Model-based Clustering of non-Gaussian Panel Data

Miguel A. Juárez and Mark F. J. Steel*
University of Warwick

Abstract

In this paper we propose a model-based method to cluster units within a panel. The underlying model is autoregressive and non-Gaussian, allowing for both skewness and fat tails, and the units are clustered according to their dynamic behaviour and equilibrium level. Inference is addressed from a Bayesian perspective and model comparison is conducted using the formal tool of Bayes factors. Particular attention is paid to prior elicitation and posterior propriety. We suggest priors that require little subjective input from the user and possess hierarchical structures that enhance the robustness of the inference. Two examples illustrate the methodology: one analyses economic growth of OECD countries and the second one investigates employment growth of Spanish manufacturing firms.

KEYWORDS: autoregressive modelling; employment growth; GDP growth convergence; hierarchical prior; model comparison; posterior propriety; skewness.

1 Introduction

Models for panel or longitudinal data are used extensively in economics and related disciplines (Baltagi, 2001; Hsiao, 2003; Nerlove, 2002), as well as in health and biological sciences (Diggle *et al.*, 2002; Weiss, 2005).

Typically, panels are formed according to some criteria (e.g. geographical, economical, demographical, etc.) with the intention of gaining strength when estimating quantities common to all individual units in the panel. However, this grouping may strongly affect inference if presumed common characteristics of the units are, in reality, quite different. In these cases, clustering units within the panel may prove useful. This will allow the units to share some common parameters, thus borrowing strength in its estimation, but to also have some cluster-specific parameters (Banfield and Raftery, 1993; Fraley and Raftery, 2002). In an economic context, Bauwens and Rombouts (2006) proposes a method for clustering many GARCH models, while Frühwirth-Schnatter and Kaufmann (2004) discuss a Bayesian clustering method for multiple time series data.

*Corresponding author: Mark Steel, University of Warwick, Department of Statistics, CV4 7AL, Coventry, UK. Email: M.F.Steel@stats.warwick.ac.uk

Even though the majority of the literature uses Gaussian models, it is often the case that data contain outliers, which can be dealt with by allowing for a thicker-than-normal tail behaviour, as well as asymmetries, which require the underlying distribution to allow a certain amount of skewness. The former issue is frequently addressed by assuming a Student distribution with ν degrees of freedom (denoted here by t_ν), usually with ν fixed at a small value. In comparison, there has been much less development in dealing with asymmetry. Hirano (2002) proposes a semiparametric framework, with a nonparametric distribution on the error term, using a Dirichlet prior. In this paper we will use fully parametric, yet flexible, models, partly based on the models described in Juárez and Steel (2006), yet allowing for clustering, and conduct inference from a Bayesian viewpoint.

As the aims of this paper are rather similar to those of Frühwirth-Schnatter and Kaufmann (2004), we briefly highlight the differences with the approach used in that paper. Firstly, our modelling allows for skewness and imposes stationarity. In addition, we use shrinkage within the clusters only for the equilibrium levels, whereas we pool for the autoregressive coefficients. Frühwirth-Schnatter and Kaufmann (2004) either shrink or pool both. The priors used in the present paper are carefully elicited and are improper, unlike the conditionally natural-conjugate prior used in Frühwirth-Schnatter and Kaufmann (2004). This implies we need to make sure that the posterior exists (we derive simple and easily verifiable conditions for propriety), but we need to elicit fewer hyperparameters and, more importantly, our priors enjoys a natural invariance (for those parameters we are improper on) with respect to transformations of the data, which lead to desirable robustness properties. Whenever we use proper priors on cluster-specific parameters, we reduce the dependence of the Bayes factors on prior assumptions by using hierarchical prior structures. Finally, we allow for the data to inform us on the tails of the error distribution, as we leave ν a free parameter.

An important contribution of this paper is the introduction of a flexible model that can be applied in a wide variety of economic contexts with a "benchmark" prior that will be a reasonable reflection of prior ideas in many applied situations. Thus, the aim is to provide a more or less "automatic" Bayesian procedure, that can be used by applied researchers without substantial requirements for prior elicitation. In addition, we provide simple and easily checkable conditions for the existence of a well-defined posterior distribution. However, we also clearly indicate the limits of such procedures, especially in terms of model comparison (or model averaging), which is formally conducted through Bayes factors. Thus, we present two prior structures, one with a flat improper prior on the long-run mean levels of each cluster, which does not require subjective prior input from the user, but does not permit model comparison between models with different numbers of components (unless the levels of all components are assumed equal). The second prior structure asks the user for a mean and a variance of the long-run equilibrium levels, and allows for model comparison. Proper priors on the model-specific parameters are given a hierarchical structure. This leads to greater flexibility, and, more importantly, reduces the dependence of posterior inference and especially Bayes factors on prior assumptions, thus inducing a larger degree of robustness. Matlab code which implements the methodology described in this paper is freely available at http://www.warwick.ac.uk/go/msteel/steel_homepage/software/.

The rest of the paper is organized as follows: Section 2 describes the basic autoregressive model

and its extension to allow for clusters within the panel, and discusses the prior specification and posterior propriety. Numerical methods for conducting inference with this model are briefly discussed in Section 3. Two data sets are analysed in Section 4 to illustrate the implementation of the model: one comprising GDP growth data for OECD countries and the other analyses employment growth in Spanish manufacturing firms. Concluding remarks are presented in Section 5. Proofs are given in Appendix A without explicit mention in the text.

2 The model

Assume that the data available, $\boldsymbol{y} = \{y_{it}\}$ form a (possibly unbalanced) panel of $i = 1, \ldots, m$ individuals for each of which we have T_i consecutive observations. We will focus on the first-order autoregressive model:

$$y_{it} = \beta_i (1 - \alpha) + \alpha y_{it-1} + \lambda^{-\frac{1}{2}} \varepsilon_{it}, \tag{1}$$

where the errors $\{\varepsilon_{it}\}$ are independent and identically distributed (iid) random quantities centred at zero with unit precision, and α is the parameter governing the dynamic behaviour of the panel. We assume that the process is stationary, *i.e.* $|\alpha| < 1$. The intercepts β_i then indicate the long-run tendencies of the observables and are often called individual effects.

In order to accommodate skewness, we assume that the error term follows a skew distribution as in Fernández and Steel (1998), defined by

$$Sf(s \mid \gamma) = \frac{2}{\gamma + \gamma^{-1}} \Big[f(s\gamma)\, 1_{[s \le 0]} + f(s\gamma^{-1})\, 1_{[s > 0]} \Big]. \tag{2}$$

where f is a unimodal probability density function with support on the real line and symmetric around zero, $1_{[x]} = 1$ if condition x holds and 0 otherwise, and $\gamma > 0$ is the skewness parameter. Clearly, for $\gamma = 1$ the density simplifies to f, and for $\gamma \neq 1$ we have skewness, characterised by $P(s > 0|\gamma) = \gamma^2/(1 + \gamma^2)$. Positive skewness corresponds to $\gamma > 1$, while negative skewness is generated by $\gamma \in (0, 1)$. Fernández and Steel (1998) derive an explicit expression for the moments in terms of the moments of f.

In order to also allow for fat tails, we will focus on skew versions of the Student-t_ν distribution, leading to

$$\mathrm{Skt}(\varepsilon \mid \gamma, \nu) = \frac{2}{\gamma + \gamma^{-1}} \frac{\Gamma[(\nu+1)/2]}{\Gamma[\nu/2]} \sqrt{\frac{1}{\nu\pi}} \left[1 + \frac{1}{\nu} \varepsilon^2 \Big(\gamma^2\, 1_{[\varepsilon \le 0]} + \gamma^{-2}\, 1_{[\varepsilon > 0]} \Big) \right]^{-\frac{\nu+1}{2}}, \tag{3}$$

where the degrees of freedom ν will be treated as a free parameter.

Thus, we will use (1) with ε_{it} distributed according to (3), for unit $i = 1, \ldots, m$, and with $T_1, \ldots, T_m$ consecutive measurements in time. This parameterisation allows for a clear interpretation of α as the parameter governing the dynamics of the panel, λ as driving the precision in the measurements and β_i as an individual location (level) or individual effect. In addition, γ will control the skewness and ν determines the tail behaviour. Since the error distribution has a mode at zero,

individual effects are interpreted as the long-run modal tendencies of the corresponding observables. In addition, the individual effects are assumed to be related according to $\beta_i \sim \mathrm{N}\left(\beta_i \mid \beta, \tau^{-1}\right)$, which is a commonly used normal random effects specification, found e.g. in Liu and Tiao (1980), Nandram and Petruccelli (1997) and Gelman (2006), where β is a common mean and τ the precision. Within a Bayesian framework, this is merely a hierarchical specification of the prior on the β_i's, which puts a bit more structure on the problem and allows us to parameterise the model in terms of β and τ, rather than all m individual effects. Finally, we assume that the initial observed value for individual i is y_{i0}, on which we condition throughout, and that the process started a long time ago.

Pooling similar time series can be beneficial when estimating a model, but when the behaviour is not homogeneous enough, the resulting pooled estimates may be misleading, as will be illustrated in the examples in the sequel. Clustering is one way to keep the advantages of pooling, while also allowing for heterogeneity within the panel (see e. g. Canova, 2004; Frühwirth-Schnatter and Kaufmann, 2004; Hoogstrate *et al.*, 2000). In order to allow for clustering within the panel, we assume that all units share a common parameter $\boldsymbol{\theta}^C$ and each has a cluster-specific parameter $\boldsymbol{\theta}^j$, for $j = 1, \ldots, K$, with K the number of clusters in the panel.

Specifically, we assume that the different behaviour may arise either from the dynamics and/or from the equilibrium level of the series. So, extending (1) to allow for different dynamics and levels for each cluster yields

$$y_{it} = \beta_i \left(1 - \alpha_j\right) + \alpha_j y_{it-1} + \lambda^{-\frac{1}{2}} \varepsilon_{it}, \tag{4}$$

with $\left|\alpha_j\right| < 1$ and

$$\beta_i \sim \mathrm{N}\left(\beta_i \mid \beta^j, \tau^{-1}\right) \; ; \qquad j = 1, \ldots, K. \tag{5}$$

Thus, $\boldsymbol{\theta}^C = \{\gamma, \nu, \lambda, \tau\}$ and $\boldsymbol{\theta}^j = \left\{\alpha_j, \beta^j\right\}$. Alternative specifications for the common and the cluster-specific parameters are straightforwardly accommodated within this framework, and in the sequel we will also consider an alternative partition of the parameters, where only the dynamics are cluster-specific, *i.e.* the model with $\beta^j = \beta$, $j = 1, \ldots, K$, leading to $\boldsymbol{\theta}^C = \{\beta, \gamma, \nu, \lambda, \tau\}$ and $\theta^j = \alpha_j$.

2.1 Prior specification

Juárez and Steel (2006) specify a prior for model (1), (3) of the product form

$$\pi\left(\alpha, \beta, \tau, \lambda, \gamma, \nu\right) = \pi\left(\alpha\right) \, \pi\left(\beta\right) \, \pi\left(\tau\right) \, \pi\left(\lambda\right) \, \pi\left(\gamma\right) \, \pi\left(\nu\right) \; , \tag{6}$$

with a standard diffuse (improper) prior for (β, λ), which is invariant with respect to affine data transformations. Theorem 1 will provide a simple sufficient condition for posterior existence under this improper prior. For τ, however, we need a proper prior and we adopt a gamma distribution with shape parameter 2 and a scale that is consistent with the observed between-group variance of the group (*i.e.* individual) means, s_β^2, by making the prior mode equal to $2/s_\beta^2$. The prior on γ is induced by a uniform prior on the skewness measure defined as one minus twice the mass to the left of the mode. The dynamics parameter α gets a rescaled Beta prior (on $(-1, 1)$), and we make the hyperparameters of this Beta distribution random, with equal gamma priors, truncated to be larger than one. The latter

truncation is important in ensuring posterior existence. The hierarchical structure of the prior on α leads to more flexibility. Finally, for ν we take a gamma prior with mass covering a large range of relevant values (prior mean 20 and variance 200). Full details are provided in Juárez and Steel (2006).

Thus, the components of the prior (6) are given by

$$\pi(\beta)\,\pi(\lambda) \propto \lambda^{-1}\,. \tag{7}$$

$$\pi\left(\tau \mid s_\beta^2\right) = \left(s_\beta^2/2\right)^2\,\tau\,\exp\left[-\frac{s_\beta^2}{2}\,\tau\right]. \tag{8}$$

$$\pi(\gamma) = 2\,\gamma\,\left(1+\gamma^2\right)^{-2}\,. \tag{9}$$

$$\pi(\alpha \mid a_\alpha, b_\alpha) = \frac{2^{1-a_\alpha-b_\alpha}}{B(a_\alpha, b_\alpha)}\,(1+\alpha)^{a_\alpha-1}(1-\alpha)^{b_\alpha-1} \qquad |\alpha|<1\,,\quad a_\alpha, b_\alpha > 1\,, \tag{10}$$

$$\pi(\nu) = \frac{\nu}{100}\,\exp[-\nu/10]\,. \tag{11}$$

with

$$\pi(a_\alpha) \propto a_\alpha\,\exp[-0.1\,a_\alpha]\,,\ a_\alpha > 1 \qquad \text{and} \qquad \pi(b_\alpha) \propto b_\alpha\,\exp[-0.1\,b_\alpha]\,,\ b_\alpha > 1\,. \tag{12}$$

In the context of our clustering model in (4) and (5), we will use a direct extension of this specification and use independent, identical priors for the cluster specific parameters, thus for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$, $\boldsymbol{a_\alpha} = (a_{\alpha_1}, \ldots, a_{\alpha_K})'$, $\boldsymbol{b_\alpha} = (b_{\alpha_1}, \ldots, b_{\alpha_K})'$ and $\boldsymbol{\beta} = (\beta^1, \ldots, \beta^K)'$ we have

$$\pi(\boldsymbol{\beta}, \lambda) \propto \lambda^{-1}, \text{ and} \tag{13}$$

$$\pi(\boldsymbol{\alpha}) = \prod_{j=1}^{K} \pi\left(\alpha_j \mid a_{\alpha_j}, b_{\alpha_j}\right) \tag{14}$$

$$\pi(\boldsymbol{a_\alpha}, \boldsymbol{b_\alpha}) = \prod_{j=1}^{K} \pi\left(a_{\alpha_j}\right)\pi\left(b_{\alpha_j}\right) \tag{15}$$

with each component prior specified as above.

In order to complete the mixture model, we need to specify a prior on the assignment of units to clusters. A common approach in the literature is to augment the data with the indicator function $S_i \in \{1, \ldots, K\}$, where $S_i = j$ means that unit i belongs to cluster j. Thus, we may write

$$p(\boldsymbol{y}_i \mid S_i, \boldsymbol{\theta}) = p(\boldsymbol{y}_i \mid \boldsymbol{\theta}^j, \boldsymbol{\theta}^C) \text{ for } S_i = j,\ j = 1, \ldots, K,$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}^C, \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K)$.

A priori we assume that independently

$$\mathrm{P}[S_i = j \mid \boldsymbol{\eta}] = \eta_j,$$

where η_j is the relative size of cluster $j = 1, \ldots, K$ and $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_K\}$. Obviously, $\boldsymbol{\eta}\cdot\boldsymbol{\iota}' = 1$ (where $\boldsymbol{\iota}$ denotes a K-dimensional vector of ones) and thus it is natural to specify the Dirichlet prior

$\pi(\boldsymbol{\eta}) = \text{Di}(\boldsymbol{\eta} \mid \boldsymbol{e})$, where we will use a "Jeffrey's type" prior with $\boldsymbol{e} = (1/2) \times \boldsymbol{\iota}$ (see Berger and Bernardo, 1992). In addition, we exclude from the sampler cluster assignments that do not lead to a proper posterior (as will be explained in Subsection 2.3). Therefore, the joint prior for $\boldsymbol{S} = \{S_1, \ldots, S_m\}$ and $\boldsymbol{\eta}$ is

$$\pi(\boldsymbol{S}, \boldsymbol{\eta}) = \prod_{i=i}^{m} \pi(S_i \mid \boldsymbol{\eta}) \; \pi(\boldsymbol{\eta}) \, I(\boldsymbol{S}) \propto \prod_{i=1}^{m} \eta_{S_i} \prod_{j=1}^{K} \eta_j^{-1/2} I(\boldsymbol{S}), \tag{16}$$

where $I(\boldsymbol{S})$ is one if the assigment gives rise to a proper posterior and zero otherwise.

Finally, note that in this finite mixture model with unknown K and common β (*i.e.* $\theta^j = \alpha_j$), the hierarchical prior structure on $\boldsymbol{\alpha}$ will induce less dependence of the Bayes factors between models with different K on the prior assumptions.

2.2 An alternative prior on $\boldsymbol{\beta}$

The long-run equilibrium levels associated with each cluster are often quantities that we possess some prior information about. If so, it may be desirable to introduce that information through an informative prior, rather than the improper uniform prior used in the previous subsection. Another reason for wanting to put a carefully elicited proper prior on $\boldsymbol{\beta}$ is that we typically want to compute Bayes factors between models with different numbers of components. If the components have a common β, that is perfectly feasible with the improper prior in (13), but in the general case where β^j's are cluster-specific, such Bayes factors are no longer defined. Of course, any proper prior on the cluster-specific parameters in $\boldsymbol{\theta}^j$ will give us Bayes factors, but we need to be very careful that the prior on $\boldsymbol{\beta}$ truly reflects reasonable prior assumptions, just like we did (in Juárez and Steel, 2006) for the prior on each α^j, since the Bayes factors will depend crucially on the particular prior used.

Staying within the product form of (13), we propose the following multivariate Normal prior for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \sim \text{N}_K \left(m \boldsymbol{\iota}, c^2 [(1-a)\,\text{I} + a \boldsymbol{\iota} \boldsymbol{\iota}'] \right), \tag{17}$$

where $c > 0$ and $-1/(K-1) < a < 1$. The prior in (17) generates an equicorrelated prior structure for $\boldsymbol{\beta}$ with prior correlation a throughout. Thus, if $a = 0$ we have independent normally distributed β^j's, but if $a \to 1$ they tend to perfect positive correlation. The main reason for allowing for nonzero a becomes obvious when we consider that (17) implies that $\beta^j \sim \text{N}(m, c^2)$, $j = 1, \ldots, K$ and $\beta^i - \beta^j \sim \text{N}(0, 2c^2(1-a))$, $i \neq j$, $i, j = 1, \ldots, K$. Thus, for $a = 0$ the prior variance of the difference between the equilibrium levels of two clusters would be twice the prior variance of the level of any cluster. This would seem counterintuitive, and positive values of a would be closer to most prior beliefs. In fact, $a = 3/4$, leading to $\text{Var}(\beta^i - \beta^j) = (1/2) \times \text{Var}(\beta^j)$ might be more reasonable.

As we typically will have a fair amount of sample information on β^j, we can go one step further and, rather than fixing a at, say, a reasonable positive value, we can keep a random and put a prior on it. This implies an additional level in the prior hierarchy and would allow us to learn about a from the data. We put a beta prior on a, rescaled to the interval $(-1/(K-1), 1)$, and posterior inference on a then provides valuable information regarding the assumption that all β^j's are equal. In particular, if we

find a lot of posterior mass close to one for a, that would imply that a model with $\beta^j = \beta$, $j = 1, \ldots, K$ (where only the α_j's differ across clusters) might be preferable to the model with cluster-specific β^j's.

As an important bonus of such a hierarchical prior structure, the sensitivity of the Bayes factors to the prior assumptions will be much reduced. In particular, in the model with cluster-specific β^j's, Bayes factors between models with different K depend on the prior on $\boldsymbol{\beta}$ mostly through the implied prior on the contrasts $\beta^i - \beta^j$. If the prior $\pi\left(\beta^i - \beta^j\right)$ is unreasonably vague (corresponding to a very far from 1), we will tend to favour smaller values of K, whereas for excessively precise $\pi\left(\beta^i - \beta^j\right)$ (*i.e.* a very close to 1), Bayes factors would point to models with more components. By changing a we can thus affect model choice, and making a largely determined by the data reduces the dependence of Bayes factors on prior assumptions.

2.3 Propriety of the posterior

Note that (13) yields an improper joint prior. Juárez and Steel (2006) proves that in the extreme case where $K = 1$ a sufficient condition for the posterior to be proper is that $\mathcal{T} > m + 1$, where $\mathcal{T} = \sum_{i=1}^{m} T_i$ is the total number of observations in the sample. The following result under the prior assumptions in Subsection 2.1 can be derived easily from their Theorem 1:

Theorem 1.
Consider the model defined by (4) *and* (5)*, with the error term distributed according to* (3)*, and the prior specification* (6) *through* (16)*. Define* $m_j = \sum_{i=1}^{m} 1_{[S_i=j]}$*, the number of units assigned to cluster* j*, and let* $\mathcal{T}_j = \sum_{i=1}^{m} T_i 1_{[S_i=j]}$ *denote the number of available observations for cluster* j*. If for every* $j = 1, \ldots, K$, $\mathcal{T}_j > m_j + 1$*, then the posterior is proper.*

This condition is not very strong and is trivial to check. It implies that $\mathcal{T} > m + K$ and is satisfied if each cluster has at least one unit with more than two observations. It excludes empty clusters and it is easy to see that the presence of empty clusters would preclude the existence of a posterior. The condition in Theorem 1 will be imposed in the sampler by truncating the prior in (16) through $I(\boldsymbol{S})$. In practice (as in our examples), this will often merely imply assigning probability zero to empty clusters.

In case we simplify the model by assuming that $\beta^j = \beta$, the condition in Theorem 1 will still be sufficient, as this restriction increases the borrowing of strength between the clusters and can only help posterior existence.

In the case where we adopt a proper prior on $\boldsymbol{\beta}$, as in Subsection 2.2, we can derive the following, even weaker, condition for posterior propriety:

Theorem 2.
Consider the model defined by (4) *and* (5)*, with the error term distributed according to* (3)*, and the prior specification* (6) *through* (16)*, but replacing the flat prior on* $\boldsymbol{\beta}$ *in* (13) *with any proper prior, such as the one in* (17)*. The posterior is proper if and only if* $\mathcal{T}_j > m_j + 1$ *holds for at least one* $j = 1, \ldots, K$.

Theorem 2 provides a necessary and sufficient condition for propriety, which is so weak that any sample with at least one unit with more than two observations will always lead to a posterior. As the prior is only improper on the precision λ, existence of the posterior can only be destroyed by having so few observations that we can find a perfect fit in all clusters. As long as we have one cluster where we can not fit the data perfectly, we have a valid Bayesian analysis. Since there are no cluster-specific parameters with an improper prior, empty clusters will now not preclude Bayesian inference.

If we assume a common level $\beta^j = \beta$, the existence condition of Theorem 2 will continue to hold, as it is a necessary condition for integrating out the precision λ.

3 Model estimation

There is a large literature on mixture models, see *e.g.* the monographs by Titterington *et al.* (1985) and McLachlan and Peel (2000). Diebolt and Robert (1994) and Marin *et al.* (2005) provide an overview from the Bayesian perspective.

3.1 Likelihood

Augmenting the data with cluster indicators S_i as described above, we can write the likelihood as

$$L(\boldsymbol{\theta}, \boldsymbol{S}) = \prod_{i=1}^{m} \prod_{j=1}^{K} p(\boldsymbol{y}_i|\boldsymbol{\theta}^C, \boldsymbol{\theta}^{S_i}),$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT_i})$ and the use of (3) and (4) leads to

$$p(y_{it}|\boldsymbol{\theta}^C, \boldsymbol{\theta}^j) = \frac{\sqrt{2/\pi}}{\gamma + \gamma^{-1}} \frac{(\nu/2)^{\nu/2}}{\Gamma[\nu/2]} \lambda^{1/2} \int_{\Re^+} \omega_{it}^{\frac{\nu-1}{2}} \int_{\Re} \exp\left[-\frac{1}{2}\omega_{it}\left(\nu + \lambda h_{it}^2\right)\right] f_N(\beta_i|\beta^j, \tau^{-1})\, \mathrm{d}\beta_i\, \mathrm{d}\omega_{it}$$

with

$$h_{it} = \left(y_{it} - \beta_i(1-\alpha_j) - \alpha_j y_{it-1}\right)\left(\gamma\, 1_{[h_{it}\leq 0]} + \gamma^{-1}\, 1_{[h_{it}>0]}\right),$$

and where $f_N(x|\mu, \zeta^{-1})$ is the pdf of a normal distribution on x with mean μ and precision ζ.

In the sampling density above, we have used the representation of the Student distribution as a gamma scale mixture of normals, which facilitates the computations. In particular, we will augment with the mixing variables ω_{it} in the sampler. We have also integrated the sampling density in (4) with the random effects distribution in (5). Again, we will include the individual effects β_i in the sampler, which is convenient and also allows for inference on each unit's individual effect.

Analytic solutions for this mixture model are not available and, thus, we will resort to Monte Carlo techniques, briefly described in the next section. When dealing with an unknown number of clusters, two alternative approaches may be followed : direct estimation in the sampler or model comparison. The first involves a Markov chain moving in spaces of different dimensions and is implemented by *e.g.* Green (1995) and Richardson and Green (1997) through reversible jump Markov chain Monte Carlo, while Stephens (2000a) and Phillips and Smith (1996) propose alternative samplers that move

between models. We will adopt the second approach, *i.e.* we fit the model for different values of K and then compute Bayes factors in order to decide which number of clusters performs best, as in Bensmail *et al.* (1997), Frühwirth-Schnatter and Kaufmann (2004) and Raftery (1996). This approach is particularly useful in cases where the clusters have a specific interpretation, as inference given a chosen number of components is immediately available.

3.2 Computational implementation

In order to conduct inference, we will use Markov chain Monte Carlo (MCMC) methods, as is now common in the Bayesian literature on finite mixture models. As most of the ideas can be found in the literature (see *e.g.* Bensmail *et al.*, 1997 and Marin *et al.*, 2005), we will not provide any details in the paper.

As pointed out by Celeux *et al.* (2000), Stephens (2000b) and Casella *et al.* (2004), a number of difficulties may arise when constructing a sampler for a mixture model. In particular, we need to take into account the multimodality of the posterior distribution caused by the invariance under permutation of the cluster labels. To overcome this problem, Diebolt and Robert (1994) propose to impose identifiability constraints, while Celeux *et al.* (2000) and Stephens (2000b) use decision-theoretical based criteria. Casella *et al.* (2004) suggest a method based on an appropriate partition of the space of augmented variables. Casella *et al.* (2002) introduce a perfect sampling scheme, which is not easily extended to non-exponential families, and Frühwirth-Schnatter (2004) proposes a random permutation scheme. A comprehensive discussion is given in Jasra *et al.* (2005).

In our setting, we are interested in differentiating between the components in terms of either dynamic or long-run behaviour. It would not be meaningful to distinguish between the clusters in terms of the weights η_j. Thus, we propose to consider scatterplots of all the draws on $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ before deciding on the labels. This will suggest which of the two sets of parameters ($\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$) are best separated between the clusters, and the one that provides the clearest separation will be used to identify the labels through an order constraint. This can then be done by simply post-processing the MCMC output. In both of the examples in this paper, this convincingly indicates that imposing an identifiability constraint through the dynamics parameter, $\boldsymbol{\alpha}$, is a natural way to identify the labels and does not seem to preclude the chain from adequately visiting the posterior support.

To perform model comparison we use the formal tool of Bayes factors[1]. The Bayes factor between any two models is simply defined as the ratio of the marginal likelihoods[2] for the models. Several ways of approximating the marginal likelihood are available in the literature, see e.g. Chen *et al.* (2000), Chib (1995), DiCiccio *et al.* (1997), Newton and Raftery (1994) and references therein. However, in our case these methods may yield poor results due to the potential multimodality of the posterior. Steele *et al.* (2003) and Ishwaran *et al.* (2001) provide alternative methods for mixture models, but

[1] Posterior odds between any two models are then immediately obtained by multiplying the prior odds with the appropriate Bayes factor. These can then be used either for model comparison or Bayesian model averaging (for inference on quantities that are not model-specific, such as predictive inference). In this paper, we will typically assume unitary prior odds.

[2] The marginal likelihood is the sampling density integrated out with the prior, and is not that easy to obtain from MCMC output.

rely either on being able to integrate out some of the parameters or on the underlying distribution being from the exponential family.

In the sequel, we will use the bridge sampler of Meng and Wong (1996), which is based on importance sampling and a simple identity, and uses a so-called bridge function that helps to link the importance function to the target distribution. Given the complexity of the target distribution, which potentially will have heavy tails and be skewed, we construct the importance function using Student-t_3 distributions, centred at the modal MCMC values, for parameters with support on $\Re$; Gamma densities with parameters matching the first two moments of the MCMC output, for positive parameters; and rescaled Betas, with parameters matching the first two moments of the chain, for the dynamics parameter and a in (17). The variance of these distributions is then doubled to aid sampling from the entire posterior support. This choice is intended to mimic the posterior as closely as possible, while still allowing for easy sampling from the importance density and easy evaluations of the importance function at the chain values. Finally, we use the iterative procedure suggested by Meng and Wong (1996) to calculate the optimal bridge function. Using other special cases of bridge sampling, such as ordinary importance sampling or the harmonic mean estimator (see DiCiccio *et al.*, 1997) always leads to the same conclusions in terms of model choice in the examples that follow.

In the particular case that one model is a simple parametric restriction of another model, we can often compute Bayes factors through the Savage-Dickey density ratio[3]. This way of computing Bayes factors is typically easier and can be more precise than using the methods estimating the marginal likelihoods mentioned above, but is not always applicable.

On the basis of various simulated data sets, we conclude that the numerical methods work well and that the priors described in Subsection 2.1 and 2.2 are reasonable and not overly informative.

4 Examples

Two real data sets are analysed in this section. The first contains GDP growth data from 29 OECD countries. The second is a panel of 738 Spanish manufacturing firms, taken from Arellano (2003, Sec. 6.7), where we model growth of employment. We use both the priors in Subsections 2.1 and 2.2. In the latter case, the induced proper prior on each long-run growth level β^j will be $\mathrm{N}(0.05, 0.03^2)$ for the GDP growth data and $\mathrm{N}(0, 0.05^2)$ for the employment growth. For the correlation parameter a in (17), we will use a uniform prior over $(-1/(K-1), 1)$ in both applications.

MCMC samplers were ran for 170,000 iterations, discarding the first 20,000 and then taking every 10th draw, ending up with an effective size of 15,000. Parameters of the Metropolis-Hastings proposal distributions were tuned as to achieve acceptance rates of around 1/3. In both examples very similar results were obtained using smaller runs of size 70,000, suggesting that convergence was achieved.

[3]The Savage-Dickey density ratio is the ratio of the posterior and the prior density values at the restriction (see Verdinelli and Wassermann, 1995). For example, the Bayes factor in favour of a symmetric model over its skewed counterpart will be $p(\gamma = 1|\text{data})/p(\gamma = 1)$.

Table 1. *OECD GDP data. Log Bayes factors for the number of clusters K. A positive number indicates support in favour of the model in the row.*

	K		
K	2	3	4
1	-47.4	45.3	28.3
2		92.7	75.7
3			-16.9

4.1 GDP of OECD countries

There is a vast literature concerned with economic growth and convergence. While there seems not to be empirical evidence of overall growth convergence (Durlauf and Johnson, 1995; Durlauf and Quah, 1999; Temple, 1999), some clusters of homogeneous growing countries or convergence clubs have been found; see e.g. Canova (2004) and Quah (1997).

Here we concentrate on GPD growth rates from 29 OECD countries, taken from the Penn World Table (Heston *et al.*, 2002), for the period 1950-2000. We define the growth of country i from time $t-1$ to t as $y_{it} = \log(x_{it}/x_{it-1})$, where x_{it} is the GDP of country i at time t.

We feel that allowing for $K \geq 5$ would not be of practical interest in this context, so we fit the model for $K = 1, 2, 3, 4$. Initially, we report results for the case of the proper prior on $\boldsymbol{\beta}$ and will indicate any differences with respect to the analysis with a flat prior on $\boldsymbol{\beta}$. At the end of this subsection, we will use the flat prior. Estimated log Bayes factors (BF) are shown in Table 1, a positive value implies support in favour of the model in the row. Interestingly, the simplest, completely pooled model is clearly preferred to $K = 3$ and $K = 4$, but the best model by far is the one with two clusters.

Figure 1 shows scatter and trace plots of the drawn values for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in the chain with two components. The original assignment of labels to drawings is indicated by the use of different shadings and symbols (grey traces and crosses for one cluster and black traces and circles for the other). It is clear from the traces that label switching has occurred (around draw 7000) and the scatterplot vividly illustrates that the dimension in which the components differ is the dynamics parameter α. Thus, we use the labelling convention according to the values of α, and impose that $\alpha_1 < \alpha_2$ in post-processing the data. This perfectly implements the separation between the two visually separate clusters in the scatterplot. Similar pictures appear for the other values of K, so we always use ordering on α to identify the component labels. It is clear from the graphs that identifying the labels through an ordering constraint on β would have made little sense, and would have left us with two (very similar) bimodal distributions on α_1 and α_2.

Figure 2 shows estimated marginal posterior densities for the model-specific parameters of the models with $K = 1, 3, 4$. Throughout, we also plot the prior density in these graphs, indicated by a dashed line. Estimation of the common parameters is virtually unaffected by the number of clusters. Thus, we only present posterior densities for the cluster-specific parameters here and will present those for the other parameters in the figure for the preferred model with $K = 2$. Comparing the plots for $\boldsymbol{\alpha}$ with different K, the effect of pooling when units are not homogeneous is apparent: the pooled model ($K = 1$) averages over the whole panel, yielding misleading inference on the dynamics and an

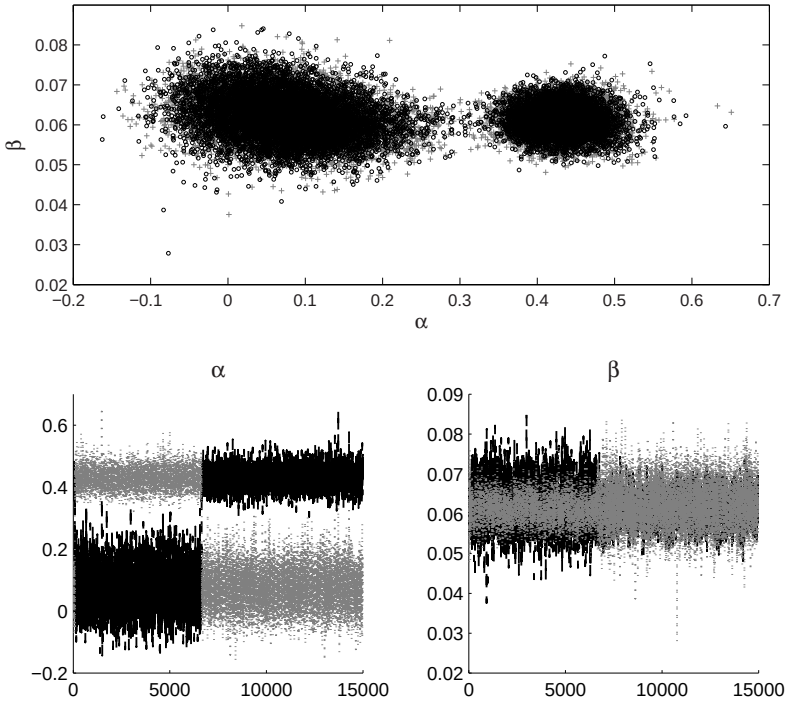

Figure 1. *OECD GDP data. Scatterplot and traces of the sampler for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, using $K = 2$. Different shadings and symbols indicate the raw assignment of labels to drawings, before post-processing.*

illusion of precise estimation. Also, it is clear from the inference on $\boldsymbol{\alpha}$ with $K = 3$ and $K = 4$ that these models contain more clusters than supported by the data, as there is no clear separation between the clusters with higher α_j (and the clusters do not distinguish between β^j either). It is reassuring that model choice through Bayes factors strongly avoids the inclusion of unwarranted clusters in our model. This illustrates, in particular, the sensible calibration of our prior assumptions.

With a decisive BF of 3.9×10^{20}, $K = 2$ is preferred over the pooled model and its superiority over the other alternatives is even more pronounced. Posterior results are displayed in Figure 3. Note that the prior on λ is improper and the scaling is, therefore, arbitrary. For this best model with two clusters, we have a fast converging club of countries, i.e. those with a small value of α, and a slow converging subset, as indicated in the top left graph of Figure 3. The posterior mean relative cluster sizes are $\{0.17, 0.83\}$. In addition, Figure 4 shows the individual membership probabilities. The first club, with an mean value of $\alpha = 0.077$ is the smallest and it is constituted by Spain, Luxembourg and Turkey with membership probabilities of over 0.85, while Mexico and Denmark have a probability of belonging to this cluster of around 0.35 (countries are ordered as in Appendix B). The other club corresponds to a mean value of $\alpha = 0.43$, indicating much slower convergence. The posterior mean and median values of β^j are 0.062 for the first cluster and 0.061 for the second, indicating that both clubs converge (at different speeds) to a long-run growth rate of around 6%. Note that the posterior distribution of α for the pooled model ($K = 1$) in Figure 2 is concentrated around an area which receives only very little probability mass from the posteriors of α_1 and α_2 in the two-component model, so its averaged nature really does not correspond to any "observed" dynamic behaviour.

Figure 3 suggests that skewness is not an important feature of this data set. Indeed, $\gamma \in (0.96, 1.11)$ with posterior probability of 0.95 for all values of K used. As a consequence, the log Savage-Dickey

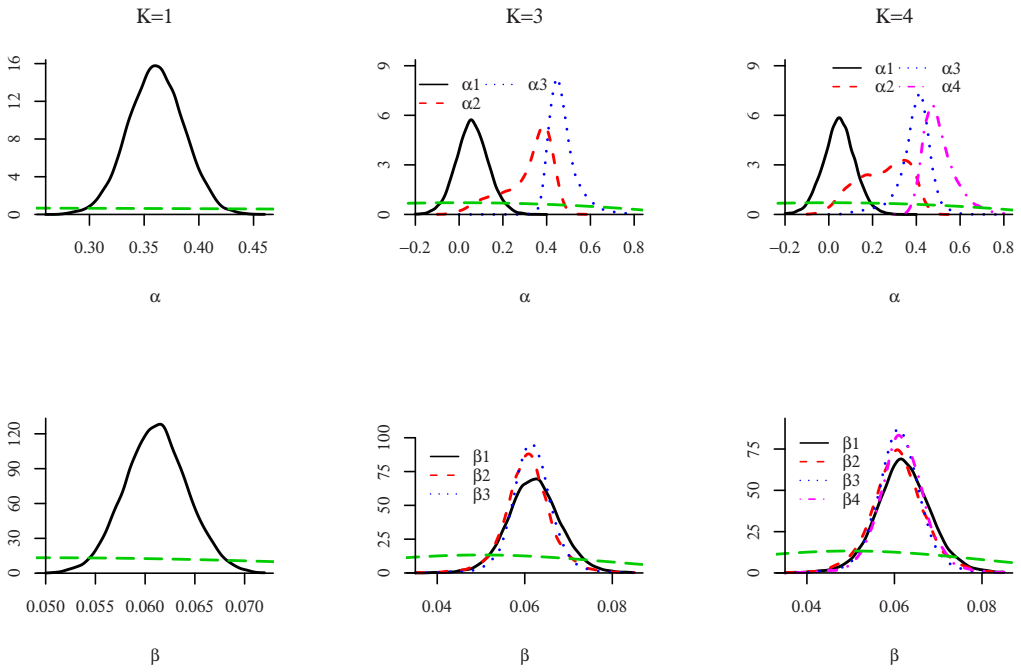

Figure 2. *OECD GDP data. Prior (light dashed) and posterior (solid or as in legend) densities for the model-specific parameters, using* $K = 1, 3, 4$.

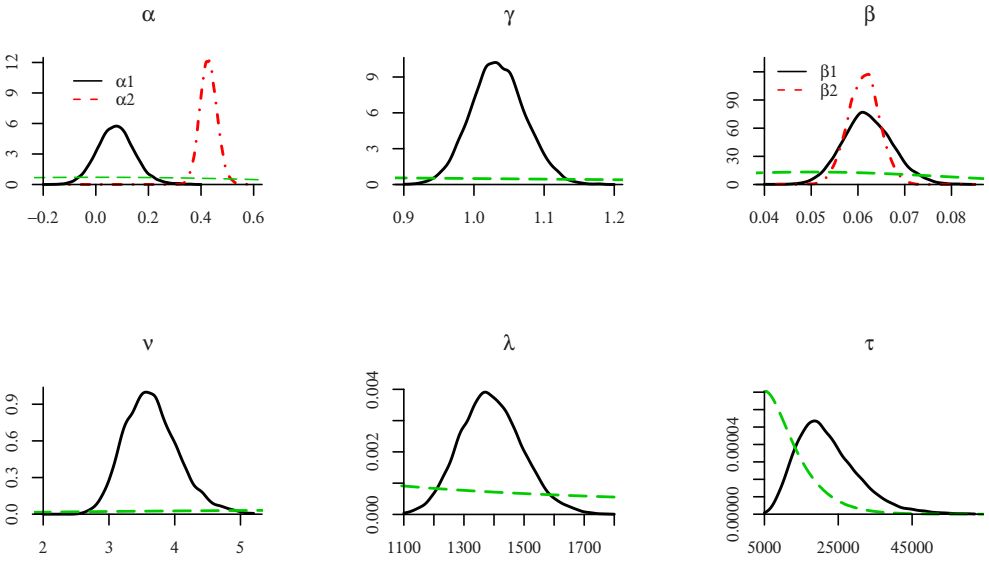

Figure 3. *OECD GDP data. Prior (dashed) and posterior (solid or as in legend) densities, using* $K = 2$.

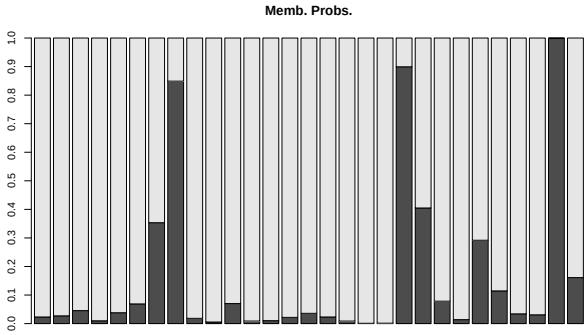

Figure 4. *OECD GDP data. Membership probabilities for $K = 2$. Countries are ordered as in Appendix B.*

density ratio in favour of $\gamma = 1$ is 2.6 (for $K = 2$, with very similar values for other K). Comparing the marginal likelihoods for the symmetric and skewed Student t models, computed using bridge sampling, leads to a log BF of 2.5 in favour of symmetry, which accords very well with the Savage-Dickey result. For the symmetric version, the rest of the estimated parameters are virtually identical. In the remainder of this subsection, we will use the symmetric t model.

Fat tails, however, are a very prominent feature of these data. Posterior inference on ν is quite concentrated on small values in all cases, typically $\nu \in (2.8, 5.0)$ with high posterior probability.

Throughout, posterior results with the flat improper prior on $\boldsymbol{\beta}$ introduced in Subsection 2.1 are virtually indistinguishable from the ones presented here, except, of course, for the fact that the Bayes factors involving the choice of K are not defined for this case.

As already indicated, the posterior distributions of the β^j are always centred around similar values. Recall that the proper prior for $\boldsymbol{\beta}$ introduced in Subsection 2.2 puts a hierarchical prior on the correlation parameter a, where a close to one would indicate similarity of the β^j's. Figure 5 displays the posterior density of a for $K = 2, 3, 4$ and clearly suggests that the data favour values close to one. This becomes especially clear for larger K where we have even more information on a in the data. Thus, we also consider a model where the only difference between the clusters is the dynamics parameters, while the long-run level is shared between the components. The Bayes factor in favour of this model with common long-run mean level versus the original two-component model is 11.9.[4]

Note that if we use a common β, we can compute Bayes factors between models with different numbers of components under a flat improper prior on β. Table 2 presents these Bayes factors for the models with symmetric t errors and common β. The ordering of models is the same as with skewed errors and cluster-specific β^j's: $K = 2$ is strongly preferred to the pooled model, which is itself a lot better that the models with 3 and 4 components.

So we finally model the data using a t_ν model (*i.e.* $\gamma = 1$) with two clusters and $\beta^j = \beta$ for $j = 1, 2$. We now use a flat improper prior on β. As illustrated in Figure 6, the resulting estimates are quite close to those for the previous case (the posteriors for λ and τ are not shown as they are indistinguishable from those in Figure 3). Also, the expected cluster sizes remain the same. A 95%

[4]The latter can easily be computed using the Savage-Dickey density ratio for the difference $\beta^2 - \beta^1$.

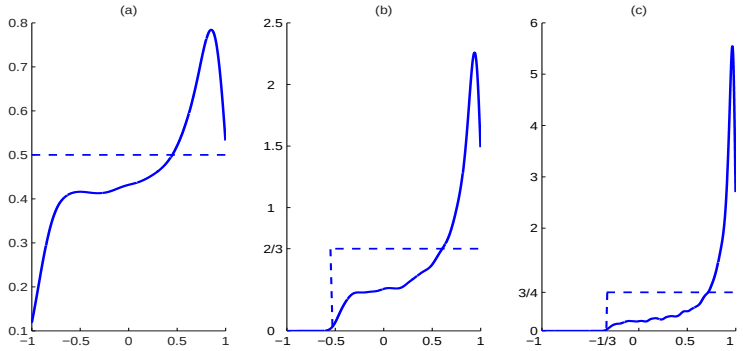

Figure 5. *OECD GDP data. Posterior (solid) and prior (dashed) densities for a in (17). From left to right, panels are for $K = 2, 3$ and 4. Note that $a \in (-1/(K-1), 1)$.*

Table 2. *OECD GDP data. Log Bayes factors for the number of clusters, using the t model with common β, for which we adopt an improper flat prior. A positive number indicates support in favour of the model in the row.*

	K		
K	2	3	4
1	-20.5	88.0	88.8
2		108.5	109.3
3			0.8

posterior credible interval[5] for the new common equilibrium level β is (0.060, 0.068) and β is now more precisely estimated, as all observations contribute.

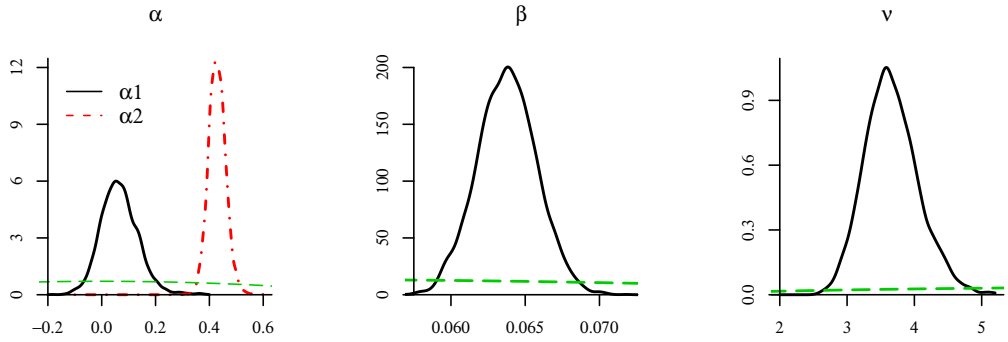

Figure 6. *OECD GDP data. Prior (dashed) and posterior (solid or as in legend) densities, using a t_ν model with $K = 2$ clusters and a common level with an improper flat prior.*

4.2 Spanish firm employment

The data set is described in the Appendix of Alonso-Borrego and Arellano (1999) and also used in Arellano (2003, Sec. 6.7). It consists of a balanced panel of 738 manufacturing companies, recorded

[5]Throughout the paper, 95% credible intervals are defined by the 2.5th and the 97.5th percentiles of the corresponding distribution.

Table 3. *Spanish firm data. Log-BF, according to the number of clusters. A positive figure indicates support in favour of the model in the row.*

K	K			
	2	3	4	5
1	824	-7	3077	5127
2		-830	2253	4304
3			3084	5134
4				2051

yearly from 1983 to 1990 and represents more than 40% of the Spanish value added in manufacturing in 1985.

In particular, we model employment growth in these firms. With our setting described in Section 2, and letting $K = 1$, we obtain the posterior densities shown in Figure 7 for the model-specific parameters. This is computed with the proper prior on β from Subsection 2.2, which will be used until we indicate otherwise. Again, inference on parameters common to models with different values of K is virtually unaffected by the choice of the number of clusters.

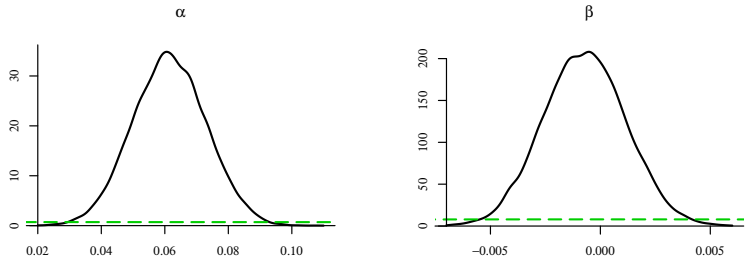

Figure 7. *Spanish firm data. Prior (dashed) and posterior (solid) densities for α and β, using $K = 1$.*

As shown in Table 3, $K = 1$ is strongly preferred to $K = 2$, $K = 4$ and $K = 5$. However, the model with three clusters performs better than the pooled model, and we will concentrate on the model with $K = 3$ in the sequel. Since the model with five clusters was not preferred to any other, we did not experiment with even larger values of K.

Figure 8 shows a scatterplot of the drawn values for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in the chain with three components, clearly illustrating that identifying the labels through ordering the values of α_j is the natural approach, just like in the previous example. The partition of the draws before inducing this labelling convention is indicated through the different hues and symbols. After imposing the ordering constraint the three clusters are well separated.

From the posterior densities in Figure 9, it is apparent that tail behaviour is extremely heavy and very well determined by this (fairly large) data set. In contrast with the GDP data, these data present substantial right skewness with (1.05, 1.13) a 95% credible interval for γ. In this case the log-BF calculated from the Savage-Dickey density ratio in favour of $\gamma = 1$ is -31. As the posterior density value at $\gamma = 1$ is quite small, the Savage-Dickey density ratio is not that easy to estimate in this case. Computing marginal likelihoods through bridge sampling leads to a log BF of -20, which corroborates

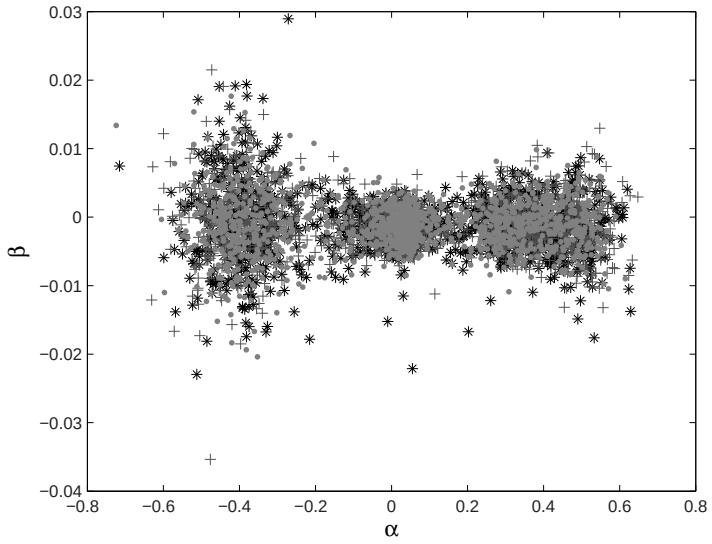

Figure 8. *Spanish firm data. Scatterplot of the sampler for α and β, using $K = 3$. Different shadings and symbols indicate the raw assignment of labels to drawings, before post-processing. For clarity, only a thinned version of the sample (one in 10) is displayed.*

Table 4. *Spanish firm data. 95% posterior credible intervals and means for the dynamics parameters α_j using $K = 3$.*

Cluster	95% Interval	Mean
1	(-0.549 , -0.200)	-0.384
2	(-0.022, 0.073)	0.028
3	(0.222, 0.551)	0.378

the massive evidence in favour of the skewed model.

The relative size of each cluster, *i.e.* the average probability of cluster membership, is {0.132, 0.651, 0.217}. Table 4 presents 95% posterior credible intervals for α_i. Thus, there are two relatively small clusters of "extreme" dynamic behaviour: one with negative α (suggesting alternating behaviour) and one with positive α (slowly converging) existing besides one big club with more or less random walk employment behaviour. In fact, the cluster displaying negative α tends to contain smaller firms, which are more volatile and often overadapt to market situations. Firms that have a high probability of belonging to the slowly converging cluster are typically larger firms which display much more stable long-term employment strategies. The firms in the main cluster cover a wide range of sizes and have, on average, experienced a small decline in employment over the sample period. Again, the effect of pooling all units to estimate the dynamics parameter is apparent from comparing Figures 7 and 9: rather than gaining strength in the process, opposites are averaged out and the spread of the dynamic behaviour is dramatically underestimated when we use $K = 1$.

We have already reported that the skewed model is strongly favoured by the data over its symmetric counterpart. In order to assess whether allowing for skewness makes a practical difference in this example, we have estimated the symmetric Student model (*i.e.* $\gamma = 1$) with 3 components. The main difference is in the locations which are represented by the β^j in our model. The posterior medians for β^j with skewness were all within (-0.0011, -0.0008), and these are now all positive, equal

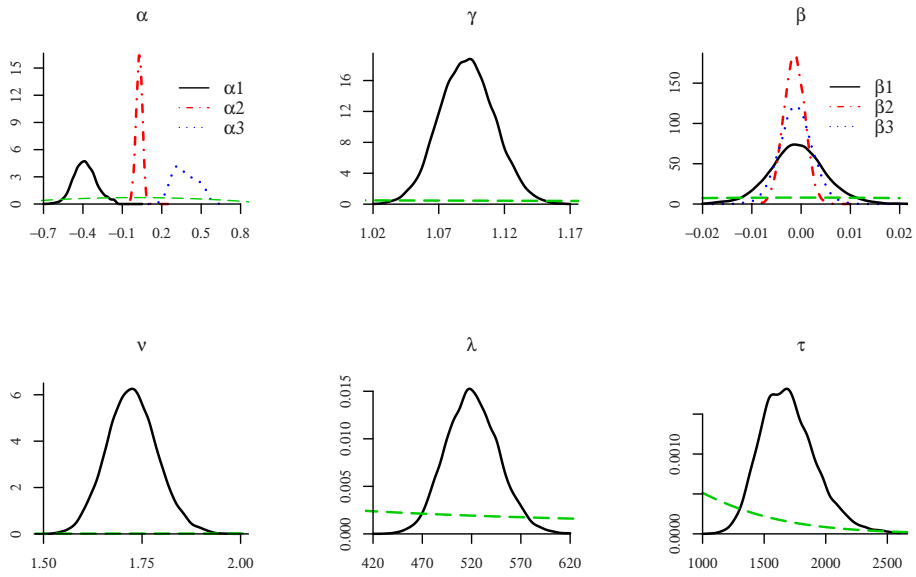

Figure 9. *Spanish firm data. Prior (dashed) and posterior (solid or as in legend) densities, using* $K = 3$.

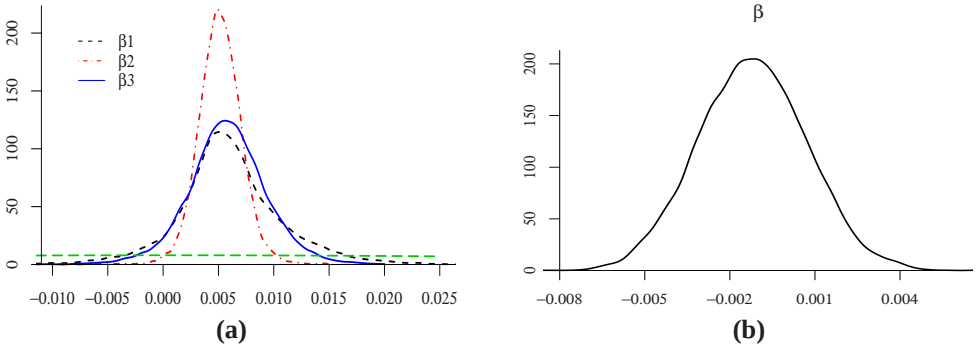

Figure 10. *Spanish firm data. (a) Estimated posterior for* $\boldsymbol{\beta}$*, using* $K = 3$ *and* $\gamma = 1$*. The (proper) prior is dashed. (b) Posterior density for common equilibrium level* β*, using* $K = 3$ *and a flat (improper) prior on* β*.*

to $\{0.0057, 0.0051, 0.0058\}$ whereas the 95% credible interval for β^2 is entirely on the positive real line. Figure 10 (a) shows these posterior densities of $\boldsymbol{\beta}$, which are clearly shifted to the right with respect to the skewed case in Figure 9. Thus, without taking into account the skewness, we would erroneously conclude that long-run employment growth is positive, whereas our skewed model assigns most probability to negative equilibrium growth of employment in Spanish manufacturing firms.

Both for the skewed and symmetric cases, the three clusters of firms converge to very similar equilibrium levels, suggesting that we might also pool this parameter to gain strength, as done in the previous example. The resulting marginal posterior for β when we impose a common equilibrium level is shown in Figure 10 (b), where we have now used the improper flat prior on β. Other parameters are virtually unaffected by this simplification of the model. In particular, the evidence in favour of right skewness does not change and a credible interval of size 0.95 for γ is (1.05, 1.13), as before. The inference on the dynamics parameters is also virtually unchanged from the previous case. The common long-run level $\beta \in (-0.005, 0.003)$ with posterior probability 0.95, very much in line with

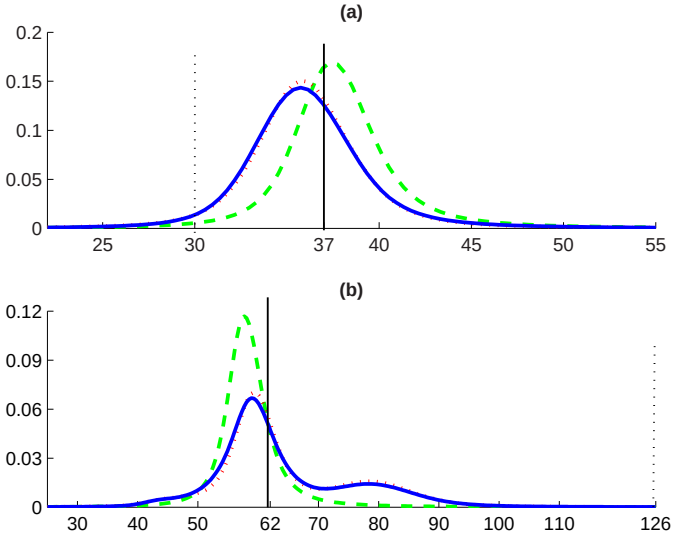

Figure 11. *Spanish firm data. Predictive distribution for 1991 for the employment of firms 433 (a) and 31 (b). Predictives are for $K = 1$ (dashed) and $K = 3$ (solid for skewed model, dotted for symmetric model). Employment numbers for 1989 and 1990 are indicated by dotted and solid vertical lines, respectively.*

the results for cluster-specific β^j's, except that inference is now a bit more precise as a consequence of borrowing strength.

Finally, we calculate the predictive distribution of the employment of two firms in the sample for 1991 (one year after the last observation in the sample), using the flat prior on the common β. As we are predicting employment itself (rather than its growth), we condition on the actual employment values in the sample years. Firms 433 and 31 are selected: the former grows from 30 to 37 employees in 1990 and in the model with $K = 3$ it is assigned to the three clusters with posterior weights $\{0.834, 0.165, 0.001\}$; the latter shrinks its employment in 1990 from 126 to 62 and has cluster probabilities $\{0.324, 0.636, 0.040\}$. Figure 11 presents these predictives for the pooled model ($K = 1$) and the model with three components (a symmetric and a skewed version). The model with $K = 1$ has a slightly positive α (see Figure 7) and will thus concentrate the predictive at a value which slightly extends the last observed movement. In the three-cluster model, Firm 433 (Figure 11 (a)) has most mass on the first cluster, which corresponds to large negative values for α (see Figure 9 and Table 4), and will thus counteract the last movement, which results in much more predictive mass on lower employment values. Firm 31 has non-negligible mass for all three clusters and this results in a multimodal predictive, with the first cluster providing predictive mass around 80 (partially counteracting the last movement) and the third (least important) cluster resulting in slightly more weight on lower values. The latter is a consequence of the large positive values for the dynamics parameters, which lead to a pronounced extrapolation of the last observed change. Finally, the second cluster (which has most of the weight) corresponds to very small, mostly positive values for α (see Table 4), which is translated in the large central mode, close to the last observed value (with a slight extrapolation of the last movement). The clusters vary mostly in terms of the dynamics parameter, so if the observed change is substantial (as is the case for firm 31), multimodality in the predictive is easily generated. It

is clear that the pooled model substantially underestimates the predictive uncertainty and can lead to dramatically different conclusions. Of course, the different firms also have different individual effects β_i, but the effect of those on the one-step ahead predictives shown is dominated by the dynamics: in the three-component model β_{433} has a posterior mean of 0.011 (corresponding to 1% growth) and the mean of β_{31} is -0.025. In case we use the symmetric three-component model ($\gamma = 1$), the posterior means of these long-run levels are changed to 0.026 and -0.018, respectively, which constitutes a rather different picture for the equilibrium situation, especially for firm 433. This would, of course, affect the predictives for long forecast horizons, but short-run forecasting with the symmetric model is not very different from that with the skewed model, as illustrated in Figure 11.

5 Conclusion

This paper deals with model-based clustering of longitudinal data, where the clusters can differ in dynamic and long-run equilibrium behaviour. We adopt flexible error distributions, allowing for fat tails and skewness, each controlled by a single (easily interpretable) parameter. Prior distributions are carefully chosen, to reflect a (commonly encountered) situation without strong prior information. Hierarchical prior structures are used to increase the robustness of our posterior results with respect to prior assumptions. Two prior structures are proposed, giving the applied user the opportunity to conduct inference with these models without spending a lot of effort on prior elicitation. Practically useful and quite mild conditions for the existence of the posterior distribution are provided. We propose to use a scatterplot of the drawn values for the dynamics and long-run level parameters to indicate a solution to the labelling problem.

It would be straightforward to extend the model to let the assignment of observations to clusters depend on covariates: *e.g.* a probit or logit specification would simply add one step to the MCMC sampler. In view of our discussion of the example on Spanish firm employment, it would, for example, be natural to use firm size as a determinant of cluster probabilities in that case. In addition, the inclusion of extra explanatory variables in the regression model in (4), with coefficients that could either be cluster-specific or common to all clusters, is relatively straightforward.

We analyse two real panel data sets: one on GDP growth of OECD countries, with only 29 individual countries, and one concerning employment growth in a much larger sample of 738 manufacturing firms. Both applications favour clustering, with the clusters characterised by different speeds of convergence to a common underlying long-run growth level β. As a consequence, modelling β as common is favoured by both data sets. Ignoring the clustering in the data would result in totally misleading inference of the dynamic behaviour, parameterised by α: the pooled model averages out the dynamic behaviour and does not properly account for the uncertainty. In both examples, the pooled posterior distribution for α is far too sharp, inducing a false sense of security. For the GDP data, where two components are preferred, the pooled model puts virtually all the posterior mass for α in between the clusters, where very little posterior mass is allocated by the clustering model. For the firm employment example, which favours three clusters, the pooled posterior on α is a slightly shifted and more concentrated version of the central cluster's dynamics, but the two other clusters are totally

overlooked by the pooled model. The effect of this is perhaps best appreciated by considering the predictive distribution: the shape, location and concentration of the latter are often very different for the pooled model, as illustrated here for the firm data. In the firm application skewness also matters; not just statistically, but also in terms of the conclusions we would draw from the data: equilibrium growth levels are quite different if we ignore the skewness, in that they would point to overall long-run employment growth rather than contraction.

Appendix A Proofs

A.1 Proof of Theorem 1

The proof relies on Theorem 1 of Juárez and Steel (2006), which states that $\mathcal{T} > m+1$ is sufficient for propriety in the case with $K = 1$. Given a particular cluster assignment $\boldsymbol{S}$, we simply apply this result to any cluster j separately, leading to $\mathcal{T}_j > m_j + 1$ as a sufficient condition for existence. The fact that the parameters in $\boldsymbol{\theta}^C$ are shared between the clusters can only help existence, so the condition in Theorem 1 is definitely sufficient. Finally, we mix over $\boldsymbol{S}$ with its proper prior. Technically, the prior on $\boldsymbol{S}$ as defined in (16) is truncated to impose this condition.

A.2 Proof of Theorem 2

We use Theorem 1 of Fernández and Steel (1998) which states that the posterior of the skew model is proper if and only if it is proper when $\gamma = 1$. Given the proper prior on γ, we can thus concentrate on the symmetric model. So we need to evaluate

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\xi}, \lambda, \boldsymbol{S}, \boldsymbol{\omega})\lambda^{-1} p(\boldsymbol{\xi}, \boldsymbol{S}, \boldsymbol{\omega})\, \mathrm{d}\lambda\, \mathrm{d}\boldsymbol{\xi}\, \mathrm{d}\boldsymbol{S}\, \mathrm{d}\boldsymbol{\omega},$$

where $\boldsymbol{\xi} = \{\gamma, \nu, \tau, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ and $\boldsymbol{\omega} = \{\omega_{1\,1}, \ldots, \omega_{1\,T_1}, \ldots, \omega_{m\,T_m}\}$. We can derive $p(\boldsymbol{y}|\boldsymbol{\xi}, \lambda, \boldsymbol{S}, \boldsymbol{\omega})$ from the fact that, given a cluster assignment, we can write the model for each cluster as $\boldsymbol{y}^j = X_j\, \boldsymbol{\zeta}_j + \lambda^{-1/2}\boldsymbol{\varepsilon}^j$, where superscripts j indicate that we consider only those observations allocated to cluster j:

$$\boldsymbol{y}^j = \begin{pmatrix} y^j_{1\,1} \\ \vdots \\ y^j_{1\,T^j_1} \\ y^j_{2\,1} \\ \vdots \\ y^j_{m_j\,T^j_m} \end{pmatrix}, \qquad X_j = \begin{pmatrix} \boldsymbol{\iota}_{T^j_1} & 0_{T^j_1, m_j-1} & & \boldsymbol{y}^j_1 \\ 0_{T^j_2,1} & \boldsymbol{\iota}_{T^j_2} & 0_{T^j_2, m_j-2} & \boldsymbol{y}^j_2 \\ \vdots & \ddots & \ldots & \vdots \\ 0_{T^j_m, m_j-1} & & \boldsymbol{\iota}_{T^j_m} & \boldsymbol{y}^j_m \end{pmatrix} \quad \text{and} \quad \boldsymbol{\zeta}_j = \begin{pmatrix} \beta^j_1(1-\alpha_j) \\ \beta^j_2(1-\alpha_j) \\ \vdots \\ \beta^j_{m_j}(1-\alpha_j) \\ \alpha_j \end{pmatrix}$$

where $\boldsymbol{y}^j_i = \left\{y^j_{i\,0}, \ldots, y^j_{i\,T^j_i-1}\right\}$, $\boldsymbol{\iota}_k$ is a k-dimensional vector of ones and $0_{A,B}$ is an $A \times B$ matrix of zeros. So, $\boldsymbol{y}^j \in \mathbb{R}^{\mathcal{T}_j}$, X_j is a matrix of size $\mathcal{T}_j \times (m_j + 1)$ and $\boldsymbol{\zeta}_j \in \mathbb{R}^{m_j+1}$.

Thus, defining $\Omega_j = \text{diag}(\omega^j_{11}, \ldots, \omega^j_{m_j T^j_{m_j}})$, we can write

$$p(\boldsymbol{y}|\boldsymbol{\xi}, \lambda, \boldsymbol{S}, \boldsymbol{\omega}) = \prod_{j=1}^{K} f_N^{\mathcal{T}_j}(\boldsymbol{y}^j | X_j\zeta_j, \lambda^{-1}\Omega_j^{-1})$$

$$\propto \lambda^{\mathcal{T}/2} \exp\left[-\frac{\lambda}{2}\sum_{j=1}^{K}(\boldsymbol{y}^j - X_j\boldsymbol{\zeta}_j)'\Omega_j(\boldsymbol{y}^j - X_j\boldsymbol{\zeta}_j)\right]\prod_{j=1}^{K}\prod_{i=1}^{m_j}\prod_{k=1}^{T^j_{m_j}}\left(\omega^j_{ik}\right)^{1/2}.$$

The expression above is integrable with respect to the prior on λ if and only if at least one of the terms in the sum above is strictly positive. This is the case if and only if the rank of X_j is larger than the dimension of θ_j for at least one $j = 1, \ldots, K$. This is equivalent to the condition in Theorem 2. Under that condition, we can integrate out λ through a Gamma conditional posterior and are left with

$$p(\boldsymbol{y}) \propto \int \left[\sum_{j=1}^{K}(\boldsymbol{y}^j - X_j\boldsymbol{\zeta}_j)'\Omega_j(\boldsymbol{y}^j - X_j\boldsymbol{\zeta}_j)\right]^{-\mathcal{T}/2} \prod_{j=1}^{K}\prod_{i=1}^{m_j}\prod_{k=1}^{T^j_{m_j}}\left(\omega^j_{ik}\right)^{1/2} p(\boldsymbol{\xi}, \boldsymbol{S}, \boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\xi}\, \mathrm{d}\boldsymbol{S}\, \mathrm{d}\boldsymbol{\omega},$$

where $\sum_{j=1}^{K}(\boldsymbol{y}^j - X_j\boldsymbol{\zeta}_j)'\Omega_j(\boldsymbol{y}^j - X_j\boldsymbol{\zeta}_j) > \sum_{j=1}^{K}\boldsymbol{y^j}'\Omega_j\boldsymbol{y}^j - \boldsymbol{y^j}'\Omega_j X_j(X_j'\Omega_j X_j)^{-1}X_j'\Omega_j\boldsymbol{y}^j$, which is strictly positive with probability one under the condition. Thus, we can integrate out $\boldsymbol{\xi}$ with its proper prior. Following the proof of Theorem 2 in Fernández and Steel (2000) $\boldsymbol{\omega}$ can be integrated out to leave a finite result for each assignment that satisfies the condition. Integrating over $\boldsymbol{S}$ with a prior that respects the condition completes the proof.

Appendix B List of OECD countries

1	Australia	11	United Kingdom	21	Mexico
2	Austria	12	Germany	22	Netherlands
3	Belgium	13	Greece	23	Norway
4	Canada	14	Hungary	24	New Zealand
5	Switzerland	15	Ireland	25	Portugal
6	Czech Republic	16	Iceland	26	Slovak Republic
7	Denmark	17	Italy	27	Sweden
8	Spain	18	Japan	28	Turkey
9	Finland	19	Republic of Korea	29	United States
10	France	20	Luxembourg		

References

Alonso-Borrego, C. and Arellano, M. (1999). Symmetrically normalised intrumental variable estimation using panel data, *Journal of Business & Economic Statistics*, **17**, 36–49.

Arellano, M. (2003). *Panel Data Econometrics*, Oxford: University Press.

Baltagi, B. (2001). *Econometric Analysis of Panel Data*, Chichester: Wiley, second ed.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803–821.

Bauwens, L. and Rombouts, J.V.K. (2006). Bayesian clustering of many GARCH models, *Econometric Reviews*, forthcoming.

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

Berger, J.O. and Bernardo, J.M. (1992). Ordered group reference priors with application to the multinomial problem, *Biometrika*, **79**, 25–37.

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

Casella, G., Mengersen, K. L., Robert, C. P. and Titterington, D. M. (2002). Perfect samplers for mixtures of distributions, *J. Roy. Statist. Soc. B*, **64**, 777–790.

Casella, G., Robert, C. P. and Wells, M. T. (2004). Mixture models, latent variables and partitioned important sampling, *Statistical Methodology*, **1**, 1–18.

Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions, *J. Amer. Statist. Assoc.*, **95**, 957–970.

Chen, M. H., Shao, Q. M. and Igrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*, New York: Springer.

Chib, S. (1995). Marginal likelihood from the Gibbs output, *J. Amer. Statist. Assoc.*, **90**, 1313–1321.

DiCiccio, J., Kass, R. E., Raftery, A. E. and Wasserman, L. (1997). Computing Bayes factors by combining simulations and asymptotic approximations, *J. Amer. Statist. Assoc.*, **92**, 903–915.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *J. Roy. Statist. Soc. B*, **56**, 363–375.

Diggle, P. J., Heagerty, P., Liand, K. Y. and Zeger, S. L. (2002). *Analysis of longitudinal data*, Oxford: University Press, second ed.

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

Durlauf, S. N. and Quah, D. T. (1999). The new empirics of economic growth, *Handbook of Macroeconomics*, vol. 1 (J. B. Taylor and M. Woodford, eds.), Amsterdam: Elsevier, pp. 235–308.

Fernández, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness, *J. Amer. Statist. Assoc.*, **93**, 359–371.

Fernández, C. and Steel, M. F. J. (2000). Bayesian regression analysis with scale mixtures of normals, *Econometric Theory*, **16**, 80–101.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *J. Amer. Statist. Assoc.*, **97**, 611–631.

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. and Kaufmann, S. (2004). Model-based clustering of multiple time series, mimeo, Johannes Kepler Universität Linz.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, **1**, 1–19.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711–732.

Heston, A., Summers, R. and Aten, B. (2002). Penn world table version 6.1, http://pwt.econ.upenn.edu/php_site/pwt_index.php. Center for International Comparisons at the University of Pennsylvania (CICUP).

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models, *Econometrica*, **70**, 781–799.

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

Hsiao, C. (2003). *Analysis of Panel Data*, 2nd ed., Cambridge: Cambridge Univ. Press.

Ishwaran, H., James, L. F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions, *J. Amer. Statist. Assoc.*, **96**, 1316–1322.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling, *Statistical Science*, **20**, 50–67.

Juárez, M. A. and Steel, M. F. J. (2006). Non-Gaussian dynamic Bayesian modelling for panel data, Working Paper 06-05, CRiSM, University of Warwick.

Liu, M. C. and Tiao, G. C. (1980). Random coefficient first-order autoregressive models, *Journal of Econometrics*, **13**, 305–325.

Marin, J. M., Mengersen, K. and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions, *Handbook of Statistics*, vol. 25 (D. Dey and C. R. Rao, eds.), Amsterdam: North-Holland, pp. 459–207.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*, New York: Wiley.

Meng, X. and Wong, W. H. (1996). Simulating ratios of normalising constants via a simple identity: A theoretical exploration, *Statistica Sinica*, **6**, 831–860.

Nandram, B. and Petruccelli, J. D. (1997). A Bayesian analysis of autoregressive time series panel data, *Journal of Business and Economic Statistics*, **15**, 328–334.

Nerlove, M. (2002). *Essays in Panel Data Econometrics*, Cambridge: University Press.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap, *J. Roy. Statist. Soc. B*, **56**, 3–48.

Phillips, D. B. and Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions, *Markov chain Monte Carlo in practice* (W. R. Gilks, S. Richardson and S. J. Spiegelhalter, eds.), Boca Raton: Chapman & Hall, pp. 215–240.

Quah, D. T. (1997). Empirics for growth distribution: stratification, polarization and convergence clubs, *Journal of Economic Growth*, **2**, 27–59.

Raftery, A. E. (1996). Hypothesis testing and model selection, *Markov chain Monte Carlo in practice* (W. R. Gilks, S. Richardson and S. J. Spiegelhalter, eds.), Boca Raton: Chapman & Hall, pp. 163–188.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components, *J. Roy. Statist. Soc. B*, **59**, 731–792. (with discussion).

Steele, R. S., Raftery, A. E. and Emond, M. J. (2003). Computing normalizing constants for finite mixture models via incremental mixture important sampling, Tech. Report 436, Department of Statistics, University of Washington.

Stephens, M. (2000a). Bayesian analysis of mixtures with an unknown number of components – An alternative to reversible jump methods, *Annals of Statistics*, **28**, 40–74.

Stephens, M. (2000b). Dealing with label switching in mixture models, *J. Roy. Statist. Soc. B*, **62**, 795–809.

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

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*, Chichester: Wiley.

Verdinelli, I., and Wasserman, L. 1995. Computing Bayes Factors using a generalization of the Savage-Dickey density ratio, *Journal of the American Statistical Association*, 90, 614-618.

Weiss, R. E. (2005). *Modeling longitudinal data*, New York: Springer.