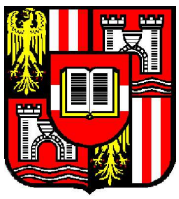


Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2004-02

Model-based Clustering of Multiple Time Series

Sylvia Frühwirth-Schnatter and Sylvia Kaufmann[a]

Juni 2004

[a]Oesterreichische Nationalbank, Economic Studies Division, e-mail sylvia.kaufmann@oenb.at, and University of Vienna, Department of Economics

Abstract

We propose to use the attractiveness of pooling relatively short time series that display similar dynamics, but without restricting to pooling all into one group. We suggest to estimate the appropriate grouping of time series simultaneously along with the group-specific model parameters. We cast estimation into the Bayesian framework and use Markov chain Monte Carlo simulation methods. We discuss model identification and base model selection on marginal likelihoods. A simulation study documents the efficiency gains in estimation and forecasting that are realized when appropriately grouping the time series of a panel. Two economic applications illustrate the usefulness of the method in analyzing also extensions to Markov switching within clusters and heterogeneity within clusters, respectively.[1] JEL classification: C11,C33,E32

Panel data, clustering, mixture modelling, Markov switching, Markov chain Monte Carlo.

1 Introduction

Let $\{y_{it}\}, t = 1, \ldots, T$ be a panel of multiple time series observed for N units $i = 1, \ldots, N$. The modelling approach pursued in this paper is based on formulating a time-series models for each univariate time series $y_i = \{y_{i1}, \ldots, y_{iT}\}$ in terms of the joint predictive density $p(y_i|\vartheta)$, where ϑ is an unknown parameter that needs to be estimated from the data. If T were large, the parameter ϑ could be estimated for each time series y_i individually. However, if T is relatively small one might use information from the other time series in the panel to estimate ϑ. In the case that all time series were generated by the same parameter, one would of course estimate ϑ from the pooled panel. This approach, however, introduces a bias, if the data-generating parameter ϑ differs substantially between all or some of the time series.

The basic idea of our model-based clustering approach is to assume that among the N multiple time series, K hidden groups are present, whereby all time series within each group, say k, are characterized by an econometric model with group-specific parameter ϑ_k. Consequently, information from all time series in the group can be used to estimate ϑ_k. An important feature of our approach is to assume that group membership of a certain time series is unknown apriori, and is estimated along with the group-specific parameters.

Consider, for illustration, a panel of time series where the main goal is forecasting a few steps ahead. In order to capture the short-term dynamics of each time series, model-based clustering is based on an AR(p) model, with the autoregressive parameters being different among K groups:

$$y_{it} = c_k + \delta_{1,k} y_{i,t-1} + \ldots + \delta_{p,k} y_{i,t-p} + \varepsilon_{it}, \tag{1}$$

[1] A draft version of the paper was presented under the title "Bayesian Clustering of Many Short Time Series" at the 57th European Meeting of the Econometric Society in 2002, at the 7th Valencia International Meeting on Bayesian Statistics in 2002 and at seminar presentations given in October 2002 at CORE, Louvain-la-Neuve, and the Erasmus University Rotterdam. The work of the first author was partly supported by the Austrian Science Foundation (FWF) under grant SFB 010 ('Adaptive Information Systems and Modelling in Economics and Management Science'). The views expressed in the paper are those of the authors and do not necessarily reflect those of the OeNB.

where $\varepsilon_{it} \sim \text{Normal}\,(0, \sigma_k^2)$. Bayesian econometric inference will then yield estimates for each $\vartheta_k = (c_k, \delta_{1,k}, \ldots, \delta_{p,k}, \sigma_k^2)$, $k = 1, \ldots, K$, as well as individual forecasts for each time series y_i, together with the posterior probability that time series y_i belongs to group k.

The idea of model-based clustering of time series is quite general and may be applied to a much broader class of time series models than only the AR(p) model. The choice of a specific model class usually will be guided by the general features prevalent in the observed time series. One example we will study in this article, is model-based clustering of dynamic regression models in which we allow for group-specific dependence on exogenous variables. Our approach bears similarities to the technique of stratifying a panel of time series by some variable prior to estimation, see e.g. Baltagi (1995). A distinctive feature of our clustering technique, however, is that group membership is estimated simultaneously with the remaining model parameters rather than determined prior to estimation.

A further example will be model-based clustering of Markov-switching autoregressive processes, introduced to econometrics by Hamilton (1989). Thereby we allow for structural breaks at unknown dates, with the additional feature that these breaks are allowed to occur at different times in the different groups. The inclusion of hidden Markov chains to allow for structural changes in a panel is related to, but different from the threshold panel data technique of Hansen (1999). Finally, we will consider model-based clustering of random effect models, where heterogeneity is present also within each group. This model is rather popular in marketing research, see for instance Lenk and DeSarbo (2000); economic applications, however, are rather rare. Canova (2004) applied a related classification technique to test for convergence in income data, and we will present additional evidence obtained with his data in the last one of our case studies.

Our approach is closely related to model-based clustering based on mixture models for non time-series data, see for instance the monograph of McLachlan and Basford (1988) and Bensmail et al. (1997). Whereas for non-time series data, distance-based clustering methods such as K-means clustering are attractive alternatives to model-based clustering, distance-based clustering methods are not easily extended to time series, hindered mainly by the difficulty of defining an appropriate distance measures between time series. The investigations of this paper will demonstrate that model-based clustering methods extend to time series in a quite natural way.

Concerning estimation, we pursue a fully Bayesian approach, using Markov chain Monte Carlo (MCMC) methods based on data augmentation, where we heavily draw from a lot a related paper on MCMC methods for mixture models, in particular from Diebolt and Robert (1994) and Frühwirth-Schnatter (2001b).

The outline of the rest of the paper is as follows. In Section 2 we formulate the general model framework. In Section 3 we discuss Bayesian estimation using MCMC methods, whereas Section 4 deals with selecting the number of groups. In Section 5 we discuss in more detail pooling within clusters using dynamic panel data models, and demonstrate the usefulness of our idea by means of a simulation study. In Section 6, we study clustering time series under regime switching, and consider clustering of industrial production growth rates of twenty-one countries as an application. In Section 7, we discuss heterogeneity within clusters, and look for convergence clubs in income data from 144 European NUTS2 units as illustration.

2 Model-based Clustering of Time Series

2.1 Model Formulation and Notation

Let $\{y_{it}\}, t = 1, \ldots, T$ be time series observed for N units $i = 1, \ldots, N$. The approach pursued in this paper is very general and is based on formulating a time-series models for each univariate time series $y_i = \{y_{i1}, \ldots, y_{iT}\}$ in terms of the joint predictive density $p(y_i|\vartheta)$, where ϑ is an unknown parameter taking values in a parameter space Θ. The predictive density may depend on observed exogenous variables, however this dependence will not be made explicit in our notation to keep it simple. Typically, $p(y_i|\vartheta)$ will be specified in terms of the one-step ahead predictive densities $p(y_{it}|y_{i,t-1}, \ldots, y_{i,t-p}, \vartheta)$:

$$p(y_i|\vartheta) = \prod_{t=p+1}^{T} p(y_{it}|y_{i,t-1}, \ldots, y_{i,t-p}, \vartheta). \tag{2}$$

Our approach may be applied to discrete-valued as well as continuous-valued time series. For discrete-valued time series, $p(y_{it}|y_{i,t-1}, \ldots, y_{i,t-p}, \vartheta)$ is a discrete probability distribution. Emphasis of the present paper, however, lies on continuous-valued time series.

To be more specific, assume that the one-step ahead predictive density arises from a normal distribution:

$$y_{it}|y_{i,t-1}, \ldots, y_{i,t-p}, \vartheta \sim \text{Normal}\left(\hat{y}_{it|t-1}(\vartheta), C_{it|t-1}(\vartheta)\right), \tag{3}$$

where $\hat{y}_{it|t-1}(\vartheta)$ and $C_{it|t-1}(\vartheta)$ are the mean and the variance of the one-step ahead predictive density, which may depend on $y_{i,t-1}, \ldots, y_{i,t-p}$ and on exogenous variables. A typical example is the AR(p) model, where $\vartheta = (c, \delta_1, \ldots, \delta_p, \sigma^2)$, with mean and variance given by $\hat{y}_{it|t-1}(\vartheta) = c + \delta_1 y_{i,t-1} + \ldots + \delta_p y_{i,t-p}$, and $C_{it|t-1}(\vartheta) = \sigma^2$, respectively. Another example is the GARCH(1,1)-model, where $\vartheta = (\gamma, \alpha, \delta)$, with mean and variance given by $\hat{y}_{it|t-1}(\vartheta) = 0$ and $C_{it|t-1}(\vartheta) = \gamma + \alpha y_{i,t-1}^2 + \delta C_{i,t-1|t-2}(\vartheta)$.

Following a well-developed tradition in Bayesian econometrics (Zellner 1971, Geweke 1993) a certain robustness against outliers in time series is achieved by choosing model densities from the t_ν-distribution rather than the normal distribution. Here, robustness against atypical time series is achieved by assuming that $p(y_i|\vartheta)$ arises from a multivariate t-distribution. To simplify estimation, we represent the t_ν-distribution as a scaled mixture of normal distributions as in Geweke (1993) and define the following one-step ahead predictive densities,

$$y_{it}|y_{i,t-1}, \ldots, y_{i,t-p}, \lambda_i, \vartheta \sim \text{Normal}\left(\hat{y}_{it|t-1}(\vartheta), C_{it|t-1}(\vartheta)/\lambda_i\right), \tag{4}$$

where

$$\lambda_i \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \tag{5}$$

and $\hat{y}_{it|t-1}(\vartheta)$ and $C_{it|t-1}(\vartheta)$ are the same as in (3).

2.2 Model-based Clustering

Let us further assume that the N observed time series in fact form K groups, whereby within each group, say k, an econometric model based on the same parameter ϑ_k for all time series could be used for inference and forecasting, or in other words, we could pool all time series within a cluster. To this aim, a latent group indicator S_i is introduced for each time series y_i, which takes a value out of the discrete set $\{1, \ldots, K\}$, i.e. $S_i = k$ if time series y_i belongs to group k. Thus, knowing S_i is equivalent to knowing the unit specific parameter:

$$p(y_i|S_i, \vartheta_1, \ldots, \vartheta_K) = p(y_i|\vartheta_{S_i}) = \begin{cases} p(y_i|\vartheta_1), & \text{if } S_i = 1, \\ \vdots & \vdots \\ p(y_i|\vartheta_K), & \text{if } S_i = K. \end{cases} \tag{6}$$

An important aspect of model (6) is that we do not assume to know a priori which time series belong to which group. For each time series, the group indicator S_i is estimated along with the group-specific parameters $\vartheta_1, \ldots, \vartheta_K$ from the data. The Bayesian classification rule (see equation (9) below) combines the information in the data with the prior information available on group indicator S_i to obtain the inference on group membership. Two sensible prior structures for S_i are discussed in the next subsection.

2.3 Modelling the Prior of the Group Indicators

In order to complete the model specification, we have to formulate a probabilistic model for the group indicators $S^N = (S_1, \ldots, S_N)$. This probabilistic model turns out to be the prior distribution of S^N within the Bayesian approach we pursue in the present paper. First, we assume that $S_1, \ldots, S_N$ are a priori independent. For each $i = 1, \ldots, N$ we can then define the prior probability, $\Pr\{S_i = k\}$, that a certain time series y_i belongs to group k. We consider here two probabilistic structures for this prior.

The first prior is the one that is commonly used in the context of mixture model, namely to assume complete prior ignorance about the group membership of a certain unit. Then the prior probability of time series y_i to belong to group k is equal to the relative size η_k of that group:

$$\Pr\{S_i = k|\eta_1, \ldots, \eta_K\} = \eta_k. \tag{7}$$

The group sizes $(\eta_1, \ldots, \eta_K)$, which obviously sum to 1, are assumed to be unknown and are estimated along with the data.

In practice, however, a unit-specific factor, which may be economic, geographic or sociopolitical, might contain a priori information on how to group the time series of a panel. Such information might be included in a priori clustering by assuming a logit-type model for $\Pr\{S_i = k\}$. For instance, to model dependence on a single unit-specific exogenous variable z_i for $K = 2$, the corresponding prior would read:

$$\begin{aligned} \Pr\{S_i = 1|\gamma_1, \gamma_2, z_i\} &= \frac{1}{1 + \exp(\gamma_1 + z_i\gamma_2)}, \\ \Pr\{S_i = 2|\gamma_1, \gamma_2, z_i\} &= \frac{\exp(\gamma_1 + z_i\gamma_2)}{1 + \exp(\gamma_1 + z_i\gamma_2)}, \end{aligned} \tag{8}$$

where (γ_1, γ_2) are unknown parameters also estimated from the data. If γ_2 equals zero, then prior (8) reduces to prior (7), with a different parameterization for the group sizes, however. Note that in this case $\eta_1 = 1/(1 + \exp(\gamma_1))$, whereas $\eta_2 = \exp(\gamma_1)/(1 + \exp(\gamma_1))$. If γ_2 is different from 0, then z_i helps to predict group membership. We may analyze $\Pr\{S_i = 1|\gamma_1, \gamma_2, z_i\}$ for all observed time series, in order to evaluate the discriminative power of z_i with respect to group membership. By testing if γ_2 is actually different from 0, as well as by analyzing the discriminate power of z_i for the observed time series, we obtain important insights into the factors that determine group membership. Prior (8) is easily extended to deal with more than one exogenous variable and with more than two groups.

3 Estimation

Estimation of the model introduced in Section 2 is carried out within a Bayesian framework through the help of Markov chain Monte Carlo methods and data augmentation methods. Markov chain Monte Carlo methods have been applied for related models with hidden groups such as mixture models by, among many authors, Diebolt and Robert (1994) and Frühwirth-Schnatter (2001b).

3.1 Estimation Using MCMC

Subsequently $\vartheta_1, \ldots, \vartheta_K$ denote the unknown model parameters in the different groups, whereas ϕ summarizes unknown parameters in the prior of the group indicators, i.e. $\phi = (\eta_1, \ldots, \eta_K)$ for the ignorance prior (7) and $\phi = (\gamma_1, \gamma_2)$ for the logit-type prior (8). To pursue the Bayesian approach, we assume that a prior $p(\vartheta_1, \ldots, \vartheta_K, \phi)$ is available. Further quantities that are estimated jointly with these model parameters are the group indicators $S^N = (S_1, \ldots, S_N)$, either under the ignorance prior (7) or the logit-type prior (8).

Estimation of $\psi = (\vartheta_1, \ldots, \vartheta_K, \phi, S^N)$ using MCMC basically iterates between the two following step:

(a) *Classification for fixed parameters*: each time series as a whole is classified into one of the K groups by sampling the group indicator S_i for all time series $i = 1, \ldots, N$ from the posterior $p(S_i|y, \vartheta_1, \ldots, \vartheta_K, \phi)$.

(b) *Estimation for a fixed classification*: conditional on known indicators, the parameters $(\vartheta_1, \ldots, \vartheta_K)$ and ϕ are conditionally independent. Estimation is carried out by sampling the group-specific parameters $\vartheta_1, \ldots, \vartheta_K$ from the posterior $p(\vartheta_1, \ldots, \vartheta_K|S^N, y)$ and the parameters ϕ relevant for prior classification from the posterior $p(\phi|S^N, y)$.

Step (a) makes use only of the predictive density $p(y_i|\vartheta)$, defined for each time series in (2), as well as the prior classification probabilities (7) or (8):

$$\Pr\{S_i = k|y, \vartheta_1, \ldots, \vartheta_K, \phi\} \propto p(y_i|\vartheta_k)\Pr\{S_i = k|\phi\}, \quad k = 1, \ldots, K. \tag{9}$$

Equation (9) shows that model-based clustering of time series is a very general method that may be applied to many different classes of time series models.

Sampling the group-specific parameters $\vartheta_1, \ldots, \vartheta_K$ in step (b) is particularly easy, if the various groups share no parameters. Then, each group parameter ϑ_k is estimated by pooling all time series that currently belong to group k:

$$p(\vartheta_k|y, S_1, \ldots, S_N) = \prod_{i:S_i=k} p(\vartheta_k|y_i). \tag{10}$$

To sample from this posterior one can make use of the many results available on MCMC estimation of particular time series models. Chib (2001) gives an excellent review on MCMC estimation of common econometric time series models; see also Chib and Greenberg (1994) in particular for ARMA models and Nakatsuma (2000) for GARCH models.

Sampling the prior parameters ϕ of the group indicators in step (b) is standard for the ignorance prior (7). The posterior of $\phi = (\eta_1, \ldots, \eta_K)$ follows a Dirichlet distribution. Under the logit-prior (8), the posterior $p(\phi|S_1, \ldots, S_N)$ is not of closed form and a Metropolis-Hastings-algorithm is used to sample ϕ, see Albert and Chib (1993b) and Scott (1999) for more details.

Under the robust predictive density (4), the scale factors $\lambda^N = (\lambda_1, \ldots, \lambda_N)$ have to be added to the vector ψ of unknown quantities, and an additional step is necessary to sample these scale factors:

(c) *Determining the weights for robust estimation and classification*: conditional on known indicators and known parameters, λ_i is sampled from the posterior $p(\lambda_i|y, \vartheta_{S_i})$ for each time series:

$$\lambda_i|y, \vartheta_{S_i} \sim \text{Gamma}\left(\frac{\nu + T - p}{2}, \frac{\nu}{2} + .5 \sum_{t=p+1}^{T} \left(\frac{(y_{it} - \hat{y}_{i,t|t-1}(\vartheta_{S_i}))^2}{C_{i,t|t-1}(\vartheta_{S_i})}\right)^2\right). \tag{11}$$

In this case, step (a) and (b) are carried out conditional on λ^N. The classification rule (9), for instance, is substituted by:

$$\Pr\{S_i = k|y, \vartheta_1, \ldots, \vartheta_K, \lambda^N, \phi\} \propto p(y_i|\vartheta_k, \lambda_i)\Pr\{S_i = k|\phi\}, \tag{12}$$

using the augmented normal distribution $p(y_i|\vartheta_k, \lambda_i)$ rather than the t-distribution $p(y_i|\vartheta_k)$.

3.2 Unit-specific inference

The MCMC draws may be used to recover individual parameters for each time series and to obtain forecasts for each individual time series. Assume that after an appropriate burn-in-phase M MCMC draws $\psi^{(m)}$, $m = 1, \ldots, M$, are retained for inference.[2] Unit-specific inference is available from these draws, especially posterior draws of the unit-specific parameters and unit-specific forecasts.

The unit-specific parameter $\tilde{\vartheta}_i$ of time series y_i may be expressed as:

$$\tilde{\vartheta}_i = \sum_{k=1}^{K} \vartheta_k I_{\{S_i=k\}}, \tag{13}$$

[2]In the following, we use the superscript (m) whenever we refer to MCMC draws, e.g. $S_i^{(m)}$ for the mth draw of the group indicator S_i.

where the indicator function $I_{\{S_i=k\}}$ takes the value 1, iff $S_i = k$ and zero otherwise. Therefore posterior draws of the unit-specific parameters are given by,

$$\tilde{\vartheta}_i^{(m)} = \sum_{k=1}^{K} \vartheta_k^{(m)} I_{\{S_i^{(m)}=k\}}, \tag{14}$$

and may be evaluated for instance by determining posterior estimates of all unit-specific parameters as the mean of these draws:

$$\mathrm{E}(\tilde{\vartheta}_i|y) \approx \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \vartheta_k^{(m)} I_{\{S_i^{(m)}=k\}}. \tag{15}$$

The unit-specific draws $\tilde{\vartheta}_i^{(m)}$, together with MCMC draws of parameters common to all series, may be used to sample future paths $\{y_{i,T+1}^{(m)}, \ldots, y_{i,T+h}^{(m)}\}$ for each time series y_i in a similar way as in Albert and Chib (1993a). For each $m = 1, \ldots, M$, $y_{i,T+h}^{(m)}$ is a Bayesian forecast sampled recursively from (4):

$$y_{i,T+h}|y_i, y_{i,T+1}^{(m)}, \ldots, y_{i,T+h-1}^{(m)} \sim \text{Normal}\left(\hat{y}_{i,T+h|T}(\tilde{\vartheta}_i^{(m)}), C_{i,T+h|T}(\tilde{\vartheta}_i^{(m)})/\lambda_i^{(m)}\right) \tag{16}$$

For each forecasting horizon h, these draws are then evaluated, for instance by considering the mean of all draws $y_{i,T+h}^{(m)}$ as a point forecast, or by deriving interval forecasts in an evident manner.

3.3 Identification and Classification

To perform posterior classification and unit-specific inference using the group-specific parameters, the groups have to be identified through some inequality constraint on the group-specific parameters, in order to avoid label switching, see Celeux et al. (2000) and Frühwirth-Schnatter (2001b) for an extensive discussion of this subtle issue.

Usually, to find an appropriate restriction, we can use scatter plots of the MCMC sample and plot marginal distributions of the model parameters. These explorative tools often give a clear inference on distinct clusters in the data. The applications in Subsections 6.2 and 7.2 will present some examples. In particular, identification schemes for Markov switching panel data models are also discussed in Frühwirth-Schnatter and Kaufmann (2004) and Kaufmann (2004).

Once the model has been identified, it is possible to classify the time series into the various groups by estimating for each time series the posterior classification probability $\Pr\{S_i = k|y\}$ from the MCMC draws:

$$\Pr\{S_i = k|y\} \approx \frac{1}{M} \#\left\{S_i^{(m)} = k\right\}.$$

4 Selecting the Number of Groups

In practice the number K of groups will be unknown. Each model specification with a fixed number K of groups will be denoted by $\mathcal{M}_K$. Following a long tradition in Bayesian econometrics initiated by Zellner (1971) in the context of selecting

regression models, the marginal likelihood will be used to select among $\mathcal{M}_1$, ..., $\mathcal{M}_{K_{max}}$. Marginal likelihoods have been applied to various complex econometric model selection problems, for instance by Frühwirth-Schnatter (1995), Shively and Kohn (1997), and Koop and van Dijk (2000) in the context of state space models and by Chib et al. (2002) to select among stochastic volatility models; see also Geweke (1995) for a general discussion of Bayesian comparison of econometric models based on marginal likelihoods. An alternative Bayesian approach, which is not considered here, to select the number of hidden groups is the reversible jump Markov Chain Monte Carlo methods used by Richardson and Green (1997).

For a fixed number of clusters K, the marginal likelihood $p(y|\mathcal{M}_K)$ is defined by:

$$p(y|\mathcal{M}_K) = \int p(y_1, \ldots, y_N|\psi, K)p(\psi)d\psi. \tag{17}$$

Analytical integration of (17) with respect to the whole parameter $\psi = (\vartheta_1, \ldots, \vartheta_K, \phi, S^N, \lambda^N)$ is not possible, but the dimension of integration can be reduced substantially by analytically integrating with respect to S^N and λ^N:

$$p(y|\mathcal{M}_K) = \int p(y_1, \ldots, y_N|\vartheta_1, \ldots, \vartheta_K, \phi, K)p(\vartheta_1, \ldots, \vartheta_K, \phi)d\vartheta_1 \ldots d\vartheta_K d\phi, \tag{18}$$

where $p(y_1, \ldots, y_N|\vartheta_1, \ldots, \vartheta_K, \phi, K)$ is the integrated likelihood, derived from (2) and (7):

$$p(y_1, \ldots, y_N|\vartheta_1, \ldots, \vartheta_K, \phi) = \prod_{i=1}^{N} p(y_i|\vartheta_1, \ldots, \vartheta_K, \phi), \tag{19}$$

where:

$$p(y_i|\vartheta_1, \ldots, \vartheta_K, \phi) = \sum_{k=1}^{K} p(y_i|\vartheta_k)\Pr\{S_i = k|\phi\}. \tag{20}$$

Even for the reduced integral (18), the computation of the marginal likelihood is a non-trivial integration problem, see for instance the discussion in Geweke (1999). Marginal likelihoods have been estimated using methods such as the candidate's formula Chib (1995) and importance sampling based on mixture approximations Frühwirth-Schnatter (1995); see also the review in Chib (2001) and the references therein for more detail. Although these methods proved to be useful for a wide range of econometric models, they are apt to fail when estimating the marginal likelihood of mixture models, as the posterior density of such a model is highly irregular due to lack of identification for these models, see Frühwirth-Schnatter (2004).

To compute the marginal likelihood (18), we follow Frühwirth-Schnatter (2004) who demonstrated that the technique of bridge sampling (Meng and Wong 1996) is a useful method of computing the marginal likelihood for mixture models. Bridge sampling generalizes the method of importance sampling which has been applied to various complex econometric inference problems by, among others, van Dijk and Kloek (1980) and Geweke (1989). Like importance sampling, bridge sampling is based on an iid sample from an importance density, however, this sample is combined with the MCMC draws from the posterior density in an appropriate way.

One might wonder, why this extension is sensible. For importance sampling it is well-known that the variance of the resulting estimator depends on the ratio of the non-normalized posterior density over the importance density which may be unbounded for poorly chosen importance densities, see e.g. Geweke (1989) and Geweke (1999). An important advantage of bridge sampling is, that the variance of the resulting estimator depends on a ratio that is bounded regardless of the tail behavior of the underlying importance density. This allows the econometrician far more flexibility in the construction of the importance density. In Frühwirth-Schnatter (2004) the importance density is constructed during MCMC sampling in an unsupervised manner as a mixture of complete data posteriors.

5 Pooling within Clusters

5.1 Dynamic Panel Data

Panel data consisting of many, rather short time series occur frequently in various areas of applied econometrics such as macroeconomics, business or marketing. To analyze the data, usually the dependent variable y_{it}, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, is regressed on a set of explanatory variables X_{it}, which may include strictly exogenous variables and/or lagged values of y_{it}. Assuming first unit-specific regression coefficients, $\tilde{\beta}_i$, the model may be formulated (Baltagi 1995):

$$y_{it} = X_{it}\tilde{\beta}_i + \varepsilon_{it}, \tag{21}$$

where the error term is either homoscedastic, $\varepsilon_{it} \sim \text{Normal}\,(0, \sigma^2)$, or conditionally heteroscedastic, $\varepsilon_{it} \sim \text{Normal}\,(0, \sigma^2/\lambda_i)$, which together with the prior (5) implies that, marginally, y_{it} follows a t_ν-distribution.

If the time series in the panel are rather short (either absolutely or relatively compared to the dimension of $\tilde{\beta}_i$ in model (21)), then estimation of $\tilde{\beta}_i$ from the individual time series $y_i = \{y_{i1}, \ldots, y_{iT}\}$, will exhibit large estimation errors. In such cases, panel data are often pooled for estimation which means that a joint parameter $\tilde{\beta}_i \equiv \beta$ is estimated for all N time series in the panel, see Garcia-Ferrer et al. (1987), Maddala (1991), Zellner and Hong (1989) and Zellner et al. (1991), Mittnik (1990), and Hoogstrate et al. (2000) for a recent review. As *all* time series are pooled, we call this technique overall pooling in what follows.

One of the main advantages of overall pooling is to borrow strength from all time series in the panel to estimate the coefficient of an individual time series. Overall pooling, however, is known to introduce a bias for unit specific coefficients, if $\tilde{\beta}_i$ were different between time series. The results reported in Hoogstrate et al. (2000) suggest that only in those cases where the parameters are "similar" enough, the gain in reducing the estimation errors may be larger than the loss due to the bias, leading in total to reduced mean squared estimation and forecasting errors. Here we suggest pooling within clusters, which inherits the appealing property of borrowing strength, without restricting the estimation to overall pooling, however.

To formalize, the general model introduced in Section 2 is specified as a dynamic regression model:

$$y_{it} = X_{it}\beta_k + \varepsilon_{it}, \quad \text{if } S_i = k, \tag{22}$$

where the clusters defined by S_i are estimated along with the model parameters using the MCMC approach described in Section 3.1. Model (22) is related to the switching regression model introduced by Quandt (1972), as the parameters switch between the units whereas for each time series the parameter remains the same over time. The later assumption will be relaxed in Section 6, where we allow additionally for switching over time according to a hidden Markov chain in order to make clustering less sensitive to structural changes.

Clustering might be performed with respect to some coefficients only. The model then reads:

$$y_{it} = X_{it}^1 \alpha + X_{it}^2 \beta_k + \varepsilon_{it}, \quad \text{if } S_i = k, \tag{23}$$

where α are parameters subject to overall pooling, whereas β_k is pooled within clusters. For the sake of identifiability we have to assume that $X_{i,\cdot}^1$ and $X_{i,\cdot}^2$ share no common columns meaning that a variable has either a *fixed* effect α or group-specific effect β_k on y_{it}.

We end this section by the rather obvious remark, that for $K = 1$, clustering based on model (22) collapses to overall pooling. By testing $K > 1$ against $K = 1$ by means of marginal likelihoods (see Section 4), we are in a position to test overall pooling against pooling within clusters and to test for the appropriate number of clusters.

5.2 A Simulation Study

5.2.1 The simulation design

We generate synthetical panels from the following dynamic regression model where reaction to an exogenous variable z_t is group-specific:

$$y_{it} = c + \delta y_{i,t-1} + \beta_k z_t + \varepsilon_{it}, \tag{24}$$

with $\varepsilon_{it} \sim \text{Normal}\,(0, \sigma^2)$, where $N = 200$, $T = 24$, $\sigma^2 = 0.1$, and $z_t \sim \text{Normal}\,(0, 0.1)$. We assume three hidden groups ($K = 3$) and combine a small group ($\eta_1 = 0.1$) with a large ($\eta_2 = 0.6$) and a medium sized ($\eta_3 = 0.3$) one.

We investigate six different scenarios of heterogeneity, by choosing different sets of group-specific parameters $(\beta_1, \beta_2, \beta_3)$. The first scenario is actually a setting of homogeneity with $\beta_1 = \beta_2 = \beta_3 = -0.45$. For the remaining five scenarios, the group-specific parameters $(\beta_1, \beta_2, \beta_3)$ are chosen such that the overall mean $\overline{\beta}$,

$$\overline{\beta} = \sum_{k=1}^{3} \beta_k \eta_k, \tag{25}$$

is identical to –0.45. However, heterogeneity measured by the variability Q of the group-specific parameters around the overall mean,

$$Q = \sum_{k=1}^{3} (\beta_k - \overline{\beta})^2 \eta_k, \tag{26}$$

increases and ranges from $Q = 0.0039$ to $Q = 0.3$, see Table 1. In all settings, the parameter of the largest group is the closest to the overall mean and the distribution

Table 1: Comparing pooling within K clusters with overall pooling by the ratio of the average mean squared estimation error $\text{MSE}_{\tilde{\beta}_i}$. A ratio bigger than 1 favors overall pooling, a ratio smaller than 1 favors pooling within K clusters. Q: heterogeneity, see (26); D: coefficient of determination, see (29).

Q	0	0.00395	0.0188	0.05	0.113	0.3
D	0	0.0500	0.2000	0.4000	0.6000	0.8
$K=2$	9.44	1.45	1.05	0.676	0.326	0.145
$K=3$	14.5	1.75	1.08	0.691	0.343	0.106
$K=4$	14.8	1.85	1.09	0.67	0.345	0.109

of heterogeneity is asymmetric, whereby the parameter of the smallest group is further away from the overall mean than the one of the medium sized group. The fixed parameters always take the values $(c, \delta) = (3.5, 0.3)$.

For each setting we simulate 100 panels for each of which we estimate model (24) for $K = 1, \ldots, 4$ using the MCMC methods described in section 3.1. Only the first $T = 20$ observations are used for estimation, while the last four are left for out-of sample evaluation. After a burn-in-phase of 1000 draws, $M = 1000$ MCMC draws are used to evaluate the estimation and the forecasting performance of pooling within clusters ($K = 2, 3, 4$) relative to overall pooling ($K = 1$).

In each scenario, we estimate clustering models (for each panel) setting K equal to $K = 1, \ldots, 4$, and evaluate the estimation and forecasting performance of pooling within clusters, i.e. $K = 2, 3, 4$, relative to overall pooling, $K = 1$.

5.2.2 Estimation performance

First of all, we consider the mean squared estimation error between the true unit-specific coefficient $\tilde{\beta}_i$,[3] which is equal to β_k if series i belongs to group k (see equation (24), and the posterior simulations $\tilde{\beta}_i^{(m)}$ (see equation (14) in Subsection 3.2). For each panel, we then compute the mean estimation error by:

$$\text{MSE}_{\tilde{\beta}_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{m=1}^{M}(\tilde{\beta}_i^{(m)} - \tilde{\beta}_i)^2. \tag{27}$$

Table 1 reports the average of the mean squared errors $\text{MSE}_{\tilde{\beta}_i}$ of the 100 simulated panels in each scenario, when K is alternatively set to $K = 2, 3, 4$, relative to overall pooling. A ratio bigger than one favors overall pooling, a ratio smaller than one favors pooling within clusters. Not surprisingly, we loose efficiency in estimating the individual parameters by introducing clusters in a case where none are present ($Q = 0$). Interestingly, a loss of efficiency is also present for $Q = 0.00395$ and $Q = 0.0188$, which is in line with the result of Hoogstrate et al. (2000) that overall pooling is preferable to unit-individual estimation when parameters, albeit being different, are quite similar across units. There is a clear gain when pooling within clusters for the last three scenarios.

To gain additional understanding of these results, we rewrite model (24) as

$$y_{it} = c + \delta y_{i,t-1} + \overline{\beta} z_t + \tilde{\varepsilon}_{it}, \tag{28}$$

[3]The notation is in analogy to subsection 3.2, in particular equation (13).

where $\overline{\beta}$ is the overall mean defined in (25) and $\tilde{\varepsilon}_{it} = \varepsilon_{it} + z_t(\tilde{\beta}_i - \overline{\beta})$ are heterogeneous errors. Whether for a given data set pooling within clusters is preferable or not, depends on how much of the variance of $\tilde{\varepsilon}_{it}$ in (28) is caused by heterogeneity among the groups. The contribution of heterogeneity usually is measured by the coefficient of determination D which is defined by the ratio of explained over total variance (see e.g. Gelfand et al. 1995):

$$D = \frac{QT\overline{z^2}}{QT\overline{z^2} + \sigma^2}, \tag{29}$$

where Q is defined in (26) and $\overline{z^2} = 1/T \sum_{t=1}^{T} z_t^2$. D obviously ranges from 0 to 1. For our synthetical data D increases from 0 to 0.8 within the six scenarios (see Table 1). If D is small, unobserved heterogeneity is not the cause of variability in $\tilde{\varepsilon}_{it}$ in (28). In this case the gain of clustering is negligible and overall pooling yields the lowest estimation error. The more D moves away from 0, the higher the share of heterogeneity in explaining the variance of $\tilde{\varepsilon}_{it}$, and pooling within clusters becomes worth the effort.

5.2.3 Forecasting Performance

We proceed with an out-of-sample evaluation, involving the true values $y_{i,T+1}, \ldots, y_{i,T+4}$ in comparison to the forecasts $y_{i,T+h}^{(m)}$ defined in Subsection 3.2, equation (16). We consider for each time series y_i in a certain panel the mean squared forecasting error, for various forecasting horizons h, and aggregate over all time series in the panel:

$$\text{MSFE}_h = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M} \sum_{m=1}^{M} (y_{i,T+h}^{(m)} - y_{i,T+h})^2. \tag{30}$$

Table 2 reports for each scenario the average of the mean squared forecasting errors MSFE_h over the 100 simulated panels for $h = 1, \ldots, 4$. We compare pooling within clusters with overall pooling through the ratio of these averages. A ratio bigger than 1 favors overall pooling, a ratio smaller than 1 favors pooling within clusters. Interestingly, when forecasting, one never looses efficiency by using a model with pooling within clusters. Whereas there is practically no gain in efficiency for scenarios of little heterogeneity, the gain increases with the amount of heterogeneity.

5.2.4 Selecting the Number of Clusters

The results reported so far indicate that we are likely to gain from introducing clusters in settings of considerable heterogeneity. For an empirical panel, however, we are faced with the problem of selecting the number of clusters. We end the simulation study by evaluating the usefulness of the marginal likelihood as a tool for choosing the number of clusters. Given the marginal likelihood of each panel for $K = 1, \ldots, 4$, in each scenario we select K according to the highest marginal likelihood. Table 3 reports the relative frequencies over the 100 simulated panels of choosing $K = 1, 2, 3, 4$.

Evidently, for short panels like the ones considered here, the selected number of clusters if often smaller than the true number of groups which is equal to one for $D = 0$ and equal to three, otherwise.

Table 2: Comparing pooling within clusters with overall pooling by the ratio of the average mean squared forecasting errors MSFE_h. A ratio bigger than 1 favors overall pooling, a ratio smaller than 1 favors pooling within K clusters. Q: heterogeneity, see (26); D: coefficient of determination, see (29).

	Q	0	0.00395	0.0188	0.05	0.113	0.3
	D	0	0.0500	0.2000	0.4000	0.6000	0.8
$K=2$	MSFE_1	1	0.998	0.995	0.985	0.958	0.889
	MSFE_2	1	1	0.999	0.987	0.962	0.896
	MSFE_3	0.998	1	0.995	0.986	0.961	0.879
	MSFE_4	1	1	1	0.987	0.963	0.882
$K=3$	MSFE_1	1	1	0.996	0.985	0.953	0.861
	MSFE_2	1	0.999	1	0.986	0.956	0.863
	MSFE_3	1	0.997	0.997	0.987	0.956	0.847
	MSFE_4	1	1	1	0.989	0.959	0.849
$K=4$	MSFE_1	0.998	0.999	0.997	0.983	0.95	0.857
	MSFE_2	1	0.999	0.999	0.985	0.956	0.864
	MSFE_3	1	0.997	0.995	0.984	0.954	0.847
	MSFE_4	1	1	1	0.989	0.957	0.85

Table 3: Frequency of selecting a model with K clusters among the 100 simulated panels.

Q	0	0.00395	0.0188	0.05	0.113	0.3
D	0	0.0500	0.2000	0.4000	0.6000	0.8
$K=1$	0.989	0.967	0.533	0.0326	0	0
$K=2$	0.0109	0.0326	0.467	0.957	0.087	0
$K=3$	0	0	0	0.0109	0.913	0.707
$K=4$	0	0	0	0	0	0.293

It is, however, very illuminating, to compare Table 3 with Table 1 and Table 2. The marginal likelihood is very reliable, when it comes to decide whether overall pooling is preferable to pooling within clusters in terms of recovering individual coefficients and forecasting performance. For those scenarios, where overall pooling is significantly more efficient in recovering the individual parameters (see columns $D=0$ and $D=0.05$ in Table 1), $K=1$ is practically selected all the time. For the third scenario, where the efficiency of both methods is comparable, one would select overall pooling in about 50 percent of the cases. Overall pooling is hardly selected in those cases where efficiency gains and an improved forecasting performance is achieved by pooling within clusters.

6 Regime Switching within Clusters

6.1 Regime Switching Dynamic Panel Data

Since the seminal paper of Hamilton (1989), Markov switching models became very popular in the analysis of macro-economic time series, as they are rather flexible, non-linear time series models that are able to capture asymmetric patterns found in many economic time series such as GDP, investment, and industrial production, see Kim and Nelson (1999), Kaufmann (2000) and Hamilton and Raj (2002) for a review.

To capture asymmetry in the panel with a Markov switching model, model (22) or (23) are extended to:

$$y_{it} = X_{it}^1 \alpha + X_{it}^2(\beta_k^G + \beta_k^R(I_{kt} - 1)) + \varepsilon_{it}, \quad \text{if } S_i = k. \tag{31}$$

For each group k, I_{kt} takes on the value 0 or 1 and follows a hidden two-state Markov chain with unknown transition matrix ξ_k. As in Section 5, pooling within clusters takes place, however pooling is toward different values, depending on the state of the group-specific indicator I_{kt}. If $I_{kt} = 1$, then pooling is toward β_k^G, if $I_{kt} = 0$ pooling is toward $\beta_k^G - \beta_k^R$.

Model (31) combines Quandt (1972) and Goldfeld and Quandt (1973), as we allow for parameters that are switching between the groups as well as for changes of the parameters over time. The inclusion of hidden Markov chains to allow for structural changes in the panel is related to the threshold panel data technique of Hansen (1999). Pooling under structural breaks, however without including a hidden Markov chain, is also discussed in Hoogstrate et al. (2000) for a panel of growth rates of real GDP of 18 OECD countries.

For model (31), the MCMC procedure discussed in Subsection 3.1 needs an additional step to sample the hidden Markov chain $I_k^T = (I_{k0}, \ldots, I_{kT})$ as well as the transition matrices $\xi^k = (\xi_{00}^k, \xi_{01}^k, \xi_{10}^k, \xi_{11}^k)$ in each group, where $\xi_{jl}^k = \Pr\{I_{kt} = l | I_{k,t-1} = j\}$ is the probability that I_{kt} will be in state l given that I_{kt} was in state j in the previous period. We apply here full conditional Gibbs sampling, meaning that the steps (a) - (c) in Subsection 3.1 are carried out conditional on known values for (I_k^T, ξ^k), $k = 1, \ldots, K$. Conditional on the actual value for S^N, (I_k^T, ξ^k) are independent across the groups. Sampling (I_k^T, ξ^k) within each group follows closely the standard MCMC methods developed for a single hidden Markov chain, see, among many others, Robert et al. (1993), Albert and Chib (1993a) McCulloch and Tsay (1994), Chib (1996) and Frühwirth-Schnatter (2001b). Specifically, a detailed description of a sampling scheme for model (31) is found in Kaufmann (2004). A variant of model (31) where the switching variable is the same for all groups is applied in Frühwirth-Schnatter and Kaufmann (2004).

It is quite a challenge to obtain the marginal likelihood for this model, as it is not possible to integrate analytically with respect to all latent variables λ^N, S^N, and I^T. Nevertheless the bridge sampling method suggested in Frühwirth-Schnatter (2004) may be extended to this model by marginalizing over λ^N and S^N, and applying the bridge sampling technique to the augmented parameter $(I^T, \alpha, \beta_1^G, \ldots, \beta_K^G, \beta_1^R, \ldots, \beta_K^R, \xi^1, \ldots, \xi^K, \eta_1, \ldots$ in a similar way as in Frühwirth-Schnatter (2001a).

6.2 Economic Application: Clustering IP Growth Rates

6.2.1 Model specification

The method proposed here is also helpful in finding answers to some questions that have been raised since the implementation of the single European currency. Are the business cycles of the euro area countries synchronized? Has this synchronization also prevailed in the past. Including overseas countries, one might investigate the relationship between European and overseas countries.

Let y_{it} represent the quarterly growth rate of industrial production of country i in period t. Some business cycle features common to European and overseas countries might be captured by specifying model (31):

$$\begin{aligned} y_{it} &= c_k^G + \delta_{1,k}^G y_{i,t-1} + \ldots + \delta_{p,k}^G y_{i,t-p} \\ &+ (I_{kt} - 1)\left(c_k^R + \delta_{1,k}^R y_{i,t-1} + \ldots + \delta_{p,k}^R y_{i,t-p}\right) + \varepsilon_{it}, \qquad S_i = k, \end{aligned}$$

where $\varepsilon_{it} \sim$ i.i.d. Normal $(0, \sigma_i^2)$ with $\sigma_i^2 = \sigma^2/\lambda_i$. In the investigation, we include 21 countries irrespective of their size, and as usually small or some catching up countries display higher volatility in their growth rates, we specify country-individual variances. With this robust specification we can also account for occasional outliers (due to changing definitions or other unexpected economic breaks).[4]

We estimate the model for a panel of quarterly growth rates covering the period from 1978 through the end of 2002 for all Western European countries and some overseas countries, in particular Australia, Canada, Japan and the United States. As such, the general specification of the model is able to capture the following features of business cycles. Each series is demeaned by its sample average growth rate. The two growth states that we expect to identify should then reproduce common periods of above-average and of below-average growth. The group- and state-specific autoregressive terms might reflect differences in the dynamics of business cycles across country groups and differences between business cycle phases within the groups.

To perform model selection, we estimate the marginal likelihood of various specifications combining $K = 1, 2, 3$, $p = 1, \ldots, 4$. Table 4 contains the estimates. We also estimate the marginal likelihood of a specification assuming that the autoregressive parameters are not group-specific (but switching). The preferred model specification is the one that allows for two groups with one group- and state-specific autoregressive parameter. To save space, the marginal likelihoods for the non-switching specification with and without group-specific dynamics are not reported here. In any case, the switching specification is clearly favored and it is interesting to note that, for all group- and state-specific parameterizations, for a given lag length, i.e. p fixed, it is always the $K = 2$ specification that is chosen, and for K fixed, $p = 1$ always performs best.

[4] Another possibility would be to define obvious outliers as missing values. This approach has been pursued in Frühwirth-Schnatter and Kaufmann (2004); note that this would have to be taken into account in the estimation of the marginal likelihood, however.

Table 4: Log of the marginal likelihood of various Markov switching model specifications with group-specific autoregressive coefficients and without group-specific autoregressive coefficients in parenthesis.

	$p=1$	$p=2$	$p=3$	$p=4$
$K=1$	-4093.55	-4097.57	-4102.61	-4099.25
	(-4094.60)	(-4096.07)	(-4098.60)	(-4091.71)
$K=2$	**-4061.33**	-4074.05	-4081.96	-4084.68
	(-4090.48)	(-4093.92)	(-4096.92)	(-4091.66)
$K=3$	-4064.67	-4082.89	-4093.65	-4101.89
	(-4098.99)	(-4098.93)	(-4101.59)	(-4098.26)

6.2.2 Model identification

To identify the model, two restrictions are necessary. First, the state indicator in each group is identified by means of the constants. State 1 ($I_{kt} = 0$) in each group will refer to below-average growth periods, i.e. the corresponding identification restriction is

$$c_k^R > 0, \qquad \forall k, \tag{32}$$

Secondly, the groups can be identified by means of the autoregressive parameter $\delta_{1,k}^G$:

$$\delta_{1,1}^G < \delta_{1,2}^G. \tag{33}$$

This restriction is motivated by the scatter plot in figure 1 which plots the simulated group-specific values c_k^G against $\delta_{1,k}^G$. Note that the restriction $c_1^G > c_2^G$ would yield the same result. Therefore, group 1 will therefore be the group of countries which grew at a higher unconditional quarterly rate over the observation period and which at the same time display less persistence in the dynamics than the countries of group 2.

6.2.3 Interpretation

Which countries fall into the tow groups? Figure 2 depicts the posterior group probabilities for each country. Note the rather clear inference on the classification for each country, the posterior group probability is 80% in only two cases, Luxembourg and Australia, otherwise the group probability is nearly 1 for one of the two groups. Australia, Canada, Japan and the US fall into the second group, and define what we call the overseas group. Interestingly and in accordance with previous studies, the United Kingdom (UK) follows more closely the overseas group. Over this long-term historical perspective, it appears that Italy and Luxembourg were also moving more closely with overseas countries. In the recent past and due to increased European integration, one would expect that these two countries would also join the bulk of all other European countries which is effectively reported in Kaufmann (2004). Quite remarkably, these countries define a single business cycle pattern already over this long-term historical perspective.

Figure 1: Scatter plot of the simulated group-specific parameters, (c_k^G against $\delta_{1,k}^G$).

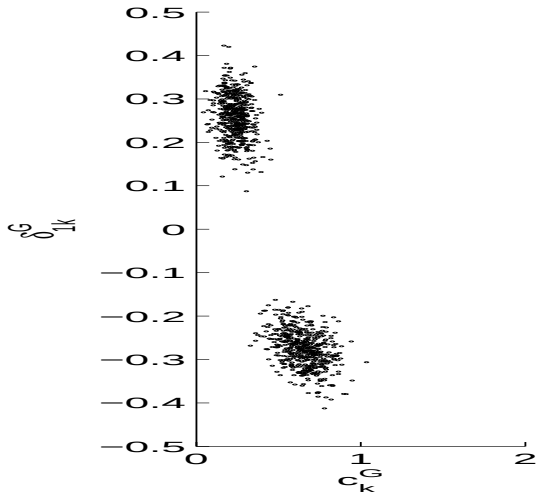


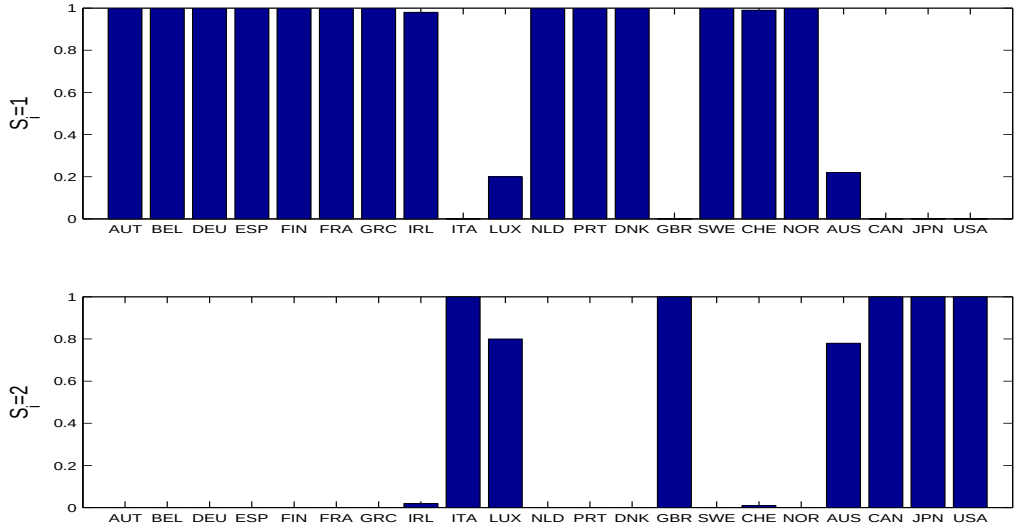
Table 5: Business cycle dating

	P	T	P	T	P	T	P	T	P	T	P	T	P	T
Europ.	80:1			82:4	85:4	87:1	90:4	93:2	95:2	96:1	98:3	99:1	01:1	
Overs.	80:1	80:3	81:3	82:4			90:2	91:1					00:4	01:4
CEPR	80:1			82:3			92:1	93:3						
NBER	80:1	80:3	81:3	82:4			90:3	91:1					01:1	01:4

What characterizes the two groups? Let us first look at the inference on the posterior state probabilities depicted in figure 3. In both groups, the inferred state indicator is able to identify the recession periods also defined by official dating institutions like the NBER and the CEPR (see Table 5). As we identify growth cycles rather than classical cycles only, some periods of growth deceleration are additionally identified for the European countries. Note the changing synchronization between the groups. Up to the 1990s, the overseas cycle was leading the European cycle up to half a year. This leading behaviour disappears during the 1990s, it seems that during this decade, the European countries were more exposed to and affected by some specific shocks (German re-unification, Asian and Russian crisis) than the overseas countries. The recent downturn affected all countries again, which is consistent with analyzes of Helbling and Bayoumi (2003) and Canova et al. (2004).

Table 6 summarizes the posterior mean of the parameters of interest along with the confidence interval in parenthesis. In both groups, the mean below-average growth rate is larger (in absolute terms) than the mean above-average growth rate. An interesting feature of the European countries (group 1) is the estimated negative autoregressive coefficient in industrial production growth rates which turns insignificant during periods of below-average growth rates. This volatile behaviour of European industrial production series may be the result of different data definitions and handling by the national statistical agencies or simply reflect measurement

Figure 2: Mean posterior group probabilities $P(S_i = k|y)$



error. This issue might be worth pursuing but is beyond the scope of the example. Finally, the persistence of above-average growth periods is much higher for the overseas countries. On average, above-average growth periods last for over 3 years, while for European countries the duration averages to about 1 and a half year. On the other hand, below-average growth periods last about half a year longer for European countries than for overseas countries.

Just to give a notion of the variance spread across the countries, figure 4 depicts boxplots of the simulated values for the country-specific variances.

7 Heterogeneity within Clusters

7.1 Heterogeneity in Dynamic Panel Data

The models discussed so far assume that no additional heterogeneity is present within the clusters. The idea of Bayesian clustering may be extended to allow for heterogeneity within each group. The appropriate group-specific time series model is then a random effects model:

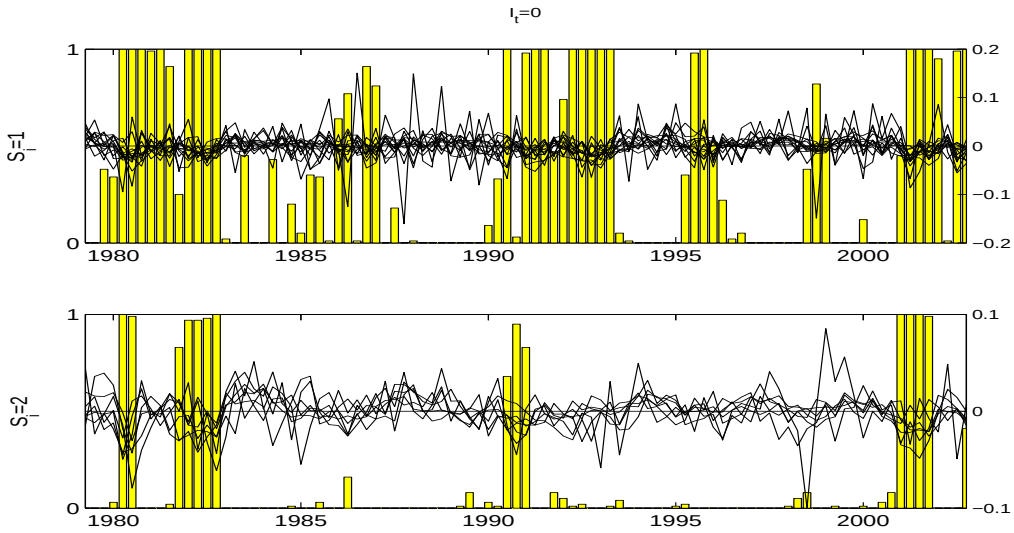
$$y_{it} = X_{it}^1 \alpha + X_{it}^2 \tilde{\beta}_i + \varepsilon_{it}, \tag{34}$$

$$\tilde{\beta}_i \sim \text{Normal}\left(\beta_k, Q_k\right), \qquad \text{if } S_i = k, \tag{35}$$

where X_{it}^1 and X_{it}^2 may again contain lagged values of y_{it}.

Model (35) is an extension of the general heterogeneity model suggested by Verbeke and Lesaffre (1996). A special version where X_{it}^1 and X_{it}^2 are not allowed to depend on lagged values of y_{it} has been applied very successfully in marketing research to deal with unobserved heterogeneity in repeated measurements (Lenk and DeSarbo 2000). Also Canova (2004) used (35) to find convergence clubs in macroeconomic panels, see Section 7.2. In (35), pooling within clusters is substituted by the softer tool of shrinkage within a cluster. The individual coefficients of each

Figure 3: Mean posterior state probabilities $P(I_{kt} = 0|y)$ for European and overseas countries.



time series in group k are pulled toward the group mean β_k, however the presence of a priori variation in $\tilde{\beta}_i$ around β_k allows for differences in the individual $\tilde{\beta}_i$ within the group. The covariance matrix Q_k influences the amount of shrinkage taking place with the limiting case of pooling within clusters if $Q_k = 0$. As the covariance matrices $Q_1, \ldots, Q_K$ are estimated simultaneously with the remaining parameters and latent indicators, the data tell us how much shrinkage is actually needed for the time series at hand.

The MCMC procedure discussed in Subsection 3.1 needs to be extended for model (35) due to the presence of the additional latent parameters $\tilde{\beta}^N = (\tilde{\beta}_1, \ldots, \tilde{\beta}_N)$. As in Lenk and DeSarbo (2000), step (a) - (c) are carried out conditional on $\tilde{\beta}^N$ and $Q_1, \ldots, Q_K$, and $\tilde{\beta}^N$ and $Q_1, \ldots, Q_K$ are sampled from the corresponding full conditional distributions given the actual sampled values of all other parameters and indicators. It has been demonstrated in Frühwirth-Schnatter et al. (2004) that this sampler may be slowly mixing, if the within group heterogeneity is small compared to σ^2. It seems preferable to apply a partially marginalized Gibbs sampler where the random effects $\tilde{\beta}^N$ are integrated out, when sampling S^N and $\alpha, \beta_1, \ldots, \beta_K$, see Frühwirth-Schnatter et al. (2004) for more details.

7.2 Economic Application: Convergence Clubs in Income Data

7.2.1 Model specification and identification

Recent theories of growth and development (Galor 1996 and Temple 1999) suggest the presence of convergence clubs in income data, i.e. a tendency of the stationary distribution to cluster around a small number of poles of attractions. Unit specific characteristics such as the initial condition of income per capita, the dispersion of the income per capita, as well as geographical location may determine, which "club ," i.e. which group the unit will finally join. Empirical studies supporting

Table 6: $K = 2$, $p = 1$: Posterior mean and confidence interval (measured as the shortest interval containing 95% of the simulated values).

	$I_t = 1$		$I_t = 0$			
	$(c^G_{S_i}, \delta^G_{1,S_i})$		$(c^G_{S_i}, \delta^G_{1,S_i}) - (c^R_{S_i}, \delta^R_{1,S_i})$		$(c^R_{S_i}, \delta^R_{1,S_i})$	
	$S_i = 1$	$S_i = 2$	$S_i = 1$	$S_i = 2$	$S_i = 1$	$S_i = 2$
c_{S_i,I_t}	0.66	0.24	-0.91	-1.42	1.57	1.65
	(0.45 0.87)	(0.11 0.37)	(-1.19 -0.67)	(-1.87 -0.92)	(1.34 1.79)	(1.18 2.09)
δ_{1,S_i,I_t}	-0.28	0.26	-0.08	0.30	-0.20	-0.04
	(-0.36 -0.20)	(0.16 0.35)	(-0.17 0.02)	(0.11 0.48)	(-0.33 -0.08)	(-0.26 0.16)
no. countries	14	7				
$\xi_{00}^{S_i}$	0.71	0.67				
conf.int.	(0.55 0.88)	(0.46 0.88)				
quarters	3.44	3.07				
$\xi_{11}^{S_i}$	0.78	0.92				
conf.int.	(0.63 0.92)	(0.85 0.98)				
quarters	4.61	12.87				

this finding are among others Durlauf and Johnson (1995) and Canova (2004). For illustration, we reanalyze the data set considered in Canova (2004)[5]. It consists of yearly per-capita income data for 144 European NUTS2 units and covers the period 1980 - 1992. Not all indicators which may determine group membership of a certain region are available at this regional level and therefore using model-based clustering is quite sensible in this case.

As far as modelling is concerned, we follow closely Canova (2004), but use a totally different approach toward econometric estimation. In contrast to the multi-step estimation procedure of Canova (2004), we use here the fully Bayesian approach discussed in detail in Frühwirth-Schnatter et al. (2004). Following Canova (2004), we allow for K groups whereby in each group heterogeneity is described by an AR(1) process with random coefficients:

$$y_{it} = \tilde{c}_i + y_{i,t-1}\tilde{\delta}_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim \text{Normal}\left(0, \sigma^2/\lambda_i\right), \tag{36}$$

where $\tilde{\beta}_i = (\tilde{c}_i, \tilde{\delta}_i)$, and

$$\tilde{\beta}_i \sim \text{Normal}\left(\beta_k, Q_k\right), \qquad \text{if } S_i = k. \tag{37}$$

y_{it} is the per-capita income of each region relative to the European average. To obtain a certain degree of robustness, we assume, $\lambda_i \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ with $\nu = 8$.

Concerning the prior on S^N, we compare the two specifications introduced in Subsection 2.3, the exchangeable prior (7) which assumes prior ignorance about group membership, and the logit-type prior (8), which includes the initial per-capita income as predictor z_i for group membership. From Table 7 we find, that a model

[5]We kindly thank Fabio Canova for making available his data set.

Figure 4: Country-specific variances, σ^2/λ_i. The box demarcates the lower and the upper quartile, the whiskers show the extent of the simulated values.

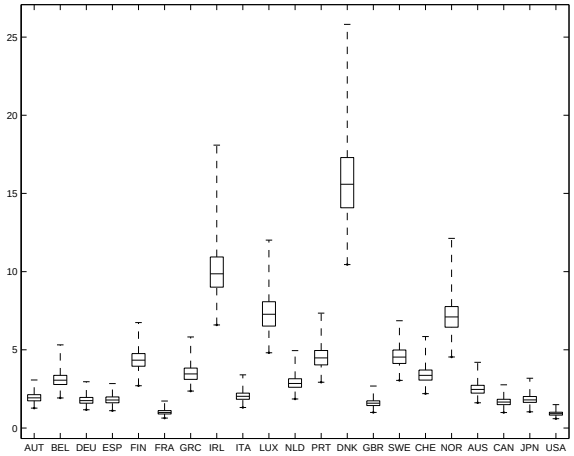


Table 7: Log of the marginal likelihood of various model specifications differing in the number of groups and in modelling the prior of group membership

No Clustering		
$K = 1$		2742.5552
Clustering		
$K = 2$	logit prior	**2748.7182**
$K = 2$	ignorance prior	2735.8733
$K = 3$	logit prior	2748.4384

with two groups which includes initial income into the prior of group membership has highest marginal likelihood among all model specification we considered.

Adding a third group does not lead to an improvement in the marginal likelihood. Also, the posterior draws in the last row of Figure 5 indicate, that it is not possible to identify more than two groups. Either the third group is empty, causing the scattered draws from the prior, or the parameters of the third group are not different from the parameters of the other two groups, making the posterior draws of the third group indistinguishable from the other two groups.

To identify the two groups, for a model with $K = 2$ based on the logit-type prior (8), we consider the posterior draws in Figure 5, where we find significant difference in the first diagonal elements of Q_k in the two groups, which determines heterogeneity in the income level $\tilde{c}_i$. Therefore we use the constraint

$$Q_{1,11} > Q_{2,11} \tag{38}$$

for identification, which corresponds to $\beta_1 < \beta_2$ when $K = 2$ when prior information on group membership is included (see Figure 5, second row).

According to the marginal likelihood in Table 7, the initial per-capita income helps to predict group membership. This is in accordance with the posterior density $p(\gamma_2|y)$ of the relevant parameter γ_2 in the logit prior (8) (see Figure 6, panel

Figure 5: Posterior draws; first row: $K = 2$, ignorance prior (7); second row: $K = 2$, logit prior (8) including initial per-capita income, last row: $K = 3$, logit prior (8) including initial per-capita income

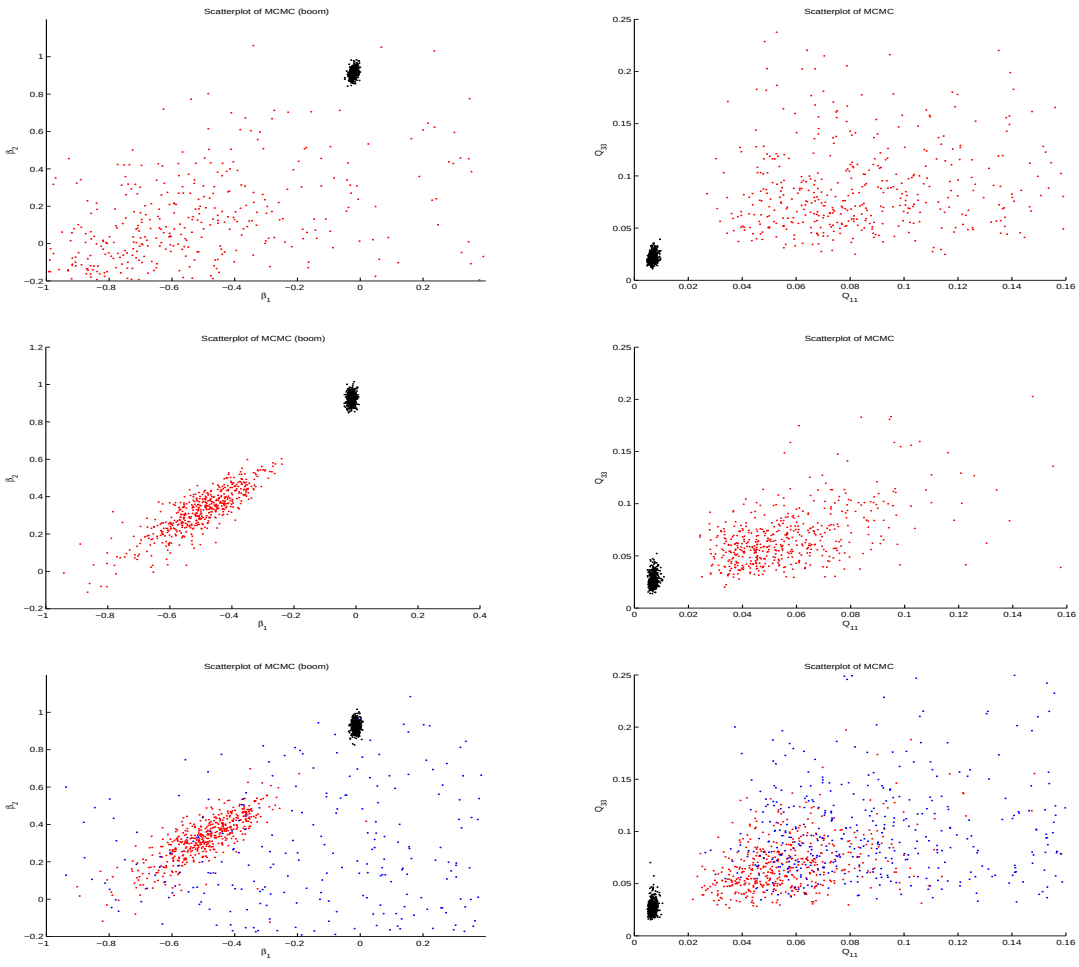
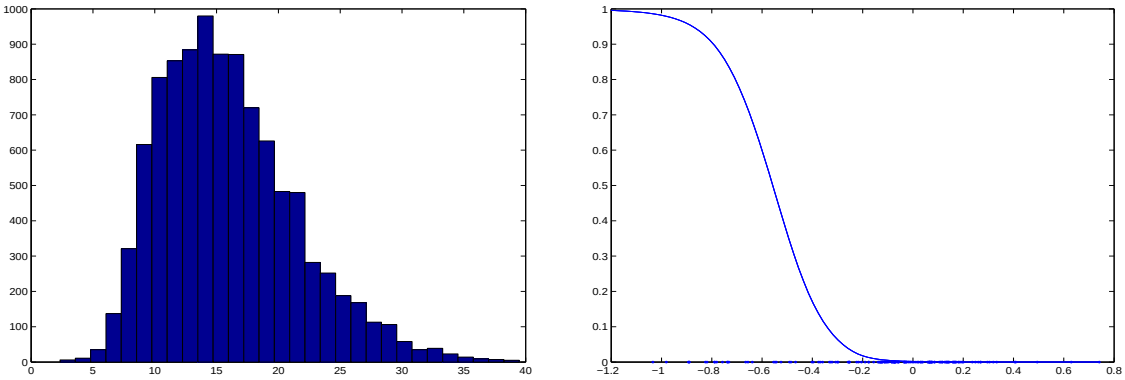
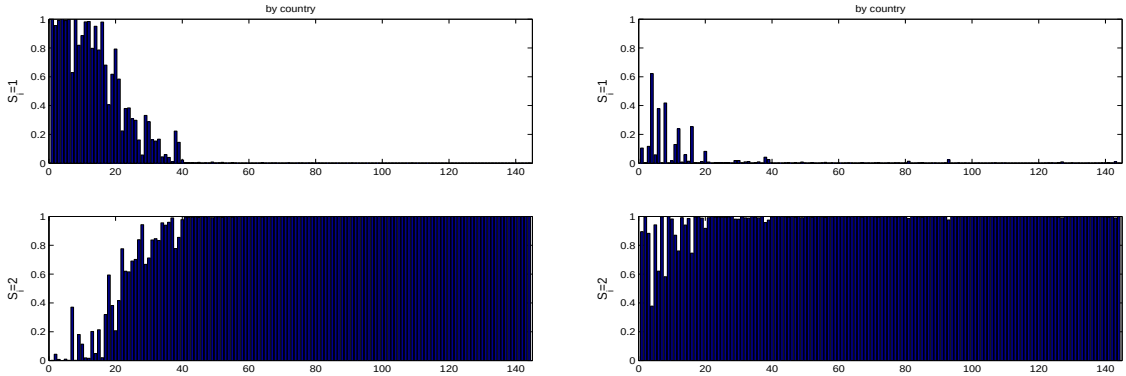


Figure 6: Posterior of γ_2 in the logit prior (8) and resulting prior group probability given initial per-capita income



(a) $p(\gamma_2|y)$ (b) $\Pr\{S_i = 1|\gamma_1, \gamma_2, z_i\}$

Figure 7: Posterior Classification of the regions, ordered according to initial per-capita income; left hand side: classification under the logit prior (8) including initial per-capita income, right hand side: classification under the ignorance prior (7)



(a)), which is strongly shifted away from 0. Nevertheless, the resulting prior group probability $\Pr\{S_i = 1|\gamma_1, \gamma_2, z_i\}$ depicted in panel (b) of Figure 6, nicely shows that initial income is not an absolute criterion for discrimination between the two groups.

It is illuminating to analyze the posterior classifications for each region in Figure 7 under the two priors. Under the ignorance prior, there is only weak information for the presence of two groups, whereas including initial per-capita income helps to identify two groups. These findings are in accordance with the marginal likelihoods in Table 7 which prefers $K = 1$ over $K = 2$ for the ignorance prior, but prefers $K = 2$ over $K = 1$ for the logit prior.

Table 8: Posterior estimates; estimates for $K = 2$ are based on the logit-type prior

	$\mathrm{E}(\tilde{c}_i)$	$\mathrm{E}(\tilde{\delta}_i)$	Q_{11}	Q_{12}	Q_{22}
$K = 1$	-0.0184	0.9330	0.0044	-0.0032	0.0377
$K = 2$, Group 1	-0.4971	0.3231	0.0565	0.0279	0.0684
$K = 2$, Group 2	-0.0139	0.9281	0.0069	0.0005	0.0273

7.2.2 Interpretation

Posterior estimates based on the identification constraint (38) are summarized in Table 8. While Canova (2004) identified four convergence clubs we are able to identify two of them. If we compare the clusters, it turns out that group 1 and group 2 of Canova (2004) (the first 23 units) nearly subsume in our group 1; group 3 and group 4 (units 24-144) gather into group 2 of our posterior inference (see table 1 of Canova 2004). Our mean posterior estimates of the intercept and the persistence in group 1 and 2 nearly correspond to the mean of both posterior estimates of group 1 and 2, and group 3 and 4 of Canova (2004), respectively. Group 1 has a lower mean intercept relative to average and a lower persistence coefficient than group 2. Concerning the dispersion for the coefficients, our estimated dispersion is lower for the parameters of group 2, while for group 1, the dispersion is again approximately the mean of the dispersion for the parameters of group 1 and 2 parameters in Canova (2004).

Therefore, the interpretation of the results can be made along the lines of Canova (2004). We find two convergence clubs with different speeds of adjustments $(1 - \tilde{\delta}_i)$ whereby group 1 is moving more quickly to the group's steady state. The higher dispersion in this group reflects the often very volatile catching-up process of countries having a below-average income initially (see Figure 2 in Canova 2004). The significant difference in the intercept $\tilde{c}_i$ and persistence $\tilde{\delta}_i$ between the groups also confirm the evidence of two convergence clubs. In the long run, the first group's mean steady state is expected to be approximately half (48%) of the regional average.

8 Concluding Remarks

In the present paper we propose to use the attractiveness of pooling time series to obtain posterior inference but without restricting to overall pooling. This means that within a panel of relatively short time series, only those which display "similar" dynamic properties are pooled to estimate the parameters of the generating process. Rather than forming the grouping prior to estimation, we propose to estimate the appropriate grouping along with the model parameters. This is achieved within the Bayesian framework applying Markov chain Monte Carlo simulation methods.

We also discuss two possibilities of designing the prior assumption on each series' group membership: Prior ignorance about group membership is reflected in a probability distribution which is proportional to the groups' relative size within the panel and unit-specific prior information on group membership reflected in a logit-type prior distribution. After estimation, we suggest to use explorative tools like scatter plots and marginal distributions to perform model identification. Model

selection, in particular with respect to the appropriate number of groups, is based on marginal likelihoods.

Three applications illustrate the usefulness of the method. First, a simulation study demonstrates that efficiency gains in estimation and forecasting may be realized when pooling time series with similar dynamics. The simulation study also reveals that pooling time series in different clusters improves efficiency, if there is a minimum heterogeneity between clusters. In the second application we introduce Markov switching within the clusters and investigate a panel of industrial production growth series of all Western European countries and Australia, Canada, Japan and the US. We find two groups of countries and the identified below-average growth states of each group relate to the European in the first group and to the overseas business cycle in the second group, respectively. Finally, we introduce heterogeneity within clusters, whereby pooling is substituted by shrinkage within each cluster. We apply the method to the data set used in Canova (2004) and find similar results on convergence clubs in regional income per capita series.

Overall, the method proposed here relates to other approaches of dimension reduction as in Forni, Hallin, Lippi, and Reichlin (2000), Stock and Watson (2002) and Canova and Ciccarelli (2004).

We want to conclude the paper by discussing potential extensions of our model. We see potential extensions to models with conditional distributions that are non-normal by the nature of the observations y_{it}, an example being panels where y_{it} is a binary indicator or a categorial variable. Clustering of non-normal time series could be carried out as outlined in this paper, with pooling all time series within a group using a logit-, probit- or multinomial model. Such a model could be applied to a large panel of firms' credit risk data.

References

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 88*, 669–679.

Baltagi, B. H. (1995). *Econometric Analysis of Panel Data*. New York: John Wiley & Sons.

Bensmail, H., G. Celeux, A. E. Raftery, and C. P. Robert (1997). Inference in model-based cluster analysis. *Statistics and Computing 7*, 1–10.

Canova, F. (2004). Testing for convergence clubs in income per-capita: A predictive density approach. *International Economic Review 45*, 49–77.

Canova, F. and M. Ciccarelli (2004). Forecasting and turning point predictions in a Bayesian panel VAR model. *Journal of Econometrics 120*, 327 359.

Canova, F., M. Ciccarelli, and E. Ortega (2004). Similarities and convergence in G-7 cycles. Documento de Trabajo 0404, Banco de España.

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*(451), 957–970.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics 75*, 79–97.

Chib, S. (2001). Markov chain Monte Carlo methods: Computation and inference. In J. Heckman and E. Leamer (Eds.), *Handbook of Ecomometrics*, Volume 5, pp. 3569–3649. Amsterdam: North Holland.

Chib, S. and E. Greenberg (1994). Bayes inference in regression models with ARMA(p, q) errors. *Journal of Econometrics 64*, 183–206.

Chib, S., F. Nardari, and N. Shephard (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics 108*, 281–316.

Diebolt, J. and C. P. Robert (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of Royal Statistical Society, Series B 56*, 363–375.

Durlauf, S. and P. Johnson (1995). Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics 10*, 365–384.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic factor model: Identification and estimation. *Review of Economics & Statistics 82*, 540–554.

Frühwirth-Schnatter, S. (1995). Bayesian model discrimination and Bayes factors for linear Gaussian state space models. *Journal of Royal Statistical Society, Series B 57*, 237–246.

Frühwirth-Schnatter, S. (2001a). Fully Bayesian analysis of switching Gaussian state space models. *Annals of the Institute of Statistical Mathematics 53*, 31–49.

Frühwirth-Schnatter, S. (2001b). 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. *The Econometrics Journal 7*, 143–167.

Frühwirth-Schnatter, S. and S. Kaufmann (2004). How do changes in monetary policy affect bank lending? An analysis of Austrian bank data. Oesterreichische Nationalbank, mimeo.

Frühwirth-Schnatter, S., R. Tüchler, and T. Otter (2004). Bayesian analysis of the heterogeneity model. *Journal of Business & Economic Statistics 22*, 2–15.

Galor, O. (1996). Convergence? Inference from theoretical models. *The Economic Journal 106*, 1056–1069.

Garcia-Ferrer, A., R. A. Highfield, F. Palm, and A. Zellner (1987). Macroeconomic forecasting using pooled international data. *Journal of Business & Economic Statistics 5*, 53–67.

Gelfand, A., S. Sahu, and B. Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika 82*, 479–488.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics 8(Supplement)*, 19–40.

Geweke, J. (1995). Bayesian comparison of econometric models. Working Papers 532, Federal Reserve Bank of Minneapolis.

Geweke, J. (1999). Using simulation methods for Bayesian econometric models: Inference, development, and communication. *Econometric Reviews 18*, 1–73.

Goldfeld, S. and R. Quandt (1973). A Markov model for switching regression. *Journal of Econometrics 1*, 3–16.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*, 357–384.

Hamilton, J. D. and B. Raj (2002). New directions in business cycle research and financial analysis. *Empirical Economics 27*, 149–162.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of Econometrics 93*, 345–368.

Helbling, T. and T. Bayoumi (2003). Are they all in the same boat? The 2000-2001 growth slowdown and the G-7 business cycle linkages. Working Paper 03/46, IMF.

Hoogstrate, A. J., F. C. Palm, and G. A. Pfann (2000). Pooling in dynamic panel-data models: An application to forecasting GDP growth rates. *Journal of Business and Economic Statistics 18*, 274–283.

Kaufmann, S. (2000). Measuring business cycles with a dynamic Markov switching factor model: An assessment using Bayesian simulation methods. *The Econometrics Journal 3*, 39–65.

Kaufmann, S. (2004). The business cycle of European countries. Evidence from a panel of GDP series. mimeo, Oesterreichische Nationalbank.

Kim, C.-J. and C. R. Nelson (1999). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. MIT Press.

Koop, G. and H. K. van Dijk (2000). Testing for integration using evolving trend and seasonals models: A Bayesian approach. *Journal of Econometrics 97*, 261–291.

Lenk, P. J. and W. S. DeSarbo (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika 65*, 93–119.

Maddala, G. S. (1991). To pool or not to pool: That is the question. *Journal of Quantitative Economics 7*, 255–264.

McCulloch, R. E. and R. S. Tsay (1994). Statistical analysis of economic time series via Markov switching models. *Journal of Time Series Analysis 15*, 523–539.

McLachlan, G. J. and K. E. Basford (1988). *Mixture models: inference and applications to clustering*. New York/ Basel: Marcel Dekker Inc.

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.

Mittnik, S. (1990). Macroeconomic forecasting using pooled international data. *Journal of Business & Economic Statistics 8*, 205–208.

Nakatsuma, T. (2000). Bayesian analysis of arma-garch models: a Markov chain sampling approach. *Journal of Econometrics 95*, 57–69.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association 67*, 306–310.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of Royal Statistical Society, Series B 59*, 731–792.

Robert, C. P., G. Celeux, and J. Diebolt (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics & Probability Letters 16*, 77–83.

Scott, S. L. (1999). Posterior sampling of multinomial logit models using latent exponential variables.

Shively, T. S. and R. Kohn (1997). A Bayesian approach to model selection in stochastic coefficient regression models and structural time series models. *Journal of Econometrics 76*, 39–52.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*, 147–162.

Temple, J. (1999). The new growth evidence. *Journal of Economic Literature 37*, 112–156.

van Dijk, H. K. and T. Kloek (1980). Further experience in Bayesian analysis using Monte Carlo integration. *Journal of Econometrics 14*, 307–328.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*, 217–221.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

Zellner, A. and C. Hong (1989). Forecasting international growth rates using Bayesian shrinkage and other procedures. *Journal of Econometrics 40*, 183–202.

Zellner, A., C. Hong, and C. Min (1991). Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameters, and pooling techniques. *Journal of Econometrics 49*, 275–304.