ given $\Pi$ and the marginal density of $\beta$ are derived in Dréze (1976) and Dréze (1977). These derivations rely on integrating out the variance-covariance matrix in the model.

Drawing from the posterior density is not straightforward because of the local non-identification issue: For $\Pi = 0$, parameter $\beta$ drops out of the posterior density kernel in

(5.23). We use the Metropolis Hastings algorithm (MH; Metropolis *et al.* (1953); Hastings (1970)) to get posterior draws from $p(\beta, \Pi, \sigma_{11}, \rho, \sigma_{22})$, using the candidate density:

$$p(\beta, \Pi, \Sigma) = p(\Pi)\, p(\beta \mid \Pi)\, p(\Sigma \mid \beta, \Pi), \tag{5.24}$$

where $p(\Pi)$ is a matrix-variate Student-$t$ density with mean $\hat{\Pi} = (x'x)^{-1}x'y_2$ and the variance-covariance matrix $\hat{s}^2(x'x)^{-1}$, where $\hat{s}^2$ is the OLS estimate for the error terms' variance ($\text{var}(\sigma_{22})$). The degrees of freedom is chosen low as 4, in order to cover all of the relevant range of $\Pi$ values.

As a second step, $\beta$ is drawn conditionally upon $\Pi$ from its conditional posterior in the IV model under flat priors:

$$M_\nu y_1 = M_\nu y_2 \beta + \eta, \tag{5.25}$$

where $M_\nu = I - \nu(\nu'\nu)^{-1}\nu'\nu$ is the projection out of the space spanned by $\nu$. Similar to $p(\Pi)$, the degrees of freedom for this density is kept small, to cover the relevant range (in the applications we use candidate densities with 4 degrees of freedom).

The *conditional* candidate density $p(\Sigma \mid \beta, \Pi)$ is the Inverted Wishart distribution with parameters matching the sample variance-covariance matrix of the error terms, and $N - k$ degrees of freedom. The true conditional posterior would result if we would choose $k = 0$, but we set $k = 10$ to cover a relatively wide range of values in the parameter space by the candidate draws.

In order to infer the posterior densities for $\rho$, we use posterior $\rho$ draws from $p(\rho, \theta_{-\rho} \mid y)$. Then the marginal posterior density of $\rho$ is approximated using non-parametric density estimation. This method is required since neither the conditional posterior nor the marginal posterior of $\rho$ has a closed form solution[6].

In Section 5.3.1, we illustrate the performance of the predictive likelihood approach using simulated datasets with varying degrees of endogeneity and instrument strength. In Section 5.3.2, we implement this approach to assess the degree of endogeneity and instrument strength in the IV model for the income-education relationship in the US states and regions.

[6]When the analytical expression for the conditional densities $p(\rho \mid \theta_{-\rho}, y)$ and $p(\rho \mid \theta_{-\rho}, y^*)$ are available, there are a number of methods to approximate the marginal posterior densities. See e.g. Silverman (1998) for kernel density estimation methods. Several methods such Chib's estimator (Chib, 1995), and the numerical integration method proposed by Verdinelli and Wasserman (1995) can be used.

### 5.3.1 Simulated data with varying degrees of endogeneity and instrument strength

We first check the performance of the model comparison approach using SDDR and predictive likelihoods on simulated data, and show that this comparison takes into account the necessity of instruments together with the degree to which instruments can explain the endogenous variable. The data are simulated according to the IV model in (5.1) and (5.2), with a single endogenous explanatory variable and a single instrument, where for $i = 1, ..., N$, $x_i \sim NID(0, 1)$, and the variance-covariance matrix for the error terms is $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

We consider 12 sets of simulated data with 1000 observations, different degrees of endogeneity and varying instrument strength. Regarding the degree of endogeneity, four specifications are considered: strong, moderate, weak endogeneity cases and finally the case of no endogeneity, where the IV model is clearly not necessary. The respective parameter settings are $\rho = 0.9$, $\rho = 0.5$, $\rho = 0.1$ and $\rho = 0$. In terms of the instrument strength, we consider strong, weak and finally irrelevant instruments (where instruments have no explanatory power in explaining the possibly endogenous explanatory variable) with the corresponding parameter settings $\Pi = 1$, $\Pi = 0.1$ and $\Pi = 0$, respectively. Parameters for the simulated datasets are shown in Table 5.1.

Next we calculate the predictive model probabilities using (5.18), where we set the prior model probabilities $M_0$ and $M_1$ to 0.5, i.e. the prior model probabilities favor both models equally. Regarding the training sample choice, we report results under three training samples, with 500, 250 and 125 observations. Furthermore, we eliminate the effect of random sampling by repeating each estimation with 20 different training samples. Notice that we consider cross-section data that do not have the natural ordering of time series data. For time series data, one can use the first subset of observations as training sample. We randomly choose the training sample, and consider 20 random choices to decrease the effect of a particular random sample selection[7]. Posterior results for all simulated datasets and training samples are reported in Table 5.2.

We first focus on the cases where there is the problem of endogeneity in the dataset, and instruments have some power as a proxy for the endogenous variable, i.e. cases where the degree of endogeneity and instrument strength are different from 'none'. Table 5.2

[7]For a time series with independent observations, using a random training sample would not be fundamentally different from using a training sample of the first observations. In a certain sense, a cross section can be interpreted as a time series of independent observations. This makes it natural to apply the predictive likelihood concept in a cross section setting, using multiple selections for the training sample.

Table 5.1: Simulated dataset specifications for the IV model

| degree of endogeneity | degree of instrument strength | parameters $\rho$ | $\Pi$ | $\beta$ | $\sigma_{11}$ | $\sigma_{22}$ |
|---|---|---|---|---|---|---|
| strong | strong | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 |
| strong | weak | 0.90 | 0.10 | 1.00 | 1.00 | 1.00 |
| strong | none | 0.90 | 0.00 | 1.00 | 1.00 | 1.00 |
| moderate | strong | 0.50 | 1.00 | 1.00 | 1.00 | 1.00 |
| moderate | weak | 0.50 | 0.10 | 1.00 | 1.00 | 1.00 |
| moderate | none | 0.50 | 0.00 | 1.00 | 1.00 | 1.00 |
| weak | strong | 0.10 | 1.00 | 1.00 | 1.00 | 1.00 |
| weak | weak | 0.10 | 0.10 | 1.00 | 1.00 | 1.00 |
| weak | none | 0.10 | 0.00 | 1.00 | 1.00 | 1.00 |
| none | strong | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| none | weak | 0.00 | 0.10 | 1.00 | 1.00 | 1.00 |
| none | none | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |

*Note:* The table summarizes the IV model specifications for simulated datasets.

shows that given the same level of instrument strength, the predictive probability for the restricted model (in which $\beta$ is estimated without making use of the instruments) is lower (higher) for simulated data with relatively strong (weak) degrees of endogeneity, regardless of the training sample choice. This result confirms that the unrestricted IV model has a higher predictive probability when there is a clear problem of endogeneity in the data.

On the other hand, given the same level of endogeneity, the model probability for the restricted model without instruments is lower (higher) for simulated data with relatively strong (weak) instruments. If the instruments are rather weak in explaining the endogenous explanatory variable, the gains from employing an IV model are questionable, and the IV model has a relatively lower model probability. We conclude that comparison of models using predictive likelihoods accounts for the trade-off between the correct model specification and precision considerations.

The cases with no endogeneity and irrelevant instruments should be considered separately, as these are the cases where the standard IV model assumptions do not hold. In case of no endogeneity, we expect the predictive model probabilities to favor the model without instruments. However, the assessment of endogeneity depends on the strength of the instruments. In the extreme case where instruments have no explanatory power,

the IV model suffers from *local non-identification*: one cannot identify the effect of the explanatory variable ($y_2$) on the explained variable ($y_1$). Similarly, the degree of endogeneity is hard to assess as the test for endogeneity relies on the fit of the IV model in the first place. For this reason, in case of irrelevant instruments, the model probability of the IV model is not exactly zero.

We conclude that model probabilities impled by predictive likelihoods are not only assessing instrument strength ($\Pi = 0$), but also the necessity of instruments. Intuitively,

Table 5.2: Simulated data: Model probabilities for the restricted model in which $\beta$, the effect of education on income, is estimated without making use of the instruments (for which the posterior mean of $\beta$ is the ordinary least-squares (OLS) estimator in (5.1))

| degree of endogeneity | degree of instrument strength | training sample size 500 | 250 | 125 |
|---|---|---|---|---|
| strong | strong | 0* | 0* | 0* |
| strong | weak | 0* | 0* | 0* |
| strong | none | 0.59 | 0.64 | 0.64 |
| moderate | strong | 0* | 0* | 0 |
| moderate | weak | 0.33 | 0.33 | 0.3 |
| moderate | none | 0.52 | 0.55 | 0.55 |
| weak | strong | 0.00 | 0.00 | 0.00 |
| weak | weak | 0.57 | 0.66 | 0.68 |
| weak | none | 0.29 | 0.26 | 0.24 |
| none | strong | 0.47 | 0.57 | 0.62 |
| none | weak | 0.64 | 0.71 | 0.78 |
| none | none | 0.44 | 0.46 | 0.45 |

*Note*: The table reports mean predictive model probabilities calculated from (5.18) for simulated datasets. For each simulated data and training sample size, 20 repetitions with random subsamples are performed, for which results are averaged. Posterior density estimates for $\rho$ are achieved by nonparametric kernel density estimation.

* denotes the cases for which both densities in (5.18) (i.e. $p(\rho = 0 \mid \tilde{y}, y^*)$ and $p(\rho = 0 \mid y^*)$) are estimated as zero by the kernel density method, at least in one of the subsamples considered. If we generated a large enough set of random draws to estimate these densities as non-zero, which would typically have to be huge, then the predictive probability of the restricted model would be estimated as (almost) 0, since $p(\rho = 0 \mid \tilde{y}, y^*)$ would be much smaller than $p(\rho = 0 \mid y^*)$). Roughly stated, in the current setting the zero value of the former is much more zero than the latter. We set the predictive probability of $M_0$ to 0 for these cases.

this comparison incorporates the degree to which instruments provide information for the data generating process, and efficiency losses from using instrumental variables when the degree of endogeneity is small. Further, when the level of endogeneity is not strong and the instruments are relatively weak, model probabilities are very close to 0.5. We therefore conclude that a *model choice* in these cases can be inappropriate, as none of the models are clearly favored. For this reason, we will analyze the application of BMA, to illustrate possible gains from averaging the results of two alternative models instead of choosing one of the models.

We estimate the posterior density of $\beta$ for using BMA, using the predictive model probabilities reported in Table 5.2. Table 5.3 reports model averaging results for simulated data. $M_0$ denotes the nested model, $M_1$ denotes the IV model and 'average' denotes the estimates obtained by model averaging using predictive model probabilities in Table 5.2. Average posterior means and variances for $\beta$ are calculated from (5.20). We report BMA results using model probabilities achieved from training samples with 250 observations (25% of the full sample size) and note that the results for other training sample choices are quite similar.

Table 5.3 shows that averaging over the IV model and the model without instruments leads to smaller variances in the $\beta$ parameter, except for the cases of strong instruments. Notice that in case of strong instruments, the efficiency loss from employing the IV model is rather negligible, hence model averaging does not improve the posterior results for $\beta$.

For the cases where instruments have no explanatory power, the IV model leads to quite large posterior standard deviations, and mean $\beta$ estimates are quite far from the true value. BMA in these cases leads to relatively more accurate results in posterior $\beta$ density. However, as the predictive model probabilities are calculated *wrongly* by the IV model, the advantage of BMA is questionable.

In the next section we apply the proposed method to Angrist and Krueger (1991) data on the income-education relationship.

Table 5.3: Simulated data: posterior $\beta$ for the IV model and the nested model without instruments, and BMA

| Degree of endogeneity | Degree of identification | Model | $\beta$ | |
|---|---|---|---|---|
| | | | Mean | Std. Dev. |
| strong | strong | $M_0$ | 1.47 | 0.00 |
| | | average | 1.05 | 0.02 |
| | | $M_1$ | 1.05 | 0.02 |
| strong | weak | $M_0$ | 1.89 | 0.00 |
| | | average | 0.89 | 9.36 |
| | | $M_1$ | 0.89 | 9.36 |
| strong | none | $M_0$ | 1.88 | 0.00 |
| | | average | 1.97 | 1.87 |
| | | $M_1$ | 2.16 | 5.55 |
| moderate | strong | $M_0$ | 1.24 | 0.00 |
| | | average | 0.98 | 0.03 |
| | | $M_1$ | 0.98 | 0.03 |
| moderate | weak | $M_0$ | 1.47 | 0.00 |
| | | average | 1.08 | 0.20 |
| | | $M_1$ | 0.90 | 0.29 |
| moderate | none | $M_0$ | 1.48 | 0.00 |
| | | average | 1.54 | 4.11 |
| | | $M_1$ | 1.62 | 9.12 |
| weak | strong | $M_0$ | 1.08 | 0.00 |
| | | average | 0.99 | 0.03 |
| | | $M_1$ | 0.99 | 0.03 |
| weak | weak | $M_0$ | 1.10 | 0.00 |
| | | average | 1.18 | 0.22 |
| | | $M_1$ | 1.37 | 0.72 |
| weak | none | $M_0$ | 1.08 | 0.00 |
| | | average | -1.78 | 24.86 |
| | | $M_1$ | -2.83 | 33.95 |
| none | strong | $M_0$ | 1.02 | 0.00 |
| | | average | 1.01 | 0.02 |
| | | $M_1$ | 0.99 | 0.04 |
| none | weak | $M_0$ | 1.03 | 0.00 |
| | | average | 1.08 | 0.09 |
| | | $M_1$ | 1.21 | 0.38 |
| none | none | $M_0$ | 1.01 | 0.00 |
| | | average | 0.88 | 22.94 |
| | | $M_1$ | 0.78 | 41.07 |

*Note*: The table reports posterior means and standard deviations for posterior $\beta$ draws under the IV model, the model without instruments and the average effect as a combination of both models, using weights achieved by the training sample with 250 observations, reported in Table 5.2.

### 5.3.2 Income-education relationship in Angrist and Krueger (1991) data

In this section, we apply the predictive likelihood approach for model comparison or combination to the Angrist and Krueger (1991) data on income and education. Angrist and Krueger (1991) data consists of men born in the US during the periods 1920-1929, 1930-1939 and 1940-1949, where the data for the first group is collected in 1970, and the data for the last two groups are collected in 1980. We use a subset of their data, consisting of men born during the period 1930-1939, including the data on weakly wages, number of completed years of education and instruments consisting of quarter of birth dummies. The data includes 51 states and 329.509 observations [8].

We first analyze the income-education relationship for each state using the model:

$$\tilde{y}_i = \alpha_1 + \tilde{x}_i\beta + \sum_{t=1}^{9} D_{t,i}\delta_t + \tilde{\epsilon}_i \tag{5.26}$$

$$\tilde{x}_i = \alpha_2 + \sum_{q=2}^{4} D_{q,i}\Pi_q + \sum_{t=1}^{9} D_{t,i}\delta_t + \tilde{\nu}_i \tag{5.27}$$

where $\tilde{y}_i$ and $\tilde{x}_i$ are the natural logarithm of the weekly wage and the number of completed years of education for the person $i$ in 1979, respectively.

In (5.26) and (5.27), $D_{t,i}$ for $t = \{1, \ldots, 9\}$ are the dummy variables for year of birth which take the value 1 if individual $i$ was born in year $1929 + t$, and 0 otherwise. $D_{q,i}$ for $q = \{2, 3, 4\}$ are the quarter of birth dummy variables which take the value 1 if individual $i$ was born in quarter $q$, and 0 otherwise. $\alpha_1$ and $\alpha_2$ are constants, and $\tilde{\epsilon}_i$ and $\tilde{\nu}_i$ are disturbances assumed to be normally distributed, and independent across individuals.

The model in (5.26) and (5.27) is similar to the model of Hoogerheide and Van Dijk (2006). For simplicity, we do not consider interactions of year dummies and quarter of birth dummies as instruments. Furthermore, the model employed here does not include state dummies, as each state is analyzed separately. Similar to Hoogerheide and Van Dijk (2006), we simplify the IV model in (5.26) and (5.27) correcting for the constant terms and exogenous year of birth dummies. Using this simplification, the IV model becomes:

$$y_i = x_i\beta + \epsilon_i, \tag{5.28}$$

$$x_i = Z_i\Pi + \nu_i, \tag{5.29}$$

where $y_i$, $x_i$ are the residuals from regressing the log weakly wage and years of education on a constant and year of birth dummies, respectively. $Z_i$ is the $3 \times 1$ vector of instru-

[8] The source of the data is the 1980 Census, 5 percent public sample, also available from http://econ-www.mit.edu/faculty/angrist/data1/data/angkru1991.

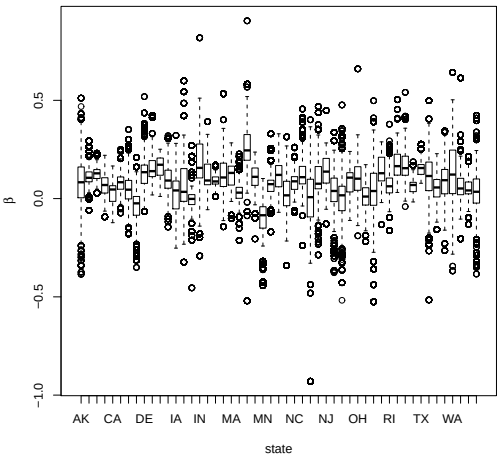
ments, obtained from regressing quarter of birth dummies on a constant and the year of birth dummies. $\epsilon_i$ and $\nu_i$ are the error terms that have a joint normal distribution, and are uncorrelated across individuals.

**IV model results for US states**

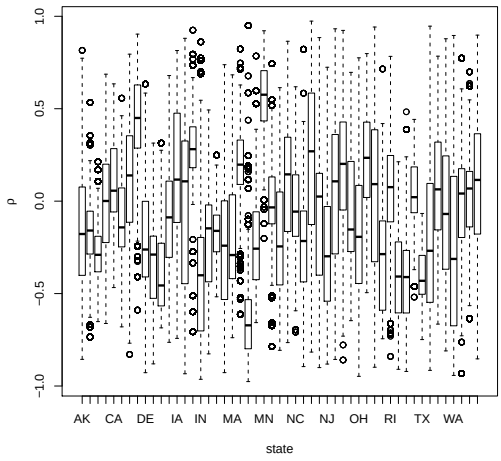
We first consider the posterior parameters of the IV model for each state. Table 5.4 reports estimated posterior means of the parameters for the IV model with the respective estimated posterior standard deviations. Posterior standard deviations for $\beta$ and $\rho$ are quite high, indicating low explanatory power of the instruments. Furthermore, posterior results are quite different across states, as reported in Hoogerheide and Van Dijk (2006).

The heterogeneity of posterior parameters across states is also presented by boxplots of the posterior draws. Boxplots of posterior $\beta$ and $\rho$ draws are given in Figure 5.1. Posterior means of $\beta$, i.e. the effect of education on income are above zero for most states, but the posterior $\beta$ densities are quite wide. Regarding the degree of endogeneity $\rho$, we also see a clear heterogeneity across states.

Figure 5.1: Income-education effects in US states: Boxplots for income effects and degree of endogeneity




Posterior $\beta$ values




Posterior $\rho$ values

Table 5.4: Income-education effects in US states: Posterior results for parameters

| | $\beta$ | $\Pi_2$ | $\Pi_3$ | $\Pi_4$ | $\rho$ |
|---|---|---|---|---|---|
| Alaska | 0.08 | 1.54 | 1.5 | -0.32 | -0.15 |
| | (0.12) | (0.76) | (0.77) | (0.87) | (0.33) |
| Alabama | 0.12 | 0.03 | 0.33 | 0.29 | -0.19 |
| | (0.05) | (0.08) | (0.07) | (0.08) | (0.21) |
| Arkansas | 0.13 | -0.18 | 0.10 | 0.44 | -0.28 |
| | (0.04) | (0.09) | (0.08) | (0.08) | (0.17) |
| Arizona | 0.07 | 0.65 | -0.03 | -0.30 | -0.02 |
| | (0.05) | (0.21) | (0.20) | (0.22) | (0.27) |
| California | 0.04 | 0.26 | 0.22 | 0.13 | 0.08 |
| | (0.06) | (0.05) | (0.05) | (0.05) | (0.26) |
| Colorado | 0.08 | 0.31 | 0.44 | 0.42 | -0.11 |
| | (0.06) | (0.11) | (0.10) | (0.11) | (0.25) |
| Connecticut | 0.05 | 0.30 | 0.12 | 0.12 | 0.10 |
| | (0.09) | (0.09) | (0.10) | (0.10) | (0.37) |
| District of Columbia | -0.04 | -0.44 | -0.41 | -0.50 | 0.43 |
| | (0.09) | (0.16) | (0.17) | (0.16) | (0.27) |
| Delaware | 0.13 | 0.56 | 0.69 | 0.15 | -0.20 |
| | (0.08) | (0.22) | (0.22) | (0.23) | (0.31) |
| Florida | 0.15 | 0.32 | 0.24 | 0.27 | -0.32 |
| | (0.07) | (0.10) | (0.09) | (0.09) | (0.26) |
| Georgia | 0.16 | -0.25 | 0.05 | 0.03 | -0.39 |
| | (0.05) | (0.05) | (0.06) | (0.06) | (0.22) |
| Hawaii | 0.10 | 0.09 | 1.55 | 0.82 | -0.09 |
| | (0.07) | (0.41) | (0.39) | (0.35) | (0.28) |
| Iowa | 0.03 | -0.04 | -0.02 | 0.09 | 0.11 |
| | (0.10) | (0.06) | (0.06) | (0.06) | (0.36) |
| Idaho | 0.05 | 0.16 | -0.01 | 0.11 | 0.02 |
| | (0.13) | (0.16) | (0.17) | (0.14) | (0.45) |
| Illinois | 0.00 | 0.07 | -0.07 | 0.06 | 0.24 |
| | (0.08) | (0.03) | (0.04) | (0.04) | (0.29) |
| Indiana | 0.16 | 0.04 | 0.08 | 0.04 | -0.32 |
| | (0.15) | (0.05) | (0.05) | (0.05) | (0.44) |
| Kansas | 0.11 | 0.30 | 0.34 | 0.19 | -0.20 |
| | (0.07) | (0.07) | (0.08) | (0.08) | (0.26) |
| Kentucky | 0.09 | 0.08 | 0.35 | 0.55 | -0.18 |
| | (0.03) | (0.07) | (0.07) | (0.07) | (0.17) |
| Louisiana | 0.12 | 0.10 | 0.26 | 0.28 | -0.23 |
| | (0.10) | (0.09) | (0.09) | (0.10) | (0.36) |
| Massachusetts | 0.12 | 0.13 | 0.17 | 0.28 | -0.20 |
| | (0.07) | (0.06) | (0.06) | (0.07) | (0.32) |
| Maryland | 0.03 | 0.38 | 0.43 | 0.34 | 0.17 |
| | (0.06) | (0.10) | (0.10) | (0.09) | (0.24) |
| Maine | 0.26 | 0.01 | 0.28 | 0.02 | -0.62 |
| | (0.13) | (0.09) | (0.10) | (0.11) | (0.26) |
| Michigan | 0.10 | 0.15 | 0.11 | 0.11 | -0.20 |
| | (0.07) | (0.03) | (0.04) | (0.04) | (0.28) |
| Minnesota | -0.10 | -0.20 | -0.20 | -0.14 | 0.55 |
| | (0.10) | (0.06) | (0.06) | (0.05) | (0.21) |
| Missouri | 0.07 | -0.08 | 0.09 | 0.00 | -0.02 |
| | (0.08) | (0.06) | (0.05) | (0.05) | (0.30) |
| Mississippi | 0.11 | 0.07 | 0.22 | 0.33 | -0.16 |
| | (0.08) | (0.08) | (0.09) | (0.08) | (0.36) |

continued on Next Page. . .

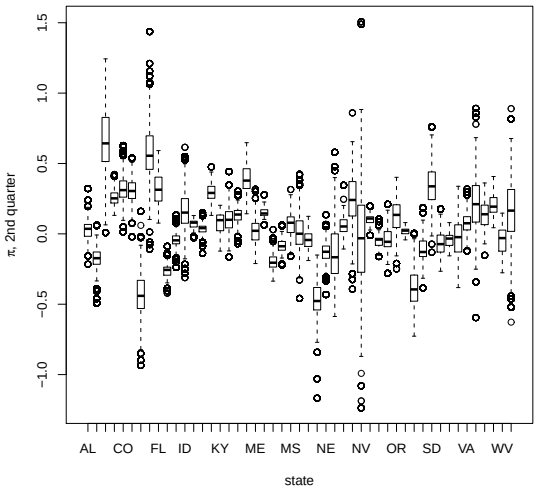
Table 5.4 continued...

| | $\beta$ | $\Pi_2$ | $\Pi_3$ | $\Pi_4$ | $\rho$ |
|---|---|---|---|---|---|
| Montana | 0.02 | 0.01 | 0.09 | 0.39 | 0.08 |
| | (0.10) | (0.13) | (0.14) | (0.14) | (0.34) |
| North Carolina | 0.09 | -0.04 | 0.13 | 0.24 | -0.06 |
| | (0.05) | (0.06) | (0.06) | (0.07) | (0.25) |
| North Dakota | 0.13 | -0.48 | -0.19 | -0.08 | -0.25 |
| | (0.09) | (0.13) | (0.14) | (0.15) | (0.30) |
| Nebraska | 0.00 | -0.13 | -0.15 | -0.12 | 0.20 |
| | (0.16) | (0.09) | (0.10) | (0.08) | (0.44) |
| New Hampshire | 0.10 | -0.14 | 0.10 | 0.20 | -0.08 |
| | (0.10) | (0.19) | (0.18) | (0.2) | (0.39) |
| New Jersey | 0.13 | 0.06 | 0.07 | 0.21 | -0.24 |
| | (0.08) | (0.06) | (0.06) | (0.08) | (0.31) |
| New Mexico | 0.04 | 0.23 | 0.11 | 0.39 | 0.06 |
| | (0.10) | (0.18) | (0.19) | (0.17) | (0.41) |
| Nevada | 0.00 | -0.03 | 0.07 | 0.85 | 0.17 |
| | (0.11) | (0.38) | (0.37) | (0.35) | (0.35) |
| New York | 0.09 | 0.10 | 0.05 | -0.05 | -0.05 |
| | (0.07) | (0.04) | (0.04) | (0.04) | (0.30) |
| Ohio | 0.11 | -0.04 | 0.05 | 0.05 | -0.17 |
| | (0.10) | (0.06) | (0.05) | (0.05) | (0.37) |
| Oklahoma | 0.01 | -0.04 | 0.16 | 0.23 | 0.21 |
| | (0.07) | (0.08) | (0.07) | (0.07) | (0.28) |
| Oregon | 0.05 | 0.12 | 0.11 | 0.06 | 0.01 |
| | (0.15) | (0.12) | (0.12) | (0.12) | (0.47) |
| Pennsylvania | 0.15 | 0.02 | 0.01 | 0.05 | -0.33 |
| | (0.07) | (0.03) | (0.03) | (0.03) | (0.26) |
| Rhode Island | 0.07 | -0.39 | 0.11 | 0.12 | 0.04 |
| | (0.07) | (0.15) | (0.17) | (0.18) | (0.3) |
| South Carolina | 0.17 | -0.11 | -0.05 | 0.31 | -0.39 |
| | (0.07) | (0.09) | (0.07) | (0.08) | (0.25) |
| South Dakota | 0.17 | 0.35 | 0.30 | 0.56 | -0.42 |
| | (0.08) | (0.14) | (0.13) | (0.15) | (0.24) |
| Tennessee | 0.06 | -0.06 | 0.19 | 0.47 | 0.07 |
| | (0.03) | (0.08) | (0.07) | (0.07) | (0.17) |
| Texas | 0.16 | -0.04 | 0.23 | 0.26 | -0.43 |
| | (0.06) | (0.05) | (0.05) | (0.07) | (0.19) |
| Utah | 0.11 | -0.02 | -0.15 | -0.25 | -0.20 |
| | (0.13) | (0.14) | (0.15) | (0.16) | (0.45) |
| Virginia | 0.06 | 0.08 | 0.30 | 0.32 | 0.05 |
| | (0.07) | (0.08) | (0.09) | (0.08) | (0.31) |
| Vermont | 0.09 | 0.22 | 0.47 | 0.33 | -0.06 |
| | (0.10) | (0.19) | (0.22) | (0.21) | (0.39) |
| Washington | 0.13 | 0.14 | 0.12 | 0.00 | -0.22 |
| | (0.15) | (0.09) | (0.08) | (0.11) | (0.47) |
| Wisconsin | 0.07 | 0.21 | 0.01 | 0.10 | -0.01 |
| | (0.08) | (0.06) | (0.08) | (0.07) | (0.28) |
| West Virginia | 0.05 | -0.04 | 0.09 | 0.27 | 0.04 |
| | (0.06) | (0.09) | (0.07) | (0.07) | (0.26) |
| Wyoming | 0.03 | 0.14 | 0.38 | -0.13 | 0.09 |
| | (0.11) | (0.23) | (0.22) | (0.25) | (0.38) |

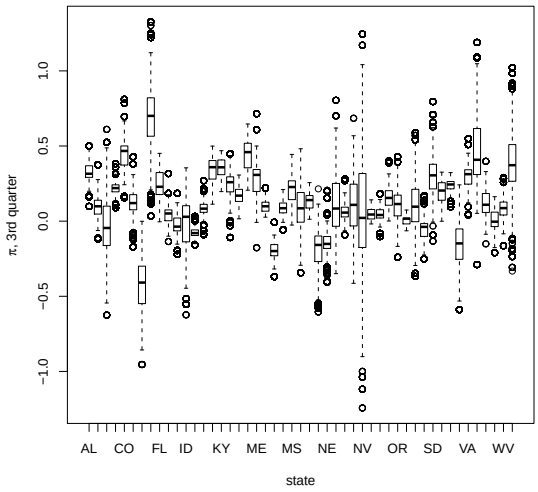
*Note*: The table reports posterior means for the parameters and posterior standard deviations (in parentheses) for each state. $\Pi_2$, $\Pi_3$ and $\Pi_4$ are the coefficients for the 2nd, 3rd and 4th quarter of birth dummies, respectively. Posterior results are achieved by 30000 draws (3000 burn-in).

Boxplots of the posterior draws for the effects of quarter of birth on years of education are given in Figure 5.2. Similar to the rest of the model parameters, we see different posterior densities across states.

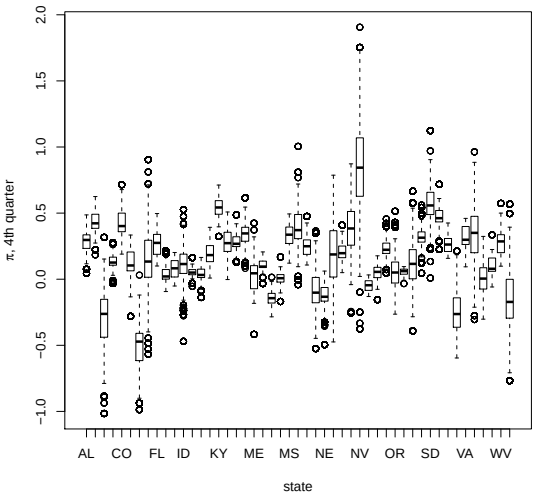
Figure 5.2: Income-education effects in US states: Boxplots for the quarter of birth effects on education




Posterior $\Pi_2$ values




Posterior $\Pi_3$ values




Posterior $\Pi_4$ values

*Note*: For illustrative purposes, boxplots of $\Pi_2$, $\Pi_3$ and $\Pi_4$ do not include results for two states: Alaska and Hawaii. The number of observations for these states are quite small, and the posterior densities for the effects of quarter of birth on education are much wider compared to rest of the states.

Percentiles for the posterior $\rho$ densities are reported in Table 5.5. 95% intervals for posterior $\rho$ densities include point '0' for all states except for Maine, Minnesota and

Texas. Hence the states' data does not show a clear endogeneity problem. We note that this can be a result of weak instruments (as posterior $\rho$ densities depend on the degree of instrument strength in the data) or the availability of too few observations per state.

Table 5.5: Income-education effects in US states: percentiles for posterior $\rho$ densities

| *Percentiles* | 5% | 50% | 95% | | 5% | 50% | 95% |
|---|---|---|---|---|---|---|---|
| Alaska | -0.65 | -0.18 | 0.46 | Montana | -0.52 | 0.14 | 0.58 |
| Alabama | -0.57 | -0.16 | 0.14 | North Carolina | -0.58 | -0.06 | 0.39 |
| Arkansas | -0.54 | -0.29 | 0.06 | North Dakota | -0.77 | -0.22 | 0.18 |
| Arizona | -0.47 | 0.00 | 0.38 | Nebraska | -0.61 | 0.27 | 0.74 |
| California | -0.39 | 0.06 | 0.50 | New Hampshire | -0.71 | 0.02 | 0.57 |
| Colorado | -0.58 | -0.14 | 0.33 | New Jersey | -0.63 | -0.30 | 0.31 |
| Connecticut | -0.59 | 0.14 | 0.68 | New Mexico | -0.63 | 0.11 | 0.75 |
| District of Columbia | -0.04 | 0.45 | 0.83 | Nevada | -0.46 | 0.20 | 0.71 |
| Delaware | -0.69 | -0.26 | 0.36 | New York | -0.48 | -0.15 | 0.44 |
| Florida | -0.74 | -0.29 | 0.18 | Ohio | -0.78 | -0.19 | 0.44 |
| Georgia | -0.61 | -0.46 | 0.01 | Oklahoma | -0.36 | 0.23 | 0.63 |
| Hawaii | -0.54 | -0.09 | 0.36 | Oregon | -0.80 | 0.09 | 0.74 |
| Iowa | -0.48 | 0.12 | 0.67 | Pennsylvania | -0.70 | -0.29 | 0.11 |
| Idaho | -0.63 | 0.11 | 0.77 | Rhode Island | -0.54 | 0.08 | 0.45 |
| Illinois | -0.51 | 0.28 | 0.62 | South Carolina | -0.72 | -0.41 | 0.06 |
| Indiana | -0.86 | -0.40 | 0.54 | South Dakota | -0.80 | -0.41 | 0.04 |
| Kansas | -0.63 | -0.15 | 0.16 | Tennessee | -0.19 | 0.02 | 0.35 |
| Kentucky | -0.51 | -0.16 | 0.07 | Texas | -0.75 | -0.43 | -0.12 |
| Louisiana | -0.76 | -0.24 | 0.32 | Utah | -0.79 | -0.27 | 0.68 |
| Massachusetts | -0.63 | -0.29 | 0.33 | Virginia | -0.44 | 0.06 | 0.55 |
| Maryland | -0.30 | 0.20 | 0.44 | Vermont | -0.75 | -0.07 | 0.59 |
| Maine | -0.90 | -0.67 | -0.15 | Washington | -0.82 | -0.31 | 0.56 |
| Michigan | -0.60 | -0.26 | 0.28 | Wisconsin | -0.49 | 0.04 | 0.44 |
| Minnesota | 0.06 | 0.58 | 0.88 | West Virginia | -0.33 | 0.07 | 0.50 |
| Missouri | -0.52 | -0.03 | 0.50 | Wyoming | -0.50 | 0.11 | 0.72 |

*Note*: The table reports posterior median and 5% and 95% percentiles for the posterior $\rho$ draws for each state. Posterior results for each state are achieved by 30000 draws (3000 burn-in).

**Predictive model probabilities and Model Averaging Results for the US states:** We next estimate predictive model probabilities for the IV model and the model without instruments for the US states. Similar to the simulated data application, prior model probabilities are chosen to be equal. We consider two sets of *training samples* for each state. The first training sample consists of 10% of the data for each state. The second training sample consists of 5% of the data for each state[9]. Furthermore, we consider the effect of random training sample choice by summarizing predictive model probabilities from 20 different random training samples in each case.

Predictive model probabilities for the model without instruments for the US states are reported in Table 5.6 for all states and both training samples. The first two columns in Table 5.6 show the average model probabilities achieved by 20 different training samples. Last two columns in Table 5.6 show model probabilities obtained by a single training sample only.

Table 5.6 shows that model probabilities achieved by different training sample sizes and across random subsets are quite close to each other. Only for a few states with relatively small number of observations, such as South Dakota, model probabilities differ substantially across training samples.

Model probabilities are quite close to 0.5 and do not indicate a specific model, except for a few states such as Texas and Tennessee. For Texas, model probabilities indicate that the IV model is necessary. For Tennessee on the other hand, we find strong evidence against the need for the IV model. We conclude that choosing one of the alternative models according to these probabilities can be quite inaccurate, and employ model averaging to infer the state-specific effects of income on education.

The *average* effects of education on income for the US states, i.e. the posterior distributions resulting from BMA, are given in Table 5.7. Model probabilities are achieved by using training samples with 5% and 10% of the observations, averaged over 20 repetitions reported in the first two columns of Table 5.6.

Posterior results for the effects of education on income achieved by model averaging are close to the IV estimates reported in Table 5.4. The main advantage of model averaging is the improved efficiency of the estimates. Standard deviations of posterior $\beta$ draws are less than half of those achieved by the IV model only.

[9]Training samples consisting of 25% of the data provide similar results.

Table 5.6: Income-education effects in US states: Predictive model probabilities of the restricted model, in which $\beta$, the effect of education on income, is estimated without making use of the instruments, for US states

| | number of obs. | 20 repetitions training sample | | single training sample training sample | |
|---|---|---|---|---|---|
| | | 10 % | 5 % | 10 % | 5 % |
| Alabama | 8536 | 0.54 | 0.57 | 0.44 | 0.41 |
| Arkansas | 5794 | 0.29 | 0.28 | 0.19 | 0.25 |
| Arizona | 1066 | 0.63 | 0.64 | 0.72 | 0.64 |
| California | 11078 | 0.83 | 0.84 | 0.84 | 0.80 |
| Colorado | 2818 | 0.64 | 0.68 | 0.67 | 0.80 |
| Connecticut | 3844 | 0.60 | 0.63 | 0.46 | 0.51 |
| District of Columbia | 1237 | 0.45 | 0.47 | 0.57 | 0.60 |
| Delaware | 598 | 0.57 | 0.55 | 0.67 | 0.60 |
| Florida | 3913 | 0.38 | 0.41 | 0.42 | 0.27 |
| Georgia | 8411 | 0.57 | 0.57 | 0.76 | 0.74 |
| Hawaii | 246 | 0.58 | 0.67 | 0.48 | 0.64 |
| Iowa | 6699 | 0.43 | 0.44 | 0.73 | 0.47 |
| Idaho | 1599 | 0.38 | 0.41 | 0.39 | 0.40 |
| Illinois | 18375 | 0.62 | 0.57 | 0.69 | 0.52 |
| Indiana | 8918 | 0.21 | 0.19 | 0.20 | 0.21 |
| Kansas | 4807 | 0.66 | 0.58 | 0.58 | 0.53 |
| Kentucky | 8933 | 0.62 | 0.67 | 0.58 | 0.62 |
| Louisiana | 5975 | 0.62 | 0.55 | 0.65 | 0.85 |
| Massachusetts | 9955 | 0.65 | 0.69 | 0.71 | 0.59 |
| Maryland | 4139 | 0.53 | 0.52 | 0.80 | 0.56 |
| Maine | 2424 | 0.32 | 0.34 | 0.39 | 0.19 |
| Michigan | 14077 | 0.59 | 0.53 | 0.58 | 0.55 |
| Minnesota | 7170 | 0.28 | 0.31 | 0.19 | 0.31 |
| Missouri | 9274 | 0.64 | 0.69 | 0.72 | 0.87 |
| Mississippi | 5864 | 0.56 | 0.56 | 0.50 | 0.47 |
| Montana | 1407 | 0.60 | 0.58 | 0.66 | 0.52 |
| North Carolina | 10798 | 0.69 | 0.69 | 0.66 | 0.72 |

continued on Next Page...

Table 5.6 continued...

| | number of obs. | 20 repetitions training sample | | single training sample training sample | |
|---|---|---|---|---|---|
| | | 10 % | 5 % | 10 % | 5 % |
| North Dakota | 2028 | 0.59 | 0.60 | 0.52 | 0.62 |
| Nebraska | 3488 | 0.41 | 0.43 | 0.39 | 0.44 |
| New Hampshire | 1200 | 0.60 | 0.60 | 0.57 | 0.44 |
| New Jersey | 8964 | 0.66 | 0.68 | 0.61 | 0.82 |
| New Mexico | 1325 | 0.51 | 0.53 | 0.53 | 0.49 |
| Nevada | 308 | 0.51 | 0.56 | 0.44 | 0.54 |
| New York | 29015 | 0.62 | 0.55 | 0.65 | 0.45 |
| Ohio | 17070 | 0.59 | 0.61 | 0.53 | 0.68 |
| Oklahoma | 6950 | 0.57 | 0.6 | 0.61 | 0.54 |
| Oregon | 2127 | 0.34 | 0.36 | 0.48 | 0.32 |
| Pennsylvania | 26385 | 0.53 | 0.49 | 0.68 | 0.53 |
| Rhode Island | 1698 | 0.61 | 0.60 | 0.49 | 0.47 |
| South Carolina | 5245 | 0.55 | 0.57 | 0.47 | 0.51 |
| South Dakota | 1754 | 0.08 | 0.09 | 0.06 | 0.24 |
| Tennessee | 8335 | 0.85 | 0.86 | 0.81 | 0.82 |
| Texas | 15932 | 0.01 | 0.01 | 0.00 | 0.01 |
| Utah | 1999 | 0.46 | 0.47 | 0.32 | 0.46 |
| Virginia | 7319 | 0.61 | 0.61 | 0.69 | 0.58 |
| Vermont | 999 | 0.67 | 0.67 | 0.70 | 0.73 |
| Washington | 3610 | 0.46 | 0.47 | 0.34 | 0.29 |
| Wisconsin | 8607 | 0.63 | 0.59 | 0.59 | 0.48 |
| West Virginia | 6412 | 0.54 | 0.56 | 0.42 | 0.58 |
| Wyoming | 706 | 0.50 | 0.53 | 0.48 | 0.45 |

*Note*: The table reports predictive probabilities of the model without instruments for the US states, achieved by 30000 draws (3000 burn-in) for two sets of training samples: 10% sample and 5% sample. Training samples are selected by random draws from the data.

* For Alaska, a 5% sample of the data consists of 3 observations only. Hence the IV estimation with three instruments for the training sample cannot implemented.

Table 5.7: Income-education effects in US states: *average* effects of education on income in US states

| State | Mean | Std. Dev. | State | Mean | Std. Dev. |
|---|---|---|---|---|---|
| Alabama | 0.11 | 0.03 | Montana | 0.04 | 0.04 |
| Arkansas | 0.11 | 0.03 | North Carolina | 0.08 | 0.02 |
| Arizona | 0.07 | 0.02 | North Dakota | 0.09 | 0.04 |
| California | 0.05 | 0.01 | Nebraska | 0.03 | 0.09 |
| Colorado | 0.07 | 0.02 | New Hampshire | 0.09 | 0.04 |
| Connecticut | 0.06 | 0.04 | New Jersey | 0.09 | 0.03 |
| District of Columbia | 0.02 | 0.05 | New Mexico | 0.05 | 0.05 |
| Delaware | 0.10 | 0.04 | Nevada | 0.03 | 0.06 |
| Florida | 0.13 | 0.05 | New York | 0.08 | 0.03 |
| Georgia | 0.12 | 0.02 | Ohio | 0.08 | 0.04 |
| Hawaii | 0.08 | 0.04 | Oklahoma | 0.04 | 0.03 |
| Iowa | 0.05 | 0.06 | Oregon | 0.05 | 0.10 |
| Idaho | 0.05 | 0.08 | Pennsylvania | 0.11 | 0.03 |
| Illinois | 0.04 | 0.03 | Rhode Island | 0.07 | 0.03 |
| Indiana | 0.15 | 0.12 | South Carolina | 0.12 | 0.03 |
| Kansas | 0.08 | 0.03 | South Dakota | 0.16 | 0.07 |
| Kentucky | 0.07 | 0.01 | Tennessee | 0.07 | 0.01 |
| Louisiana | 0.09 | 0.04 | Texas | 0.16 | 0.06 |
| Massachusetts | 0.09 | 0.03 | Utah | 0.09 | 0.07 |
| Maryland | 0.05 | 0.03 | Virginia | 0.06 | 0.03 |
| Maine | 0.21 | 0.09 | Vermont | 0.08 | 0.04 |
| Michigan | 0.08 | 0.03 | Washington | 0.10 | 0.09 |
| Minnesota | -0.06 | 0.08 | Wisconsin | 0.06 | 0.03 |
| Missouri | 0.07 | 0.03 | West Virginia | 0.05 | 0.03 |
| Mississippi | 0.09 | 0.04 | Wyoming | 0.04 | 0.06 |

*Note:* The table reports means and standard deviations of effect of education on income, resulting from BMA, for the US states, achieved by 30000 draws (3000 burn-in) for two sets of training samples: 10% sample and 5% sample. Training samples are selected by random draws from the data.

* For Alaska, 5% sample of the data consists of 3 observations only. Hence the IV estimation with three instruments for the training sample cannot implemented.

**Regional patterns in income-education relationship**

In the previous subsection, we showed that the income-education relationship in the US states shows clear heterogeneity in terms of the degree of endogeneity, and the rest of the parameters. We next analyze these differences in regional data. We apply the IV model in (5.28) and (5.29) to 9 divisions for the Angrist and Krueger (1991) data, according to the Census Bureau designated areas. The US divisions, together with the regions are summarized in Table 5.8.

Table 5.8: US regions and divisions

| Division | States | Number of observations |
|---|---|---|
| | *Northeast Region* | |
| 1. New England | Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island and Vermont | 20120 |
| 2. Middle Atlantic | New Jersey, New and Pennsylvania | 64364 |
| | *Midwest Region* | |
| 3. East North Central | Illinois, Indiana, Michigan, Ohio and Wisconsin | 67047 |
| 4. West North Central | Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota | 35220 |
| | *South Region* | |
| 5. South Atlantic | Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia and West Virginia | 48072 |
| 6. East South Central | Alabama, Kentucky, Mississippi and Tennessee | 31668 |
| 7. West South Central | Arkansas, Louisiana, Oklahoma and Texas | 34651 |
| | *West Region* | |
| 8. Mountain | Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah and Wyoming | 11228 |
| 9. Pacific | Alaska, California, Hawaii, Oregon and Washington | 17139 |

Table 5.9 reports posterior results of the IV model for US divisions. The posterior results for education effects on income are quite different across divisions. Especially for the West North Central region, the posterior standard deviation is quite high, indicating the relatively very weak instruments in this region. The differences in posterior results of the IV model between US divisions are shown in detail by the boxplots of posterior

draws in Figures 5.3 and 5.4. Similar to previous findings, we do not see similar posterior densities across divisions.

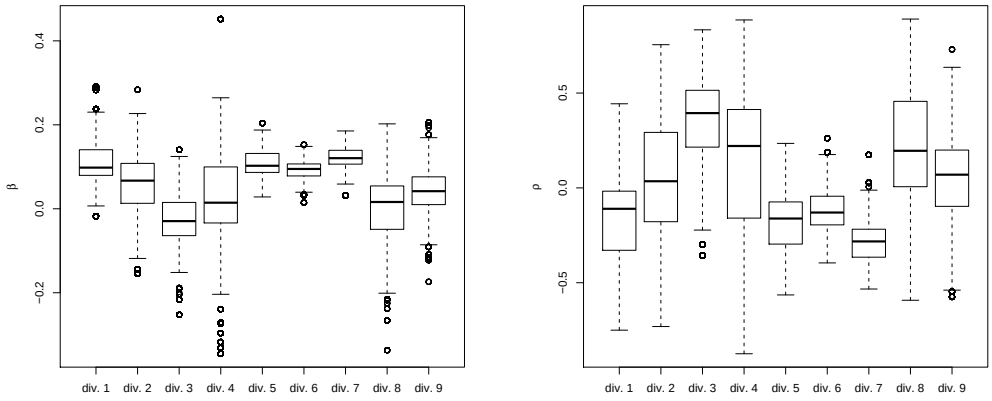
Table 5.9: Income-education effects in US divisions: parameter estimates

| | $\beta$ | $\Pi_2$ | $\Pi_3$ | $\Pi_4$ | $\rho$ |
|---|---|---|---|---|---|
| New England Division | 0.11 | 0.09 | 0.17 | 0.21 | -0.16 |
| | (0.05) | (0.04) | (0.04) | (0.04) | (0.23) |
| Middle Atlantic Division | 0.07 | 0.07 | 0.03 | 0.03 | 0.03 |
| | (0.07) | (0.02) | (0.02) | (0.03) | (0.31) |
| East North Central Division | -0.03 | 0.07 | 0.02 | 0.08 | 0.36 |
| | (0.08) | (0.02) | (0.02) | (0.02) | (0.25) |
| West North Central Division | 0.02 | -0.06 | 0.01 | 0.02 | 0.15 |
| | (0.13) | (0.04) | (0.04) | (0.04) | (0.40) |
| South Atlantic Division | 0.11 | -0.01 | 0.14 | 0.22 | -0.18 |
| | (0.03) | (0.03) | (0.03) | (0.03) | (0.16) |
| East South Central Division | 0.09 | 0.03 | 0.27 | 0.41 | -0.13 |
| | (0.02) | (0.04) | (0.04) | (0.04) | (0.12) |
| West South Central Division | 0.12 | -0.04 | 0.20 | 0.30 | -0.29 |
| | (0.02) | (0.04) | (0.04) | (0.03) | (0.11) |
| Mountain Division | 0.01 | 0.20 | 0.14 | 0.18 | 0.21 |
| | (0.08) | (0.05) | (0.06) | (0.06) | (0.30) |
| Pacific Division | 0.04 | 0.23 | 0.21 | 0.11 | 0.08 |
| | (0.05) | (0.04) | (0.04) | (0.04) | (0.23) |

*Note*: The table reports posterior means for 9 US divisions, achieved by 30000 draws (3000 burn-in). Estimated standard errors are reported in parentheses.

Posterior results for the degree of endogeneity, $\rho$, for all divisions are reported in Table 5.10. Similar to the state results, 90% intervals for posterior $\rho$ contain 0 for all divisions but West South Central Division. Hence, except for the West South Central Division, the degree of endogeneity may not be severe for all divisions. We note that the degree of endogeneity may be different, as posterior $\rho$ values for the rest of the southern regions may be around 0.

Figure 5.3: Income-education effects in US divisions: Boxplots for income effects and degree of endogeneity




Posterior $\beta$ values Posterior $\rho$ values

*Note*: The divisions in the figures correspond to Table 5.8.

Table 5.10: Income-education effects in US divisions: percentiles for posterior $\rho$ densities

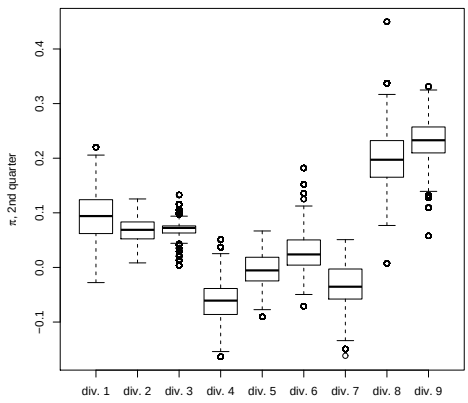
| *Percentiles* | 5% | 50% | 95% |
|---|---|---|---|
| New England Division | -0.54 | -0.11 | 0.20 |
| Middle Atlantic Division | -0.54 | 0.04 | 0.55 |
| East North Central Division | -0.19 | 0.39 | 0.68 |
| West North Central Division | -0.47 | 0.22 | 0.78 |
| South Atlantic Division | -0.43 | -0.16 | 0.10 |
| East South Central Division | -0.31 | -0.13 | 0.08 |
| West South Central Division | -0.46 | -0.28 | -0.12 |
| Mountain Division | -0.34 | 0.20 | 0.68 |
| Pacific Division | -0.24 | 0.07 | 0.50 |

*Note*: The table reports posterior results for $\rho$ for the IV model for 9 US divisions, achieved by 30000 draws (3000 burn-in).

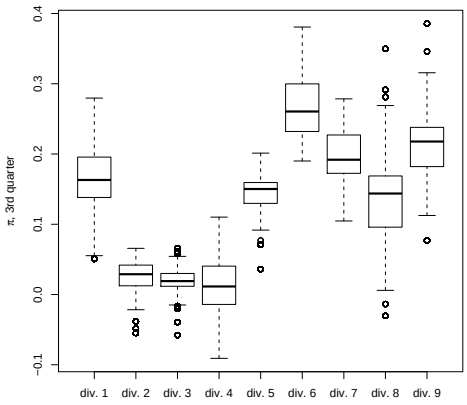
Table 5.11 reports achieved model probabilities for the model without instruments for all divisions, where the training samples consist of 25% of the observations[10]. Model

[10]We experimented with the model using smaller training samples, and the results are quite insensitive to the training sample size.

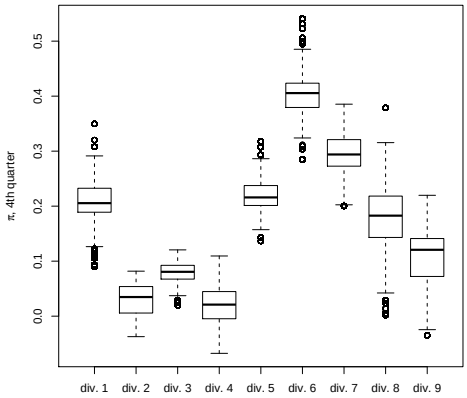
Figure 5.4: Income-education effects in US divisions: Boxplots for the quarter of birth effects on education




Posterior $\Pi_2$ values




Posterior $\Pi_3$ values




Posterior $\Pi_4$ values

*Note*: The divisions in the figures correspond to Table 5.8

probability for the nested model without instruments is far from 0.5 only for two regions: East North Central Division, and West South Central Division. In East North Central Division, the model without instruments is favored by model probabilities. In West South Central Division however, the IV model is favored according to model probabilities. This division includes states with relatively stronger instruments such as Arkansas and Texas. However, we do not find a direct link between instrument strength and preference for the IV model. For example in other divisions which include states with relatively strong instruments, such as East South Central division, we do not see a clear preference for the IV model.

Table 5.11: Income-education effects in US divisions: predictive model probabilities for the restricted model in which $\beta$, the effect of education on income, is estimated without making use of the instruments

| | probability of $M_0$ |
|---|---|
| New England Division | 0.66 |
| Middle Atlantic Division | 0.66 |
| East North Central Division | 0.85 |
| West North Central Division | 0.48 |
| South Atlantic Division | 0.52 |
| East South Central Division | 0.56 |
| West South Central Division | 0.06 |
| Mountain Division | 0.70 |
| Pacific Division | 0.54 |

*Note*: The table reports predictive model probabilities ratios for 9 US divisions, achieved by 30000 draws (3000 burn-in). For each division, the training sample size is 25% of the full sample size, and the training sample is selected by random draws from the data.

We conclude that the US divisions also show substantial heterogeneity in terms of the strength of instruments and degrees of endogeneity. Apart from two regions, model comparison results in terms of the necessity of instruments are similar to the individual state results: For most divisions, it is hard to choose one of the alternative models, hence averaging over the alternative models is likely to provide gains in efficiency in this dataset as well.

The results of model averaging are shown in Table 5.12, where we report the posterior results for the effects of education on income for BMA using model probabilities reported in Table 5.11, together with the posterior results for the IV model and the model without instruments. Table 5.12 shows that posterior means for $\beta$ are not very different across alternative models and model averaging results, except for the East North Central Division. The main advantage of model averaging is seen in the considerably smaller posterior standard deviations achieved by combining the alternative models.

For the US data on the income-education relationship, we conclude that there is substantial heterogeneity in the income-education relationship across states and regions. We document that differences between states and regions are characterized by differing instru-

ment strengths, as reported by Hoogerheide and Van Dijk (2006). Despite this finding, our results show that the degree of endogeneity is also different across states and regions.

Table 5.12: Income-education effects in US divisions: posterior $\beta$ for the IV model, model without instruments, and BMA

| | | Posterior $\beta$ summary | |
|---|---|---|---|
| Division | Model | Mean | Std. Dev. |
| New England Division | $M_0$ | 0.08 | 0.00 |
| | average | 0.09 | 0.02 |
| | $M_1$ | 0.11 | 0.05 |
| Middle Atlantic Division | $M_0$ | 0.07 | 0.00 |
| | average | 0.07 | 0.02 |
| | $M_1$ | 0.07 | 0.07 |
| East North Central Division | $M_0$ | 0.06 | 0.00 |
| | average | 0.05 | 0.01 |
| | $M_1$ | -0.03 | 0.08 |
| West North Central Division | $M_0$ | 0.06 | 0.00 |
| | average | 0.04 | 0.07 |
| | $M_1$ | 0.02 | 0.13 |
| South Atlantic Division | $M_0$ | 0.07 | 0.00 |
| | average | 0.09 | 0.02 |
| | $M_1$ | 0.11 | 0.03 |
| East South Central Division | $M_0$ | 0.07 | 0.00 |
| | average | 0.08 | 0.01 |
| | $M_1$ | 0.09 | 0.02 |
| West South Central Division | $M_0$ | 0.07 | 0.00 |
| | average | 0.12 | 0.02 |
| | $M_1$ | 0.12 | 0.02 |
| Mountain Division | $M_0$ | 0.06 | 0.00 |
| | average | 0.04 | 0.02 |
| | $M_1$ | 0.01 | 0.08 |
| Pacific Division | $M_0$ | 0.06 | 0.00 |
| | average | 0.05 | 0.03 |
| | $M_1$ | 0.04 | 0.05 |

*Note*: The table reports posterior means and standard deviations for $\beta$ for 9 US divisions. IV model results, denoted by $M_1$, are based on 30000 draws (3000 burn-in draws).

The dataset shows differing, and mostly weak power of quarter of birth in explaining education. This finding, in combination with the not so severe problem of endogeneity makes it hard to assess whether IV model should be preferred over a simpler and more parsimonious linear regression model without instruments. Hence we conclude that averaging over these alternative models is a reasonable way to deal with model uncertainty. Finally, we show that the combination of SDDR and predictive likelihoods provides a straightforward method to assess the degree of endogeneity and the relevance of the IV model.

## 5.4 IV model with plausible exogeneity and multiplicative effects of covariates

In this section, we analyze a different dataset on the income-education relationship, where the available instrument for the explanatory endogenous variable, years of education, is the level of individuals' father's education. Specifically, we apply the combination of SDDR and predictive likelihoods in model comparison, to assess the possibility of differing effects of education on income depending on certain covariates.

The data is taken from German Socio-Economic Panel Survey (SOEP) at the German Institute for Economic Research (DIW) Berlin, and was analyzed recently in Hoogerheide *et al.* (2010). The data contains individuals' annual income, number of hours worked, years of education, gender, fathers' education, and several other factors that possibly affect income, for the period between 1984–2004. We analyze a cross sectional subset of these panel data for the period 1984–2004 by using the first observation available for each individual, providing 17196 observations. Alternatively, the panel dataset including all observations can also be analyzed. In this case correlation between individuals' data over time has to be taken into account. For simplicity, we do not consider this extension. Note that a fixed effects model would not be useful, as education level is constant for most individuals over time; further, the validity of a simple random coefficient model is typically rejected. The use of cross-sectional data circumvents the problems that arise with using such panel datasets, which are outside the scope of this chapter and could harm the validity of estimation results. This allows us to focus on the core issue of family background variables as possibly endogenous instruments for education.

We consider the income-education relationship for the German SOEP data, focusing on two aspects: possible multiplicative effects of certain covariates, and possible direct effects of father's education on income. We further take into account possible endogeneity

ofinstruments, as recently suggested by Conley *et al.* (2008). The IV model incorporating *plausible exogeneity* and multiplicative effects of control variables on education effects is summarized in Section 5.4.1.

### 5.4.1 IV model with plausible endogeneity and heterogenous marginal effects

The IV model with possible direct effects of instruments on income (Conley *et al.*, 2008) and multiplicative effects of (some) conditioning variables and education is:

$$y_i = x_i\beta_1 + x_i u_i'\beta_2 + z_i^*\gamma_1 + z_i^* u_i'\gamma_2 + w_i'\delta_1 + \epsilon_i \tag{5.30}$$

$$, x_i = z_i'\delta_2 + v_i, \tag{5.31}$$

for $i = 1, \ldots n$, where for individual $i$, $y_i$ is the natural logarithm of hourly income; $x_i$ is the endogenous variable, years of education; $z_i^*$ is the instrument, father's years of secondary education; $u_i$ are exogenous variables that affect income both directly, and through returns to education; $w_i$ consists of the control variables; $z_i = \begin{pmatrix} z_i^* & w_i \end{pmatrix}$ consists of the instruments ($z_i^*$) and control variables ($w_i$); $\epsilon_i$ and $v_i$ are the error terms.

In the IV model in (5.30) and (5.31), error terms are independently normally distributed: $(\epsilon_i, v_i)' \sim NID\,(0, \Sigma_{\epsilon,v})$. Further, control variables $w$ (of which a subset forms $u$) and instruments, which together form $z$, are assumed to be independent of the error terms in both equations. Note that the model with exact exogeneity of instruments is a special case of (5.30) and (5.31) with $\gamma_1 = \gamma_2 = 0$.

We define variables with multiplicative effects on education as follows: $u_i = \begin{pmatrix} D_i^1 & D_i^2 \end{pmatrix}$, where $D_i^1$ takes the value 1 for male individuals, and 0 for female individuals; $D_i^2$ takes the value 1 if the survey was conducted after year 2000, 0 otherwise[11]. Following Hoogerheide *et al.* (2010), control variables $w_i$ include years of work experience (linear and squared term), (log) income from assets, marital status, unemployment duration before employment, and a set of dummy variables for male individuals, living in West Germany, German nationals, industries, and annual dummy variables for the period 1985–2004. Equation (5.31) can be generalized including $u_i$ and $z_i^* u_i$ on the right hand side, hence allowing for mean education levels and the effect of father's education on the individual's education to differ across time periods and genders. For simplicity we do not consider this extension.

[11]The cut-off point for the year dummy can be estimated together with the model parameters. For now, we consider the simple case where the timing of the change in education returns to income is fixed.

Define $\tilde{x}_i = \begin{pmatrix} x_i & x_i u_i' \end{pmatrix}$, $\beta = \{\beta_1, \beta_2'\}$, $\tilde{z}_i^* = \begin{pmatrix} z_i^* & z_i^* u_i' \end{pmatrix}$ and $\gamma = \begin{pmatrix} \gamma_1 & \gamma_2' \end{pmatrix}'$. Equation (5.30) becomes:

$$y_i = \tilde{x}_i \beta + \tilde{z}_i^* \gamma + w_i' \delta_1 + \epsilon_i. \tag{5.32}$$

We assume that $\gamma$ is proportional to $\beta$, i.e. there is a scalar factor $\tilde{\gamma}$ such that $\gamma = \beta\tilde{\gamma}$. Hence equation (5.32) becomes:

$$y_i = \tilde{x}_i \beta + \tilde{z}_i^* \beta\tilde{\gamma} + w_i' \delta_1 + \epsilon_i, \tag{5.33}$$

for $i = 1, \ldots, n$.

Let $\Psi_{\epsilon,v} = \Sigma_{\epsilon,v}^{-1}$, and $\theta = \{\beta, \gamma, \delta_1, \delta_2, \Psi_{\epsilon,v}\}$. Likelihood of the model in (5.33) and (5.31) is:

$$\begin{aligned} p(y, x \mid w, z, \theta) = (2\Pi)^{-n/2} \mid \Psi_{\epsilon,v} \mid^{n/2} \\ \times \exp\left( -\frac{1}{2} \sum_{i=1}^{n} \begin{pmatrix} y_i - \tilde{x}_i'\beta - \tilde{z}_i^*\beta\tilde{\gamma} - w_i\delta_1 \\ x_i - z_i\delta_2 \end{pmatrix}' \Psi_{\epsilon,v} \begin{pmatrix} y_i - \tilde{x}_i'\beta - \tilde{z}_i^*\beta\tilde{\gamma} - w_i\delta_1 \\ x_i - z_i\delta_2 \end{pmatrix} \right). \end{aligned} \tag{5.34}$$

We specify a proper prior for $\tilde{\gamma}$ as it is necessary for identification, and consider two prior specifications for $\tilde{\gamma}$ where the first is mentioned by Conley *et al.* (2008) and the second is an extension:

(i) a normal prior distribution $\tilde{\gamma} \sim N\left(\underline{\mu}_{\tilde{\gamma}}, \underline{\sigma}_{\tilde{\gamma}}^2\right)$,

(ii) a truncated normal distribution $\tilde{\gamma} \sim TN_{\tilde{\gamma} \subset \boldsymbol{A}}\left(\underline{\mu}_{\tilde{\gamma}}, \underline{\sigma}_{\tilde{\gamma}}^2\right)$, that is, $N\left(\underline{\mu}_{\tilde{\gamma}}, \underline{\sigma}_{\tilde{\gamma}}^2\right)$ truncated to a space $\boldsymbol{A}$. For example, $\boldsymbol{A}$ can be defined as $(0, \infty)$.

We define the following priors for the rest of the parameters: flat prior for $\delta_1, \delta_2 \propto 1$, uninformative proper normal priors for $\beta$, $\beta \sim N\left(\underline{\mu}_\beta, \underline{\Sigma}_\beta\right)$, for $\Psi$ an uninformative, limiting case of the Wishart distribution: $\Psi \propto \mid \Psi \mid^{-3/2}$. Hence the joint prior for $\theta$ is:

$$p(\theta) \propto \exp\left( -\frac{1}{2} \left(\beta - \underline{\mu}_\beta\right)' \underline{\Sigma}_\beta^{-1} \left(\beta - \underline{\mu}_\beta\right) \right) \mid \Psi \mid^{-3/2}. \tag{5.35}$$

We diverge from the Jeffreys priors used in Section 5.3, as the general IV model proposed for the SOEP data allows for multiplicative effects of certain covariates with level of education. For this extended IV model, the use of alternative, proper but non-informative prior has the advantage of facilitating the use of Gibbs sampler. Moreover, Jeffreys prior is not found to have an advantage in this dataset: For a benchmark model with no multiplicative effects, posterior results under prior (5.35) and Jeffreys priors were similar. In both cases, posterior draws for the effect of father's education on own education are above 0. This result is in line with the literature on the high explanatory

power of family background variables as instruments. The data does not suffer from weak instruments, hence employing Jeffreys priors does not have a clear advantage over using the non-informative normal prior.

Combining (5.34) and (5.35), the posterior density kernel is:

$$p(\theta \mid y, x, w, z) \propto \mid \Psi \mid^{(n-3)/2} \exp\left(-\frac{1}{2}\left(\beta - \underline{\mu}_\beta\right)' \underline{\Sigma}_\beta^{-1}\left(\beta - \underline{\mu}_\beta\right)\right) \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\begin{pmatrix} y_i - \tilde{x}_i'\beta - \tilde{z}_i^*\beta\tilde{\gamma} - w_i\delta_1 \\ x_i - z_i\delta_2 \end{pmatrix}' \Psi_{\epsilon,v}\begin{pmatrix} y_i - \tilde{x}_i'\beta - \tilde{z}_i^*\beta\tilde{\gamma} - w_i\delta_1 \\ x_i - z_i\delta_2 \end{pmatrix}\right). \tag{5.36}$$

In order to draw parameters from the posterior density in (5.36), we apply the Gibbs sampler of (Rossi *et al.*, 2005, ch. 5), augmented with a step for simulating $\tilde{\gamma}$. The posterior simulator is explained in detail in 5.A.1.

### 5.4.2 Income-education relationship in the SOEP data

For the IV model in (5.30) and (5.31), we analyze possible effects of gender and the time dummy on income returns to education using two alternative specifications: equal marginal effects of education on income regardless of gender, and equal marginal effects regardless of the time period considered. That is, we test the two hypotheses of 'no difference between genders' and 'no difference between time periods' separately, at both instances allowing the other difference to be present in the model.

In (5.30) and (5.31), we define $\beta_2 = \begin{pmatrix}\beta_2^m & \beta_2^t\end{pmatrix}$ as the coefficients of the male and time dummy variables multiplied by education. The corresponding alternative models are: $M_0^m : \beta_2^m = 0$, $M_1^m : \beta_2^m \neq 0$; and $M_0^t : \beta_2^t = 0$, $M_1^t : \beta_2^t \neq 0$ respectively. For both model comparisons, we assume equal prior model probabilities, hence predictive model probabilities boil down to estimating the marginal posterior densities of $\beta_2^m$ and $\beta_2^t$ evaluated at point 0 for the training sample and the full sample. For $\beta$, we choose a normal uninformative prior, $\beta \sim N(0, I_3)$.

We consider three models with differing instrument exogeneity assumptions: a model with exact exogeneity of instruments ($\tilde{\gamma} = 0$), a model allowing for a *small* direct effect of instruments on income defining the prior $\tilde{\gamma} \sim N(0, 0.1)$, and the model allowing for a *small*, *typically positive* direct effect of instruments on income having the same sign as own education's effect, using the prior $\tilde{\gamma} \sim TN_{\tilde{\gamma}>0}(0, 0.1)$.

Table 5.13 reports predictive model probabilities achieved by SDDR and predictive likelihoods for the restricted models under different exogeneity specifications and three training samples consisting of 5%, 10% and 25% of the data. Model probabilities are

quite insensitive to the training sample choice. Hence combining SDDR and predictive likelihoods is found to provide a robust method for model comparison in this dataset as well.

Table 5.13: Predictive model probabilities for the nested models: constant returns to education for male and female respondents and constant returns to education over the sample period

| model | training sample percentage | $M_0^t$ | $M_0^m$ |
|---|---|---|---|
| exact exogeneity | 5% | 0.000 | 0.257 |
| | 10% | 0.000 | 0.128 |
| | 25% | 0.001 | 0.202 |
| plausible exogeneity* | 5% | 0.000 | 0.229 |
| | 10% | 0.000 | 0.124 |
| | 25% | 0.000 | 0.120 |
| plausible exogeneity, $\tilde{\gamma} > 0$ | 5% | 0.061 | 0.246 |
| | 10% | 0.004 | 0.123 |
| | 25% | 0.004 | 0.153 |

*Note*: The table reports predictive model probabilities under different training samples for two restricted models: $M_0^t$ denotes the model where the effect of education on income is the same across time periods; $M_0^m$ denotes the model where the effect of education on income is the same across males and females. Posterior results in all cases but (*) are based on 10000 draws (1000 burn-in). For (*) we use 50000 draws (5000 burn-in) draws since the Gibbs sampler for this model was relatively slower to converge.

First, regarding the time dummy, Table 5.13 shows that model probabilities for the restricted model, with no change in returns to education, are very close to zero. Second, for the restricted model with equal effects of education on income for male and female individuals, model probabilities in all cases are around 0.2, indicating different effects of education on income between male and female individuals.

Next, we consider the effects of education on income according to time and male dummies for different model specifications. Posterior results for the effects of education are reported in Table 5.14.

Table 5.14: Posterior results for the effects of education on income

| | Mean and standard deviation | | Percentiles of posterior distribution | |
|---|---|---|---|---|
| | Mean | Std. Dev. | 2.5% | 97.5% |
| *exact exogeneity* | | | | |
| education | 0.0724 | 0.0055 | 0.0615 | 0.0831 |
| education×time | 0.0163 | 0.0034 | 0.0096 | 0.0230 |
| education×male | -0.0073 | 0.0029 | -0.0128 | -0.0016 |
| *plausible exogeneity** | | | | |
| education | 0.0827 | 0.0142 | 0.0573 | 0.1111 |
| education×time | 0.0169 | 0.0036 | 0.0099 | 0.0237 |
| education×male | -0.0073 | 0.0029 | -0.0131 | -0.0017 |
| *plausible exogeneity,* $\tilde{\gamma} > 0$ | | | | |
| education | 0.0659 | 0.0070 | 0.0522 | 0.0792 |
| education×time | 0.0160 | 0.0034 | 0.0092 | 0.0227 |
| education×male | -0.0072 | 0.0028 | -0.0128 | -0.0017 |

*Note*: The table reports posterior results for the effects of education on income under the assumptions of exogenous instruments and alternatively under the assumption of possible direct effects of instruments. Posterior results are based on 10000 draws (1000 burn-in). For the model shown by * we use 50000 draws (5000 burn-in) draws since the Gibbs sampler for this model was considerably slow to converge.

Posterior 95% intervals for the effect of time dummy are above zero, indicating higher returns to education for the period 2001–2004 compared to the period 1984–2000. Regarding the gender effects, we find lower returns to education for men compared to women.

Notice that the male dummy is also included as a control variable in the model, and the posterior results for the linear effect of this control variable on income are above 0[12]. Hence higher returns to education for women only indicates a mitigating effect of higher education on the existing gender wage gap. The wage gap between men and women are lower for higher levels of education.

Finally, we note that allowing for possible direct effects of father's education on income does not alter the education effects on income substantially. Specifically, posterior results

[12]Posterior results for the linear gender effect, together with the effects of other control variables are reported in Table 5.A.2.1.

for returns to education are very close for the cases of exact exogeneity and positive direct effects of father's education on income. For the case with unrestricted direct effects of father's education, the posterior standard deviations for returns to education are only slightly higher.

We conclude that for the German SOEP data, a more general IV model should be employed, allowing for different returns to education according to gender. Regarding the possible endogeneity of instruments,we show that the results concerning income-education effects are rather insensitive to the possibility of endogenous instruments, i.e. assuming that with 95% confidence the direct effect of father's education is smaller than 20% of own education's effect on income. Furthermore, the results suggest that the combination of SDDR and predictive likelihoods provides a simple but appropriate method to check the existence of such differing effects in the Bayesian context.

## 5.5 Conclusion

We present a general framework for model comparison in the standard Bayesian treatment of IV models under non-informative priors, where the sensitivity of model probabilities to the prior choice can be avoided using predictive likelihoods. The method is applied to simulated datasets, and two different datasets on the income-education relationship. We show that this method can be used to assess some of the highly debated issues in IV models, such as the trade-off between choosing the correct model and precision, and differing effects of the endogenous explanatory variable across subsamples of data.

The empirical contribution of this work is two-fold. First, we show that the income-education relationship in the US states and divisions, using quarter of birth of individuals to form instruments, is subject to significant heterogeneity in terms of the degree of endogeneity in education levels, as well as the strength of instruments. For these data, we conclude that the precision losses from employing an IV model can be quite severe. We rather propose a model averaging approach to infer returns to education, for which the weights in model averaging take into account the degree of endogeneity and the instrument strength.

Second, we analyze the income-education relationship for the German SOEP data, where education levels are instrumented by father's education. The model comparison technique we employ indicates that returns to education are different across subsamples of data. In particular, we find decreasing gender discrimination in earnings for individuals with higher education, and higher returns to education for the recent time period.

Future work will be on robustness checks for the proposed method under more flexible distributions for the disturbances, and the sensitivity of the results to the different prior choices, such as flat priors and the hierarchical prior proposed by Chamberlain and Imbens (1996). Further, we will consider Bayesian model averaging over models with different structural break specifications, allowing for different times for structural breaks and more structural breaks.

# 5.A Appendices

## 5.A.1 Derivations of the conditional posterior densities for the IV model with plausible endogeneity and multiplicative covariate effects

For the IV model in (5.30) and (5.31), under the prior in (5.35), conditional posterior densities for parameters $\theta = \{\beta, \gamma, \delta_1, \delta_2, \Psi_{\epsilon,v}\}$ are derived using Rossi *et al.* (2005), chapter 5, and Conley *et al.* (2008).

Let $\theta_{-\nu}$ denote the set of all parameters in $\theta$ except for $\nu$. Conditional distributions for the Gibbs sampler are as follows:

(i) $\Psi_{\epsilon,v} \mid y, x, w, z, \theta_{-\Psi_{\epsilon,v}} \sim \text{Wishart}\,(n, V_\Psi)$ with

$$V_\Psi = \left[\sum_{i=1}^n \begin{pmatrix} y_i - \tilde{x}_i'\beta - \tilde{z}_i^*\beta\tilde{\gamma} - w_i\delta_1 \\ x_i - z_i\delta_2 \end{pmatrix} \begin{pmatrix} y_i - \tilde{x}_i'\beta - \tilde{z}_i^*\beta\tilde{\gamma} - w_i\delta_1 \\ x_i - z_i\delta_2 \end{pmatrix}'\right]^{-1} \tag{5.37}$$

(ii) $(\beta, \gamma_1')' \mid y, x, w, z, \theta_{-(\beta,\gamma_1)} \sim N\,(\mu_{\beta,\gamma_1}, V_{\beta,\gamma_1})$ with

$$V_{\beta,\gamma_1} = \left[\begin{pmatrix} \underline{\Sigma}_\beta^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \left(\sigma^2_{\epsilon|v}\right)^{-1} \sum_{i=1}^n \begin{pmatrix} (\tilde{x}_i + \tilde{z}_i^*\tilde{\gamma})\,(\tilde{x}_i + \tilde{z}_i^*\tilde{\gamma})' & (\tilde{x}_i + \tilde{z}_i^*\tilde{\gamma})\,\tilde{w}_i' \\ \tilde{w}_i\,(\tilde{x}_i + \tilde{z}_i^*\tilde{\gamma})' & \tilde{w}_i\tilde{w}_i' \end{pmatrix}\right]^{-1} \tag{5.38}$$

$$\mu_{\beta,\gamma_1} = V_{\beta,\gamma_1}\left[\begin{pmatrix} \underline{\Sigma}_\beta^{-1}\underline{\mu}_\beta \\ 0 \end{pmatrix} + \left(\sigma^2_{\epsilon|v}\right)^{-1} \sum_{i=1}^n \begin{pmatrix} \tilde{x}_i \\ w_i \end{pmatrix} \left(y_i - \mu_{\epsilon_i|v_i}\right)\right] \tag{5.39}$$

where $\mu_{\epsilon_i|v_i} = (x_i - z_i'\delta_2)\,\sigma_{\epsilon,v}/\sigma_v^2$, $\sigma_{\epsilon|v} = \sigma_\epsilon^2 - \sigma_{\epsilon,v}^2/\sigma_v^2$ with

$$\Sigma_{\epsilon,v} = \begin{bmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon,v} \\ \sigma_{\epsilon,v} & \sigma_v^2 \end{bmatrix}.$$

(iii) $\delta_2 \mid y, x, w, z, \theta_{-\delta_2} \sim N\,(\mu_{\delta_2}, V_{\delta_2})$, with

$$V_{\delta_2} = \left[\left(\sigma^2_{v|\epsilon}\right)^{-1} \sum_{n=1}^n z_i z_i'\right]^{-1} \tag{5.40}$$

$$\mu_{\delta_2} = V_{\delta_2}\left[\left(\sigma_{v|\epsilon}\right)^{-1} \sum_{i=1}^n z_i\left(x_i - \mu_{v_i|\epsilon_i}\right)\right] \tag{5.41}$$

where $\mu_{v_i|\epsilon_i} = (y_i - \tilde{x}'\beta - w_i'\delta_1)\,\sigma_{\epsilon,v}/\sigma_\epsilon^2$ and $\sigma^2_{v|\epsilon} = \sigma_v^2 - \sigma^2_{v,\epsilon}/\sigma_\epsilon^2$.

(iv) $\tilde{\gamma} \mid y, x, w, z, \theta_{-\tilde{\gamma}} \sim N\,(\mu_{\tilde{\gamma}}, V_{\tilde{\gamma}})$, with

$$V_{\tilde{\gamma}} = \left[\underline{\sigma}_{\tilde{\gamma}}^{-1} + \left(\sigma^2_{\epsilon|v}\right)^{-1} \sum_{i=1}^n \left(\tilde{z}_i^*\beta\right)^2\right] \tag{5.42}$$

$$\mu_{\tilde{\gamma}} = V_{\tilde{\gamma}}\left[\underline{\sigma}_{\tilde{\gamma}}^{-1}\underline{\mu}_{\tilde{\gamma}}^{-1} + \left(\sigma^2_{\epsilon|v}\right)^{-1} \sum_{i=1}^n \begin{pmatrix} \tilde{x}_i \\ w_i \end{pmatrix} \left(y_i - \mu_{\epsilon_i|v_i}\right)\right] \tag{5.43}$$

under the prior $\tilde{\gamma} \sim N\left(\underline{\mu}_{\tilde{\gamma}}, \underline{\sigma}^2_{\tilde{\gamma}}\right)$.

For the truncated prior, this step becomes:

$$\tilde{\gamma} \mid y, x, w, z, \theta_{-\tilde{\gamma}} \sim TN_{\tilde{\gamma} \subset \boldsymbol{A}}\left(\mu_{\tilde{\gamma}}, V_{\tilde{\gamma}}\right)$$

For the benchmark model under the assumption of exact exogeneity of instruments, conditional densities (i) to (iii) hold with the restriction $\tilde{\gamma} = 0$, conditional density of $\tilde{\gamma} = 0$ in step (iv) is ignored in Gibbs sampling.

## Appendix A.2 German SOEP data: posterior results for the effects of conditioning variables in the IV model

Table 5.A.2.1 reports the posterior results for the effects of conditioning variables on income for the IV model under the assumption of exact exogeneity of instruments ($\tilde{\gamma} = 0$). The results with plausibly exogenous instruments are very similar, and hence are not reported in detail.

Table 5.A.2.1: Posterior results for the effects of conditioning variables on income for the IV model in case of exact exogeneity of the instrument

| | Mean | Std. Dev. | 2.5% | 97.5% |
|---|---|---|---|---|
| Experience | 0.03 | 0.01 | 0.02 | 0.04 |
| Experience$^2$/10 | -0.01 | 0.00 | -0.02 | -0.01 |
| Non-German | -0.91 | 0.08 | -1.06 | -0.76 |
| Married | 0.28 | 0.04 | 0.20 | 0.36 |
| Unemployment years | -0.15 | 0.02 | -0.19 | -0.12 |
| Wealth from assets | 0.16 | 0.01 | 0.13 | 0.18 |
| West Germany | -0.29 | 0.05 | -0.38 | -0.19 |
| Male | 0.53 | 0.04 | 0.46 | 0.61 |
| *Industry dummy variables*** | | | | |
| Agriculture | -1.56 | 0.14 | -1.83 | -1.28 |
| Manufacturing | -1.31 | 0.05 | -1.41 | -1.21 |
| Hotel and restaurant | -2.01 | 0.12 | -2.24 | -1.77 |
| Firm services | -0.61 | 0.07 | -0.76 | -0.47 |
| Financial sector | -0.59 | 0.09 | -0.76 | -0.41 |
| Construction | -1.72 | 0.07 | -1.86 | -1.57 |
| Transportation | -1.54 | 0.10 | -1.74 | -1.35 |
| Health | -0.72 | 0.07 | -0.85 | -0.59 |
| Culture | -0.58 | 0.16 | -0.88 | -0.26 |
| Sports | -1.64 | 0.06 | -1.77 | -1.52 |

continued on Next Page...

Table 5.A.2.1 continued...

| | Mean | Std. Dev. | 2.5% | 97.5% |
|---|---|---|---|---|
| *Year dummy variables*** | | | | |
| 1985 | -0.31 | 0.09 | -0.49 | -0.14 |
| 1986 | -0.41 | 0.13 | -0.65 | -0.16 |
| 1987 | -0.11 | 0.14 | -0.40 | 0.17 |
| 1988 | -0.16 | 0.15 | -0.47 | 0.13 |
| 1989 | -0.32 | 0.12 | -0.56 | -0.09 |
| 1990 | -0.37 | 0.15 | -0.67 | -0.08 |
| 1991 | -0.56 | 0.15 | -0.85 | -0.27 |
| 1992 | 0.31 | 0.10 | 0.12 | 0.51 |
| 1993 | 0.12 | 0.11 | -0.11 | 0.34 |
| 1994 | 0.26 | 0.13 | -0.01 | 0.51 |
| 1995 | 0.22 | 0.12 | -0.01 | 0.46 |
| 1996 | 0.29 | 0.16 | -0.02 | 0.61 |
| 1997 | 0.12 | 0.14 | -0.15 | 0.40 |
| 1998 | 0.31 | 0.11 | 0.09 | 0.53 |
| 1999 | 0.25 | 0.15 | -0.04 | 0.55 |
| 2000 | 0.36 | 0.08 | 0.19 | 0.52 |
| 2001 | 0.22 | 0.11 | 0.00 | 0.44 |
| 2002 | 1.36 | 0.09 | 1.16 | 1.54 |
| 2003 | 0.56 | 0.13 | 0.30 | 0.81 |
| 2004 | 0.57 | 0.12 | 0.34 | 0.81 |

*Note*: The table reports posterior parameters for the IV model under exact exogeneity of instruments ($\tilde{\gamma} = 0$), achieved by 10000 iterations (1000 burn-in iterations).

* Education variable is instrumented by father's education.

** Reference category for industry is *other*.

*** Reference category for year is 1984.

# Nederlandse Samenvatting (Summary in Dutch)

De keuze voor een bepaald statistisch model in kwantitatieve economische analyses wordt bepaald door de achterliggende onderzoeksvraag en de specifieke structuur van de te analyseren data. Het vermijden van sterke aannamen in het model is cruciaal om te komen tot consensus tussen resultaten van studies en voor de betrouwbaarheid en accuraatheid van de resultaten.

Een mogelijke methode om het maken van sterke aannamen in het modelleren te vermijden is door gebruik te maken van flexibele modellen die relatief meer nadruk leggen op de data in plaats van te vertrouwen op a priori specificaties. Dergelijke flexibele modellen worden in verschillende gebieden van economisch onderzoek toegepast. Een specifiek voorbeeld hiervan is de analyse van de economische vooruitgang van landen, gemeten aan de hand van het reële Bruto Binnenlands Product (BBP) per hoofd van de bevolking, waarvoor het goed gedocumenteerd is dat patronen in het BBP van landen een sterke mate van heterogeniteit vertonen in zowel het niveau als ook de groei.

Een tweede methode om de effecten van onderliggende modelaannamen te reduceren is om verschillende modellen, die mogelijkerwijs geschikt zijn voor de data, te beschouwen en de mate te bepalen waarin ieder van deze modellen de data adequaat beschrijft. Vervolgens kunnen deze modellen gecombineerd worden om expliciet rekening te houden met modelonzekerheid. In de economie wordt deze benadering vooral gebruikt bij het construeren van voorspellingen.

Dit proefschrift bestaat uit twee delen. Het eerste deel worden nieuwe econometrische modellen met een voldoende mate van flexibiliteit ontwikkeld om verschillende vormen en mate van heterogeniteit in (de relaties tussen) economische variabelen te beschrijven. In het tweede deel worden nieuwe gereedschappen ontwikkeld om te beoordelen in hoeverre modellen geschikt zijn voor het beschrijven van de data.

Het eerste deel van dit proefschrift, bestaande uit de hoofdstukken 2 en 3, beschrijft nieuwe statistische modellen voor het analyseren van de economische groei van landen.

Verschillende studies documenteren dat de aanname van homogeniteit in economische groeirelaties onrealistisch is. Landen vertonen verschillen in groeipatronen, en er zijn verschillende factoren bepalend voor de economische ontwikkeling. Deze twee hoofdstukken ontwikkelen nieuwe methoden om verschillen in groei tussen landen te analyseren en de oorzaken van deze verschillen te verklaren.

Het tweede deel van dit proefschrift, bestaande uit de hoofdstukken 4 en 5, richt zich op problemen gerelateerd aan het beoordelen van de beschrijvingskracht van statistische modellen en het vergelijken van beschrijvingskracht van modellen, met toepassingen op verschillende datasets. In deze hoofdstukken worden nieuwe gereedschappen ontwikkeld om accuraat de beschrijvingskracht van een model te beoordelen vanuit een Bayesiaans perspectief. Hoofdstuk 4 verschaft een overzicht van de beschikbare methoden om de modelwaarschijnlijkheid te schatten, wat het hoofddoel is van het Bayesiaans vergelijken van modellen. Tevens worden de standaard methoden voor het schatten van de modelwaarschijnlijkheid verbeterd door gebruik te maken van recente simulatiemethoden om dichtheden te benaderen, (Hoogerheide *et al.*, 2007b; Meng and Schilling, 2002).

Tot slot wordt in Hoofdstuk 5 een alternatieve methode voorgesteld om modellen te evalueren die de effecten van onderwijs op inkomen analyseren. Deze effecten worden meestal geanalyseerd met Instrumentele Variabele (IV) modellen. In plaats van de conventionele modelwaarschijnlijkheid voor model evaluatie wordt in dit hoofdstuk gekeken naar de voorspelwaarschijnlijkheid. Het doel van het toepassen van deze aanpak voor modelwaardering is om te laten zien dat sommige bezwaren tegen het gebruik van standaard modelwaarschijnlijkheden in IV modellen kan worden vermeden. Hierbij wordt vooral gekeken naar zaken zoals de mate van endogeniteit, sterkte van het instrument, mogelijke ongeldigheid van de instrumenten, en verschillen in effecten van onderwijs op inkomen in deelsteekproeven van de data.

# Bibliography


Acemoglu, D. (1998), Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality, *Quarterly Journal of Economics*, 113, 1055–1089.

Acemoglu, D., S. Johnson, and J. A. Robinson (2001), The Colonial Origins of Comparative Development: An Empirical Investigation, *American Economic Review*, 91, 1369–1401.

Aghion, P., P. Howitt, and D. Mayer-Foulkes (2005), The Effect of Financial Development on Convergence: Theory and Evidence, *Quarterly Journal of Economics*, 120, 173–222.

Albert, J. H. and S. Chib (1993a), Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts, *Journal of Business & Economic Statistics*, 11, 1–15.

Albert, J. H. and S. Chib (1993b), Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, 669–679.

Albouy, D. Y. (2008), The Colonial Origins of Comparative Development: An Investigation of the Settler Mortality Data.

Alfo, M., G. Trovato, and R. J. Waldmann (2008), Testing for country heterogeneity in growth models using a finite mixture approach, *Journal of Applied Econometrics*, 23, 487–514.

Angrist, J. D., G. Imbens, and D. B. Rubin (1996), Identification of causal effects using instrumental variables, *Journal of the American Statistical Association*, 91.

Angrist, J. D. and A. B. Krueger (1991), Does Compulsory School Attendance Affect Schooling and Earnings?, *The Quarterly Journal of Economics*, 106, 979–1014.

Ardia, D., N. Baştürk, L. F. Hoogerheide, and H. K. Van Dijk (2010), A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood, *Tinbergen*

*Institute Discussion Paper*, 04–059, *Computational Statistics & Data Analysis*, forthcoming.

Ardia, D., L. F. Hoogerheide, and H. K. Van Dijk (2008), *AdMit: Adaptive Mixture of Student-t Distributions in R*.

Ardia, D., L. F. Hoogerheide, and H. K. Van Dijk (2009a), Adaptive mixture of Student-t distributions as a flexible candidate distribution for efficient simulation: The R package **AdMit**, *Journal of Statistical Software*, 29, 1–32.

Ardia, D., L. F. Hoogerheide, and H. K. Van Dijk (2009b), AdMit: Adaptive Mixtures of Student-t Distributions, *The R Journal*, 1, 25–30.

Ardia, D., L. F. Hoogerheide, and H. K. Van Dijk (2009c), To bridge, to warp or to wrap: A Comparative Study of Monte Carlo Methods forEfficient Evaluation of Marginal Likelihoods, *Tinbergen Institute Discussion Paper*, 09–017.

Ardic, O. P. (2006), The Gap Between the Rich and the Poor: Patterns of Heterogeneity in the cross-country Data, *Economic Modelling*, 23, 538–555.

Ausín, M. C. and P. Galeano (2007), Bayesian estimation of the Gaussian mixture GARCH model, *Computational statistics & data analysis*, 51, 2636–2652.

Barro, R. J. (1991), Economic Growth in a Cross Section of Countries, *The Quarterly Journal of Economics*, 106, 407–443.

Barro, R. J. (1996), Determinants of Economic Growth: A Cross-Country Empirical Study, *NBER Working Papers*.

Barro, R. J. (1999), Determinants of Democracy, *Journal of Political Economy*, 107, 158–183.

Barro, R. J. and J.-W. Lee (2000), International Data on Educational Attainment: Updates and Implications, CID Working Papers 42, Center for International Development at Harvard University.

Bartlett, M. S. (1957), A comment on DV Lindley's statistical paradox, *Biometrika*, 44, 533.

Baştürk, N., L. F. Hoogerheide, and H. K. Van Dijk (2010), Measuring returns to education: Bayesian analysis using weak or possibly endogenous instrumental variables, *unpublished working paper*.

Baştürk, N., R. Paap, and D. Van Dijk (2008), Structural differences in economic growth: An endogenous clustering approach, *Tinbergen Institute Discussion Paper*, 085–4, *Applied Economics*, forthcoming.

Baştürk, N., R. Paap, and D. Van Dijk (2010), Financial development and convergence clubs, *Econometric Institute Report*, 2010–52.

Bates, D. M. and D. G. Watts (1988), *Nonlinear regression analysis and its applications*, Wiley New York.

Bates, J. M. and C. W. J. Granger (1969), The combination of forecasts, *OR*, 20, 451–468.

Baumol, W. J. (1986), Productivity Growth, Convergence, and Welfare: What the Long-run Data Show, *The American Economic Review*, 76, 1072–1085.

Bauwens, L., C. S. Bos, H. K. Van Dijk, and R. D. Van Oest (2004), Adaptive radial-based direction sampling: some flexible and robust Monte Carlo integration methods, *Journal of Econometrics*, 123, 201–225.

Bauwens, L. and J. V. K. Rombouts (2010), On marginal likelihood computation in change-point models, *Computational Statistics & Data Analysis*, forthcoming.

Beck, T. (2008), The Econometrics Of Finance And Growth, *Research Working papers*, 1, 1–45.

Beck, T., A. Demirguc-Kunt, and R. Levine (2000), A New Database on Financial Development and Structure, *The World Bank Economic Review*, 14, 597–605.

Beck, T. and R. Levine (2004), Stock Markets, Banks, and Growth: Panel Evidence, *Journal of Banking & Finance*, 28, 423–442.

Ben-David, D. (1994), Convergence Clubs and Diverging Economies, *CEPR Discussion Papers*.

Bernard, A. B. and S. N. Durlauf (1995), Convergence in International Output, *Journal of Applied Econometrics*, 10, 97–108.

Bianchi, C. and M. Menegatti (2007), On the Potential Pitfalls in Estimating Convergence by Means of Pooled and Panel Data, *Applied Economics Letters*, 14, 963–967.

Bloom, D. E., D. Canning, and J. Sevilla (2003), Geography and Poverty Traps, *Journal of Economic Growth*, 8, 355–378.

Bond, S. R., A. Hoeffler, and J. R. W. Temple (2001), GMM Estimation of Empirical Growth Models, *C.E.P.R. Discussion Papers*, 3048.

Bound, J., D. A. Jaeger, and R. M. Baker (1995), Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association*, 90, 443–450.

Bowden, R. J. and D. A. Turkington (1990), *Instrumental variables*, Cambridge Univ Press.

Bozdogan, H. (1987), Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions, *Psychometrika*, 52, 345–370.

Canova, F. (2004), Testing for Convergence Clubs in Income per Capita: a Predictive Density Approach, *International Economic Review*, 45, 49–77.

Carter, C. K. and R. Kohn (1994), On Gibbs Sampling for State Space Models, *Biometrika*, 81, 541.

Celeux, G., F. Forbes, C. Robert, and D. Titterington (2006), Deviance Information Criteria for Missing Data Models, *Bayesian Analysis*, 1, 651–674.

Chamberlain, G. and G. Imbens (1996), Hierarchical Bayes Models with Many Instrumental Variables, *NBER Technical Working Papers*.

Chen, M. H., Q. M. Shao, and J. G. Ibrahim (2000), *Monte Carlo methods in Bayesian computation*, Springer Verlag.

Chib, S. (1995), Marginal Likelihood from the Gibbs Output., *Journal of the American Statistical Association*, 90.

Chib, S. and I. Jeliazkov (2001), Marginal likelihood from the Metropolis-Hastings output, *Journal of the American Statistical Association*, 96, 270–281.

Clyde, M. and E. I. George (2004), Model uncertainty, *Statistical Science*, 19, 81–94.

Collier, P. and J. W. Gunning (1999), Explaining African Economic Performance, *Journal of Economic Literature*, 37, 64–111.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2008), Plausibly Exogenous, *unpublished working paper*.

Cuaresma, C. J. and G. Doppelhofer (2007), Nonlinearities in cross-country growth regressions: A Bayesian Averaging of Thresholds (BAT) approach, *Journal of Macroeconomics*, 29, 541–554.

Davis, L. S., A. L. Owen, and J. Videras (2009), Do All Countries Follow the Same Growth Process?, *Journal of Economic Growth*, 265–286.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38.

Dickey, J. M. (1971), The weighted likelihood ratio, linear hypotheses on normal location parameters, *The Annals of Mathematical Statistics*, 42, 204–223.

Dréze, J. H. (1976), Bayesian limited information analysis of the simultaneous equations model, *Econometrica: Journal of the Econometric Society*, 44, 1045–1075.

Dréze, J. H. (1977), Bayesian regression analysis using poly-t densities, *Journal of Econometrics*, 6, 329–354.

Durlauf, S. N. (2000), Econometric Analysis and the Study of Economic Growth : a Skeptical Perspective, Working papers 10.

Durlauf, S. N. (2007), Foreword to Special Journal of Macroeconomics Issue on Nonlinearities in Economic Growth, *Journal of Macroeconomics*, 29, 451–454.

Durlauf, S. N. and P. A. Johnson (1995a), Multiple Regimes and Cross-Country Growth Behaviour, *Journal of Applied Econometrics*, 10, 365–384.

Durlauf, S. N. and P. A. Johnson (1995b), Multiple Regimes and Cross-country Growth Behaviour, *Journal of Applied Econometrics*, 10, 365–384.

Easterly, W. and R. Levine (1997), Africa's Growth Tragedy: Policies and Ethnic Divisions, *The Quarterly Journal of Economics*, 112, 1203–1250.

Eklund, J. and S. Karlsson (2007), Forecast combination and model averaging using predictive measures, *Econometric Reviews*, 26, 329–363.

Evans, P. (1998), Using Panel Data to Evaluate Growth Theories, *International Economic Review*, 39, 295–306.

Feller, W. (1943), On a general class of contagious distributions, *The Annals of mathematical statistics*, 14, 389–400.

Ferguson, T. S. (1973), A Bayesian analysis of some nonparametric problems, *The annals of statistics*, 1, 209–230.

Friedman, M. (1992), Do Old Fallacies Ever Die?, *Journal of Economic Literature*, 30, 2129–2132.

Frühwirth-Schnatter, S. (2001), Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models, *Journal of the American Statistical Association*, 96, 194–209.

Frühwirth-Schnatter, S. (2004), Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques, *Econometrics Journal*, 7, 143–167.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer Verlag.

Frühwirth-Schnatter, S. and S. Kaufmann (2008), Model-Based Clustering of Multiple Time Series, *Journal of Business & Economic Statistics*, 26, 78–89.

Frühwirth-Schnatter, S. and H. Wagner (2008), Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling, *Computational Statistics & Data Analysis*, 52, 4608–4624.

Galor, O. (1996), Convergence? Inferences from Theoretical Models, *The Economic Journal*, 106, 1056–1069.

Gelfand, A. E. and D. K. Dey (1994), Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 501–514.

Gelman, A. and X. L. Meng (1998), Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, *Statistical Science*, 13, 163–185.

Geman, S. and D. Geman (1984a), Stochastic Relaxation, Gibbs Distributions, and the Bayesian Relation of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geman, S. and D. Geman (1984b), Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell*, 6, 721–741.

Geweke, J. (1989), Bayesian inference in econometric models using Monte Carlo integration, *Econometrica: Journal of the Econometric Society*, 1317–1339.

Geweke, J. (1999), Using simulation methods for Bayesian econometric models: inference, development, and communication, *Econometric Reviews*, 18, 1–73.

Geweke, J. (2007), Interpretation and Inference in Mixture Models: Simple MCMC Works, *Computational Statistics & Data Analysis*, 51, 3529–3550.

Geweke, J. (2010), *Complete and incomplete econometric models*, Princeton Univ Pr.

Geweke, J. and G. Amisano (2010), Comparing and evaluating Bayesian predictive distributions of asset returns, *International Journal of Forecasting*.

Geyer, C. J. (1992), Practical markov chain monte carlo, *Statistical Science*, 473–483.

Goldberger, A. S. (1972), Structural equation methods in the social sciences, *Econometrica: Journal of the Econometric Society*, 40, 979–1001.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press.

Hamilton, J. D. and M. T. Owyang (2009), The Propagation of Regional Recessions, *FRB of St. Louis Working Paper Series*, 2009–013A.

Hammersley, J. M. and D. C. Handscomb (1964), *Monte carlo methods*, Methuen.

Han, C. and B. P. Carlin (2001), Markov chain Monte Carlo methods for computing Bayes factors, *Journal of the American Statistical Association*, 96, 1122–1132.

Hansen, B. E. (2000), Sample Splitting and Threshold Estimation, *Econometrica: Journal of the Econometric Society*, 68, 575–604.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57, 97–109.

Heston, A., R. Summers, and A. B. (2009), The Penn World Table Version 6.3, *Center for International Comparisons for International Comparisons of Production, Income and Prices*.

Hobijn, B. and P. H. Franses (2000), Asymptotically Perfect and Relative Convergence of Productivity, *Journal of Applied Econometrics*, 15, 59–81.

Hoogerheide, L., J. H. Block, and R. Thurik (2010), Family background variables as instruments for education in income regressions: a Bayesian analysis, *Tinbergen Institute Discussion Paper*, 075–3.

Hoogerheide, L., F. Kleibergen, and H. K. Van Dijk (2007a), Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data, *Journal of Econometrics*, 138, 63–103.

Hoogerheide, L. F., J. F. Kaashoek, and H. K. Van Dijk (2007b), On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks, *Journal of Econometrics*, 139, 154–180.

Hoogerheide, L. F., J. F. Kaashoek, and H. K. Van Dijk (2007c), On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks, *Journal of Econometrics*, 139, 154–180.

Hoogerheide, L. F. and H. K. Van Dijk (2006), A reconsideration of the Angrist-Krueger analysis on returns to education, *Econometric Institute Report*.

Hop, J. P. and H. K. Van Dijk (1992), SISAM and MIXIN: Two algorithms for the computation of posterior moments and densities using Monte Carlo integration, *Computational Economics*, 5, 183–220.

Jedidi, K., H. S. Jagbal, and D. W. S (1997), Finite Mixture Structural Equation Models for Response-based Segmentation and Unobserved Heterogeneity, *Marketing Science*, 16, 39–59.

Kalaitzidakis, P., T. P. Mamuneas, and T. Stengos (2001), Measures of Human Capital and Nonlinearities in Economic Growth, *Journal of Economic Growth*, 6, 229–254.

Kass, R. E. and A. E. Raftery (1995), Bayes factors, *Journal of the American Statistical Association*, 90, 773–795.

Kim, C. J. and C. R. Nelson (1999), *State-space Models with Regime Switching: Classical and Gibbs-sampling Approaches with Applications*, vol. 1, The MIT press.

Kleibergen, F. and H. K. Van Dijk (1994a), Bayesian analysis of simultaneous equation models using noninformative priors, *Tinbergen Institute Discussion Paper*.

Kleibergen, F. and H. K. Van Dijk (1994b), On the shape of the likelihood/posterior in cointegration models, *Econometric Theory*, 10, 514–551.

Kleibergen, F. and H. K. Van Dijk (1998), Bayesian simultaneous equations analysis using reduced rank structures, *Econometric Theory*, 14, 701–743.

Kloek, T. and H. K. Van Dijk (1978), Bayesian estimates of equation system parameters: an application of integration by Monte Carlo, *Econometrica: Journal of the Econometric Society*, 46, 1–19.

Koop, G. (2003), *Bayesian econometrics*, Wiley, New York.

Laud, P. W. and J. G. Ibrahim (1995), Predictive model selection, *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 247–262.

Leamer, E. E. (1978), *Specification searches: ad hoc inference with nonexperimental data*, Wiley, New York.

Lee, K. and B.-Y. Kim (2009), Both Institutions and Policies Matter but Differently for Different Income Groups of Countries: Determinants of Long-Run Economic Growth Revisited, *World Development*, 37, 533–549.

Levine, R. (2004), Finance and Growth: Theory and Evidence, *NBER Working Paper Series*.

Loayza, N. V. and R. Ranciere (2006), Financial Development, Financial Fragility, and Growth, *Journal of Money, Credit and Banking*, 38, 1051–1076.

Mankiw, N. G. and D. Romer (1992), A Contribution to the Empirics of Economic Growth, *Quarterly Journal of Economics*, 107, 407–437.

Marske (1967), Biomedical Oxygen Demand Data Interpretation Using Sums of Squares Surface, unpublished master's thesis, University of Wisconsin.

Meng, X. L. and S. Schilling (2002), Warp bridge sampling, *Journal of Computational & Graphical Statistics*, 11, 552–586.

Meng, X. L. and W. H. Wong (1996), Simulating ratios of normalizing constants via a simple identity: A theoretical exploration, *Statistica Sinica*, 6, 831–860.

Méon, P. G. and L. Weill (2008), Does Financial Intermediation Matter for Macroeconomic Efficiency?, *Working Papers of LaRGE (Laboratoire de Recherche en Gestion et Economie)*.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953), Equation of state calculations by fast computing machines, *The journal of chemical physics*, 21, 1087.

Miazhynskaia, T. and G. Dorffner (2006), A comparison of Bayesian model selection based on MCMC with an application to GARCH-type models, *Statistical papers*, 47, 525–549.

Min, C. and A. Zellner (1993), Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates, *Journal of Econometrics*, 56, 89–118.

Mira, A. and G. Nicholls (2004), Bridge estimation of the probability density at a point, *Statistica Sinica*, 14, 603–612.

Newey, W. K. and K. D. West (1987), A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica: Journal of the Econometric Society*, 55, 703–708.

Newton, M. A. and A. E. Raftery (1994), Approximate Bayesian inference with the weighted likelihood bootstrap, *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–48.

Paap, R., P. H. Franses, and D. Van Dijk (2005), Does Africa Grow Slower than Asia, Latin America and the Middle East? Evidence from a new Data-based Classification Method, *Journal of Development Economics*, 77, 553–570.

Paap, R. and H. K. Van Dijk (1998), Distribution and Mobility of Wealth of Nations, *European Economic Review*, 42, 1269–1293.

Parker, S. C. and C. M. Van Praag (2006), Schooling, Capital Constraints, and Entrepreneurial Performance, *Journal of Business and Economic Statistics*, 24, 416–431.

Pesaran, M. H. (2007), A Pair-wise Approach to Testing for Output and Growth Convergence, *Journal of Econometrics*, 138, 312–355.

Phillips, P. C. B. and D. Sul (2009), Economic Transition and Growth, *Journal of Applied Econometrics*, 24, 1153–1185.

Psacharopoulos, G. and H. A. Patrinos (2004), Returns to investment in education: a further update, *Education economics*, 12, 111–134.

Quah, D. T. (1993), Galton's Fallacy and Tests of the Convergence Hypothesis, *The Scandinavian Journal of Economics*, 95, 427–443.

Quah, D. T. (1996), Empirics for Economic Growth and Convergence, *European Economic Review*, 40, 1353–1375.

Rioja, F. and N. Valev (2004), Finance and the Sources of Growth at Various Stages of Economic Development, *Economic Inquiry*, 42, 127–140.

Ritter, C. and M. A. Tanner (1992), Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler, *Journal of the American Statistical Association*, 87, 861–868.

Rossi, P. E., G. Allenby, and R. McCulloch (2005), *Bayesian Statistics and Marketing*, John Wiley and Sons.

Sachs, J. D. and A. M. Warner (1997), Sources of Slow Growth in African Economies, *Journal of African Economies*, 6, 335–376.

Sargan, J. D. (1958), The estimation of economic relationships using instrumental variables, *Econometrica: Journal of the Econometric Society*, 26, 393–415.

Schwarz, G. (1978), Estimating the Dimension of a Model, *The Annals of Statistics*, 6, 461–464.

Silverman, B. W. (1998), *Density estimation for statistics and data analysis*, Chapman & Hall/CRC.

Sims, C. A. (1980), Macroeconomics and reality, *Econometrica: Journal of the Econometric Society*, 1–48.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van der Linde (2002), Bayesian Measures of Model Complexity and Fit, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 583–639.

Tanner, M. A. and W. H. Wong (1987), The calculation of posterior distributions by data augmentation, *Journal of the American statistical Association*, 82, 528–540.

R Development Core Team (2008), R: A language and environment for statistical computing, *R Foundation for Statistical Computing, Vienna, Austria*.

Trostel, P., I. Walker, and P. Woolley (2002), Estimates of the economic return to schooling for 28 countries, *Labour Economics*, 9, 1–16.

Turner, R. T. (2000), Estimating the Propagation Rate of a Viral Infection of Potato Plants via Mixtures of Regressions, *Journal Of The Royal Statistical Society Series C*, 49, 371–384.

Van Dijk, H. K. (1999), Some Remarks on the simulation revolution in Bayesian econometric inference, *Econometric Reviews*, 18, 105–112.

Van Dijk, H. K. and T. Kloek (1980), Further experience in Bayesian analysis using Monte Carlo integration, *Journal of Econometrics*, 14, 307–328.

Verdinelli, I. and L. Wasserman (1995), Computing Bayes Factors using a generalization of the Savage-Dickey Density Ratio., *Journal of the American Statistical Association*, 90.

Wedel, M. (2002), Concomitant Variables in Finite Mixture Models, *Statistica Nederlandica*, 56, 362–375.

Zeevi, A. J. and R. Meir (1997), Density estimation through convex combinations of densities: approximation and estimation bounds, *Neural Networks*, 10, 99–109.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.

Zellner, A. and F. C. Palm (2004), *The structural econometric time series analysis approach*, Cambridge University Press.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit 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:

439 Y. -Y. TSENG, *Valuation of travel time reliability in passenger transport.*
440 M. C. NON, *Essays on Consumer Search and Interlocking Directorates.*
441 M. DE HAAN, *Family Background and Children's Schooling Outcomes.*
442 T. ZAVADIL, *Dynamic Econometric Analysis of Insurance Markets with Imperfect Information.*
443 I. A. MAZZA, *Essays on endogenous economic policy.*
444 R. HAIJEMA, *Solving large structured Markov Decision Problems for perishable-inventory management and traffic control.*
445 A. S. K. WONG, *Derivatives in Dynamic Markets.*
446 R. SEGERS, *Advances in Monitoring the Economy.*
447 F. M. VIEIDER, *Social Influences on Individual Decision Making Processes.*
448 L. PAN, *Poverty, Risk and Insurance: Evidence from Ethiopia and Yemen.*
449 B. TIEBEN, *The concept of equilibrium in different economic traditions: A Historical Investigation.*
450 P. HEEMEIJER, *Expectation Formation in Dynamic Market Experiments.*
451 A. S. BOOIJ, *Essays on the Measurement Sensitivity of Risk Aversion and Causal Effects in Education.*
452 M. I. LÒPEZ YURDA, *Four Essays on Applied Microeconometrics.*
453 S. MEENTS, *The Influence of Sellers and the Intermediary on Buyers' Trust in C2C Electronic Marketplaces.*
454 S. VUJIĆ, *Econometric Studies to the Economic and Social Factors of Crime.*
455 F. HEUKELOM, *Kahneman and Tversky and the Making of Behavioral Economics.*
456 G. BUDAI-BALKE, *Operations Research Models for Scheduling Railway Infrastructure Maintenance.*
457 T. R. DANIËLS, *Rationalised Panics: The Consequences of Strategic Uncertainty during Financial Crises.*
458 A. VAN DIJK, *Essays on Finite Mixture Models.*
459 C. P. B. J. VAN KLAVEREN, *The Intra-household Allocation of Time.*
460 O. E. JONKEREN, *Adaptation to Climate Change in Inland Waterway Transport.*
461 S. C. GO, *Marine Insurance in the Netherlands 1600-1870, A Comparative Institutional Approach.*
462 J. NIEMCZYK, *Consequences and Detection of Invalid Exogeneity Conditions.*
463 I. BOS, *Incomplete Cartels and Antitrust Policy: Incidence and Detection*
464 M. KRAWCZYK, *Affect and risk in social interactions and individual decision-making.*

465 T. C. LIN, *Three Essays on Empirical Asset Pricing.*

466 J. A. BOLHAAR, *Health Insurance: Selection, Incentives and Search.*

467 T. FARENHORST-YUAN, *Efficient Simulation Algorithms for Optimization of Discrete Event Based on Measure Valued Differentiation.*

468 M. I. OCHEA, *Essays on Nonlinear Evolutionary Game Dynamics.*

469 J. L. W. KIPPERSLUIS, *Understanding Socioeconomic Differences in Health An Economic Approach.*

470 A. AL-IBRAHIM, *Dynamic Delay Management at Railways: A Semi-Markovian Decision Approach.*

471 R. P. FABER, *Prices and Price Setting.*

472 J. HUANG, *Education and Social Capital: Empirical Evidences from Microeconomic Analyses.*

473 J. W. VAN DER STAATEN *Essays on Urban Amenities and Location Choice.*

474 K. M. LEE, *Filtering Non Linear State Space Models: Methods and Economic Applications.*

475 M. J. REINDERS, *Managing Consumer Resistance to Innovations.*

476 A. PARAKHONYAK, *Essays on Consumer Search, Dynamic Competition and Regulation.*

477 S. GUPTA, *The Study of Impact of Early Life Conditions on Later Life Events: A Look Across the Individual's Life Course.*

478 J. LIU, *Breaking the Ice between Government and Business: From IT Enabled Control Procedure Redesign to Trusted Relationship Building.*

479 D. RUSINOVA, *Economic Development and Growth in Transition Countries.*

480 H. WU, *Essays on Top Management and Corporate Behavior.*

481 X. LUI, *Three Essays on Real Estate Finance.*

482 E. L. W. JONGEN, *Modelling the Impact of Labour Market Policies in the Netherlands.*

483 M. J. SMIT, *Agglomeration and Innovations: Evidence from Dutch Microdata*

484 S. VAN BEKKUM, *What is Wrong With Pricing Errors? Essays on Value Price Divergence.*

485 X. HU, *Essays on Auctions.*

486 A. A. DUBOVIK, *Economic Dances for Two (and Three).*

487 A. M. LIZYAYEV, *Stochastic Dominance in Portfolio Analysis and Asset Pricing.*

488 B. SCHWAAB, *Credit Risk and State Space Methods.*