Model-based Clustering of non-Gaussian Panel Data Based on Skew-*t* Distributions

Miguel A. Juárez and Mark F. J. Steel*

University of Warwick

Abstract

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, equilibrium level and the effect of covariates. 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 and possess hierarchical structures that enhance the robustness of the inference. We apply our methodology to GDP growth of European regions and to 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), 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 their estimation, but to also have some cluster-specific parameters. In particular, we will consider model-based clustering, which is based on a formal statistical framework (Banfield and Raftery, 1993; Fraley and Raftery, 2002). In an economic context, Bauwens and Rombouts (2007) propose a method for clustering many GARCH models, while Frühwirth-Schnatter and Kaufmann (2008) discuss a Bayesian clustering

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

method for multiple time series data. From a frequentist perspective, Lin and Ng (2007) propose nonparametric model-based clustering methods for panel data with fixed effects.

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 heavier-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 in Juárez and Steel (2006), yet allowing for clustering and additional covariates, and conduct inference from a Bayesian viewpoint.

As the aims of this paper are rather similar to those of Frühwirth-Schnatter and Kaufmann (2008), 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 (2008) either shrink or pool both (although Frühwirth-Schnatter and Kaufmann, 2006 pool only part of the parameters in a somewhat related model). The prior used in the present paper is carefully elicited and is improper, unlike the conditionally natural-conjugate prior used in Frühwirth-Schnatter and Kaufmann (2008). This implies we need to make sure that the posterior exists (we derive a simple and easily verifiable condition for propriety), but we need to elicit fewer hyperparameters and, more importantly, our prior enjoys a natural invariance (for the parameter we are improper on) with respect to affine transformations of the data, which leads to desirable robustness properties. In addition, 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. The prior structure asks the user for a mean and a variance of the parameters describing long-run equilibrium levels, and allows for comparison (or averaging) of models with different numbers of clusters through Bayes factors. 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/`, along with the data sets used in the applications.

The rest of the paper is organized as follows: Section 2 describes the model and discusses the prior specification and posterior propriety. Numerical methods for conducting inference with this model are briefly discussed in Section 3. Section 4 discusses the analysis of simulated data, which are used to assess the model performance in terms of clustering. Two real data sets are analysed in Section 5 to illustrate the implementation of the model: one comprising per-capita GDP growth data for European regions and the other describes employment growth in Spanish manufacturing firms. Concluding remarks are presented in Section 6.

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. In addition, we observe a vector, $\boldsymbol{x}_{it} = (x_{it}^1, \ldots, x_{it}^p)'$, of covariates. We will focus on the first-order autoregressive model:

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

where the errors $\{\varepsilon_{it}\}$ are independent and identically distributed random quantities with mode at zero and unit precision, α is the parameter governing the dynamic behaviour of the panel and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ is a vector of coefficients related to p explanatory variables in $\boldsymbol{x}_{it}$. We assume that the process is stationary, *i.e.* $|\alpha| < 1$. The parameters β_i are individual effects. Since the error distribution has zero mode, these individual effects can be interpreted as reflecting differences in the long-run modal tendencies for the corresponding individuals. 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 condition throughout on the initial observed values, y_{i0}, and assume that the process started a long time ago.

In order to accommodate skewness while retaining a unique mode at zero, we assume that the error term follows a skew distribution as in Fernández and Steel (1998). Thus, given a unimodal probability density function f which has support on the real line and is symmetric around zero, we consider

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

where $1_{[A]} = 1$ if condition A 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 $P(x > 0 \mid \gamma) = \gamma^2/(1+\gamma^2)$. 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. Of course, we could use other ways of introducing skewness, such as in Jones and Faddy (2003) and Azzalini and Capitanio (2003), but we prefer the approach adopted here because it retains a zero mode and because of its inferential simplicity and the clear interpretation of the extra parameter γ. The latter also facilitates prior elicitation.

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

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

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

The basic model then consists of (1) with ε_{it} distributed according to (3). In this model, we can clearly interpret α as the parameter governing the dynamics of the panel, λ as the observational precision, β_i as the individual long-run level and $\boldsymbol{\mu}$ as a measure of the long-run modal effect of the covariates on the observable. In addition, γ will control the skewness and ν determines the tail behaviour.

As discussed before, 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 applications 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, 2008; Hoogstrate *et al.*, 2000). In order to allow for clustering within the panel, we assume that all units share a common parameter vector, say, $\boldsymbol{\theta}^C$ and each has a cluster-specific set of parameters in $\boldsymbol{\theta}^j$, 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, the coefficients of the covariates or from the equilibrium level of the series. So, extending (1) to allow for different dynamics, covariate effects and levels for each cluster yields

$$y_{it} = \beta_i(1 - \alpha^j) + \alpha^j y_{it-1} + (1 - \alpha^j)\boldsymbol{\mu}^j \boldsymbol{x}_{it} + \lambda^{-\frac{1}{2}}\varepsilon_{it}, \tag{4}$$

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 $\boldsymbol{\theta}^j = \left\{\alpha^j, \beta^j, \boldsymbol{\mu}^j\right\}$. The interpretation of the cluster-specific parameters is as follows: α^j characterizes the autoregressive dynamics, while the long-run average equilibrium level is given by β^j, provided we standardize $\boldsymbol{x}_{it}$ to have mean zero for each unit. Finally, the equilibrium level at each time point will also depend on $\boldsymbol{x}_{it}$ through the coefficients in $\boldsymbol{\mu}^j$.

Note that we have specified common values for the precisions in (4) and (5), as well as the non-normality

parameters of the error distribution in (3). This reflects both our judgement that these are unlikely to be parameters of interest and the relative difficulty of learning from data about the non-normality parameters, especially ν. Making more parameters cluster-specific is perfectly feasible, but we feel the current specification is a good way to focus attention on differences between the clusters that we can easily interpret. Finally, we can also consider alternative partitions of the parameters, where *e.g.* only the dynamics are cluster-specific, *i.e.* the model with $\beta^j = \beta, \boldsymbol{\mu}^j = \boldsymbol{\mu}$, $j = 1, \ldots, K$, leading to $\boldsymbol{\theta}^C = \{\beta, \boldsymbol{\mu}, \gamma, \nu, \lambda, \tau\}$ and $\theta^j = \alpha^j$.

2.1 Prior specification

We specify a product form prior for our clustering model in (4) and (5), combined with (3)

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, M, \tau, \lambda, \gamma, \nu) = \pi(\boldsymbol{\alpha})\, \pi(\boldsymbol{\beta})\, \pi(M)\, \pi(\tau)\, \pi(\lambda)\, \pi(\gamma)\, \pi(\nu)\ , \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha^1, \ldots, \alpha^K)'$, $M = (\boldsymbol{\mu}^1, \boldsymbol{\mu}^2, \ldots, \boldsymbol{\mu}^K)$ and $\boldsymbol{\beta} = (\beta^1, \ldots, \beta^K)'$ denote the cluster-specific parameters, the prior of which will be discussed in the next subsection.

We adopt a standard diffuse (improper) prior for λ, which is invariant with respect to affine transformations. Theorem 1 will provide a simple 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 $2/s_\beta^2$ (this distribution is denoted by $\mathrm{Ga}(2, s_\beta^2/2)$). 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. Full details and further motivation for these choices are provided in Juárez and Steel (2006). Thus, we adopt

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

$$\tau \sim \mathrm{Ga}(2, s_\beta^2/2) \tag{8}$$

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

The degrees of freedom parameter ν is often not that clearly determined by the data, so we consider three different priors. Firstly, we take a $\mathrm{Ga}(2, 1/10)$ prior for ν with mass covering a large range of relevant values (prior mean 20 and variance 200). This prior leads to the probability density function (pdf)

$$\pi_1(\nu) = \frac{\nu}{100}\ \exp[-\nu/10]\ , \tag{10}$$

which has a mode at 10 and allows for all prior moments to exist. We also consider a hierarchical prior by taking

an exponential prior on the scale parameter of a gamma distribution with shape parameter 2, which leads to the pdf

$$\pi_2(\nu) = 2d\frac{\nu}{(\nu+d)^3}\ . \tag{11}$$

This introduces a parameter $d > 0$, which controls the mode $(d/2)$ and the median $((1+\sqrt{2})d)$. The tail is now too heavy to allow for a mean. Finally, in the context of Student-t regression models, Fonseca *et al.* (2006) derive a Jeffreys' prior for ν, which has an "objective" flavour and performs well in terms of frequentist coverage. This prior is proper with pdf

$$\pi_3(\nu) \propto \left[\left(\frac{\nu}{\nu+3}\right)\left(\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right)\right]^{\frac{1}{2}}\ , \tag{12}$$

where $\psi'(\cdot)$ is the trigamma function. This prior has the same right tail behaviour as π_2, not allowing for a mean, but has quite different behaviour close to zero, as it is unbounded as ν tends to zero. The median is always equal to 0.55. Thus, the prior on ν is given by either one of (10), (11) or (12).

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

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

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, $\boldsymbol{\eta}'\boldsymbol{\iota} = 1$ (where $\boldsymbol{\iota}$ denotes a K-dimensional vector of ones) and thus it is natural to specify the Dirichlet prior $\pi(\boldsymbol{\eta}) = \mathrm{Di}(\boldsymbol{\eta} \mid \boldsymbol{e})$, 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

$$\pi(\boldsymbol{S},\boldsymbol{\eta}) = \prod_{i=i}^{m} \pi(S_i \mid \boldsymbol{\eta})\ \pi(\boldsymbol{\eta})\, I(\boldsymbol{S}) \propto \prod_{i=1}^{m} \eta_{S_i} \prod_{j=1}^{K} \eta_j^{-1/2} I(\boldsymbol{S}), \tag{13}$$

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

2.2 The prior on the cluster-specific parameters

An important reason for wanting to put a carefully elicited proper prior on cluster-specific parameters is that we typically want to compute Bayes factors between models with different numbers of components. If the components

have, say, a common β and $\boldsymbol{\mu}$, that would be perfectly feasible with a flat improper prior on $(\beta, \boldsymbol{\mu})$, but in the general case where β^j's and/or $\boldsymbol{\mu}^j$'s are cluster-specific, such Bayes factors would no longer be 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 $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and M truly reflects reasonable prior assumptions, since the Bayes factors will depend crucially on the particular prior used.

Within each cluster, the dynamics parameter α gets a rescaled Beta prior (on $(-1, 1)$), and we make the hyper-parameters of this Beta distribution random, with equal gamma priors. This hierarchical structure of the prior on α leads to more flexibility. In particular, we adopt

$$\pi(\alpha \mid a_\alpha, b_\alpha) = \frac{2^{1-a_\alpha-b_\alpha}}{B(a_\alpha, b_\alpha)} (1+\alpha)^{a_\alpha-1}(1-\alpha)^{b_\alpha-1} \qquad |\alpha| < 1 \tag{14}$$

with

$$a_\alpha \sim \text{Ga}(2, 1/10) \qquad \text{and} \qquad b_\alpha \sim \text{Ga}(2, 1/10)\ . \tag{15}$$

The implied marginal prior on α is roughly bell-shaped with $\text{P}(|\alpha| < 0.5) = 0.65$ and $\text{P}(|\alpha| > 0.9) = 0.03$, in line with reasonable prior beliefs for our (and most) applications. In the context of our clustering model we will use independent and identical priors for the dynamics parameters, thus, defining $\boldsymbol{a}_{\boldsymbol{\alpha}} = (a_{\alpha^1}, \ldots, a_{\alpha^K})'$ and $\boldsymbol{b}_{\boldsymbol{\alpha}} = (b_{\alpha^1}, \ldots, b_{\alpha^K})'$, we have

$$\pi(\boldsymbol{\alpha} \mid \boldsymbol{a}_{\boldsymbol{\alpha}}, \boldsymbol{b}_{\boldsymbol{\alpha}}) = \prod_{j=1}^{K} \pi\left(\alpha^j \mid a_{\alpha^j}, b_{\alpha^j}\right) \tag{16}$$

$$\pi(\boldsymbol{a}_{\boldsymbol{\alpha}}, \boldsymbol{b}_{\boldsymbol{\alpha}}) = \prod_{j=1}^{K} \pi(a_{\alpha^j})\, \pi(b_{\alpha^j}) \tag{17}$$

where each component prior is specified as above. Note that this hierarchical prior structure on $\boldsymbol{\alpha}$ will make the Bayes factors between models with different K less dependent on the prior assumptions.

The long-run equilibrium levels associated with each cluster are often quantities that we possess some prior information about. Within the product form of (6), we propose the following multivariate normal prior for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \sim \text{N}_K\left(\boldsymbol{\beta} \mid m\,\boldsymbol{\iota}, c^2\left[(1-a)\,\text{I} + a\,\boldsymbol{\iota}\,\boldsymbol{\iota}'\right]\right), \tag{18}$$

where $c > 0$ and $-1/(K-1) < a < 1$. The prior in (18) 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 clear when we consider that (18) 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 and μ^j's differ across clusters) might be preferable to the model with cluster-specific β^j's.

We will specify a similar prior structure on the coefficients M. In order to be able to interpret these coefficients, we will standardise each of the p covariates to have mean zero and variance one for each individual unit. Then, we will set the mean of the prior at $\mathbf{0}$ and use a similar covariance structure for the K cluster-specific coefficients of regressor l, grouped in $\boldsymbol{\mu}_l = (\mu_l^1, \ldots, \mu_l^K)'$, leading to

$$\boldsymbol{\mu}_l \sim \text{N}_K \left(\boldsymbol{\mu}_l \mid \mathbf{0}, c_l^2 \left[(1 - a_l) \, \text{I} + a_l \, \boldsymbol{\iota} \boldsymbol{\iota}' \right] \right) , \quad l = 1, \ldots, p, \tag{19}$$

where we choose $c_l > 0$ and we specify a rescaled beta prior for each $a_l \in (-1/(K-1), 1)$.

As an important bonus of such a hierarchical prior structure, the sensitivity of the Bayes factors to the prior assumptions will be much reduced. For example, 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.

The prior in this subsection is similar to that specified in Deschamps (2006) for the regression coefficients in a Markov switching model, although there the same prior is also used for the dynamics parameters (thus precluding stationarity).

2.3 Propriety of the posterior

Note that (7) yields an improper joint prior, so we need to verify the existence of the posterior. 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. We can derive the following necessary and sufficient condition for posterior propriety:

Theorem 1.

Consider the model defined by (4) *and* (5)*, with the error term distributed according to* (3)*, and the prior specific-*

ation as described in Subsections 2.1 and 2.2. The posterior is proper if and only if $\mathcal{T}_j > m_j + p + 1$ *holds for at least one* $j = 1, \ldots, K$.

The condition of Theorem 1 is so weak that any sample with at least one unit with more than $p+2$ observations will always lead to a proper 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 not preclude Bayesian inference. The condition in Theorem 1 will be imposed in the sampler by truncating the prior in (13) through $I(\boldsymbol{S})$.

If we assume a common level $\beta^j = \beta$ and/or a common $\boldsymbol{\mu}^j = \boldsymbol{\mu}$, the existence condition of Theorem 1 will continue to hold, as it is a necessary condition for integrating out the precision λ.

In fact, we can also prove existence under improper flat priors on $\boldsymbol{\beta}$ and M under a slightly stronger condition. More details and proofs of these results can be found in an earlier version of this paper at `http://www.warwick.ac.uk/go/msteel/steel_homepage/techrep/clustnew.pdf`.

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), Marin *et al.* (2005) and, in particular, Frühwirth-Schnatter (2006) provide an exhaustive discussion from the Bayesian perspective.

3.1 Likelihood

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

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

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT_i})$ and the use of (3), (4) and (5) 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

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

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 (see Geweke, 1993), 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 computationally 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 (2000*a*) 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 (2008) 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, Marin *et al.*, 2005 and Frühwirth-Schnatter, 2006), we will not provide much detail here. We have blocked the sampler into separate steps for each of the (vector) parameters, and use Gibbs steps for the precisions λ and τ, the membership probabilities and indicators $\boldsymbol{\eta}$ and $\boldsymbol{S}$ and the auxiliary mixing variables $\{\omega_{it}\}$. The long run equilibrium parameters $\boldsymbol{\beta}$ and M are drawn with random walk Metropolis steps with independent t_3 proposals, with the scale chosen so as to obtain an appropriate acceptance rate. For all other parameters we use Metropolis-Hastings steps from proposals with the mode equal to the previous draw, tuning the free parameter to achieve the desired acceptance rates. We adopt independent rescaled Beta proposals for the components of the dynamics parameter $\boldsymbol{\alpha}$ and the correlations a and $\{a_l\}$. Independent gamma proposal distributions are used for the dynamics hyperparameters $\boldsymbol{a}_\alpha$ and $\boldsymbol{b}_\alpha$, the skewness parameter γ and the degrees of freedom ν.

As pointed out by Celeux *et al.* (2000), Stephens (2000*b*) 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 (2000*b*) use decision-theoretical 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. Using the analytical structure of the posterior

distribution, Frühwirth-Schnatter (2001) proposes a random permutation scheme, while Geweke (2007) introduces the permutation-augmented simulator, a deterministic modification of the usual MCMC sampler. Comprehensive discussions are found in Jasra *et al.* (2005) and Frühwirth-Schnatter (2006).

In our setting, we are interested in differentiating between the components in terms of dynamics, long-run behaviour or covariate effects. 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 $(\boldsymbol{\alpha}, \boldsymbol{\beta}, M)$ before deciding on the labels. This will suggest which of the sets of parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ or M) 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 real-data examples in this paper, this indicates that imposing an identifiability constraint through the dynamics parameter, $\boldsymbol{\alpha}$, is a natural way to identify the labels.

To perform model comparison we use the formal tool of Bayes factors. 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). The Bayes factor between any two models is simply defined as the ratio of the marginal likelihoods. The marginal likelihood is the sampling density integrated out with the prior, and is not immediately obtained from MCMC output. Several ways of approximating the marginal likelihood are available in the literature, see e.g. 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.* (2006) and Ishwaran *et al.* (2001) provide alternative methods specifically designed for mixture models. Here we compute the marginal likelihood based on a particular permutation of the cluster labels, obtained from post-processing the output as explained above. This implies that we need to correct the marginal likelihood by a factor $K!$ (the number of possible permutations), as we have effectively underestimated the prior density by the same factor. As explained in Frühwirth-Schnatter (2004), this leads to a very precise estimate for well-separated clusters. In case the clusters are less well separated, the appropriate correction factor will be in $(1, K!)$ and this procedure will give us an upper bound to the actual marginal likelihood. More precise estimation for such cases can be based on the method proposed in Frühwirth-Schnatter (2004), but this would not change any conclusions in the applications studied here.

In the sequel, we will compute the marginal likelihood using the bridge sampler of Meng and Wong (1996). This method was used and extensively discussed in Frühwirth-Schnatter (2004) in a related context. DiCiccio *et al.* (1997) and Frühwirth-Schnatter (2006) provide comprehensive discussions. Bridge sampling generalizes importance sampling and combines sampling from the posterior distribution with that from an importance function. The marginal likelihood can be approximated by a ratio of sample averages; one from the importance function and another from the posterior. These sample averages both involve a so-called bridge function, which needs to be chosen

subject to an integrability constraint. An important advantage of bridge sampling is its robustness with respect to the relative tail behaviour of the importance function. 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 Beta distributions, with parameters matching the first two moments of the chain, for the dynamics parameter as well as the correlations a in (18) and a_l in (19). The variance of these distributions is then doubled to aid sampling from the entire posterior support. This choice is intended to mimic the posterior closely, while still allowing for easy sampling from the importance density. 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, which is the ratio of the posterior and the prior density values at the restriction (see Verdinelli and Wasserman, 1995). For example, the Bayes factor in favour of a symmetric model over its skewed counterpart will be $p(\gamma = 1|\text{data})/p(\gamma = 1)$. 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 (*e.g.* when the restriction corresponds to a boundary or limit of the parameter space).

4 Simulated Data and Clustering Performance

On the basis of various simulated data sets, we conclude that the numerical methods work well and that the prior described in Subsections 2.1 and 2.2 is reasonable and not overly informative. Inference based on both simulated and real data indicates that there is very little difference between the three different priors for ν. In particular, none of the results reported in the paper was noticeably affected by this prior choice. The only difference we identified was for situations where the data are close to normality, due to the variations in right-hand tail behaviour. In particular, the sampler then mixes less well with the fatter tailed priors $\pi_2(\nu)$ and $\pi_3(\nu)$, as a consequence of the combination of a relatively flat likelihood with a very fat prior tail. Since any value of ν above 50 or so is practically indistinguishable from normality, we are not too interested in minor differences in the far right-hand tail, and will only report results with the prior $\pi_1(\nu)$ in the sequel.

We now use simulated data to highlight the ability of the model to correctly identify the clusters. This will help our understanding of the properties and limitations of the model.

Data were generated from the following baseline model. We take $K = 2$, $m = 80$, $T = 10$ and $y_{i0} = 0$ for all

$i = 1, \ldots, m$. Note that we do not start from "equilibrium" conditions, making it more challenging for the model to adequately cluster in terms of $\boldsymbol{\beta}$. $p = 3$ covariates were generated from a uniform distribution and then standardised. We use the parameter values,

$$\gamma = 0.85\,, \qquad \nu = 5\,, \qquad \lambda = 200\,, \qquad \tau = 2000 \qquad M' = \begin{pmatrix} -0.05 & 0.2 & 0.01 \\ -0.1 & 0.2 & 0.01 \end{pmatrix},$$

which means the clusters are always distinguished by a very small difference in M. Throughout, membership probabilities are $\{0.3, 0.7\}$.

The prior used is as described in Subsections 2.1 and 2.2, with $m = 0$ and $c = 0.3$ in (18), $c_l = 1$ in (19), and a uniform prior on correlations a and a_l.

With these simulated data, we ran MCMC chains of 50,000, discarding the first 10,000 and recording every 10th draw. The computational cost for each chain was about 1.4 hrs of CPU time using our Matlab code on a single-core Xeon processor with a clock-speed of 3 GHz (we could run up to four runs in parallel using a workstation with two processors). Running considerably longer chains led to virtually identical results.

4.1 Effect of the error distribution

An interesting question is whether the non-Gaussian error distribution has a large effect on our estimates. We generate data from the skew-t model described above. In order to further distinguish between the clusters, we take $\boldsymbol{\alpha} = (0.1, 0.3)'$ while we use identical long-run levels $\boldsymbol{\beta} = (0.02, 0.02)'$. Posterior inference on the parameters using the correct skew-t model is well concentrated around the values used to generate the data with a relatively small spread. Neglecting fat tails mostly affects the inference on the observational precision, λ, as can be expected. Neglecting skewness as well by estimating the usual Gaussian model has an additional large effect on the equilibrium levels, which are shifted downward by over 60% of the length of the 95% credible intervals (CI's) (throughout, CI's are taken from the 2.5^{th} to the 97.5^{th} percentiles). Again, this is as expected, since we attempt to capture a negatively skewed distribution ($\gamma < 1$) by a symmetric one, which will underestimate the mode.

To assess the classification performance, we consider the average probability of mis-classification, defined as $\mathrm{P}(S_i \neq j \mid y_i \text{ is generated by cluster } j)$ averaged over all m units. Table 1 summarizes our findings for these data, as well as another dataset generated as above but with $\nu = 2$ (which is not an unusual value in light of our real data applications below). Clearly, not accounting for either of the non-Gaussian aspects of the error distribution worsens the clustering performance, especially for $\nu = 2$.

Table 1. *Synthetic skewed and fat-tailed data. Average mis-classification probabilities for each model using two different data sets (generated with different tail behaviour).*

	$\nu = 5$	$\nu = 2$
Normal	0.16	0.54
skew-Normal	0.15	0.37
skew-t	0.14	0.16

4.2 Effect of the distance between clusters

We examine two ways in which the clusters differ. First we fix the dynamics parameters both at 0.1 and vary the equilibrium locations by taking $\beta^1 = 0$ and varying β^2. We express the normalised difference as $\Delta = \tau^{1/2}(\beta^2 - \beta^1)$. Figure 1(a) indicates the average probability of mis-classification, as defined above. We see that the ability to correctly classify the data increases with the distance between the clusters, as expected. Despite the fact that the data for both clusters have the same starting value, we already have a significant improvement in the clustering ability when the long run levels are one (random effect) standard deviation apart.

Then we fix $\beta^1 = \beta^2 = 0.02$ and $\alpha^1 = 0$ and let α^2 vary. Figure 1(b) shows that we can quite accurately distinguish the clusters for $\alpha^2 < -0.2$ or $\alpha^2 > 0.4$. In this case, we also repeat the experiment with a smaller time dimension, $T = 5$. Of course, this decreases the model performance somewhat, but we still have quite good clustering properties for values of α^2 far away from the extremes. In both samples, it appears to be easier to identify the clusters for a given $|\alpha^2|$ if $\alpha^2 < 0$, as the implied alternating behaviour is quite noticeable.

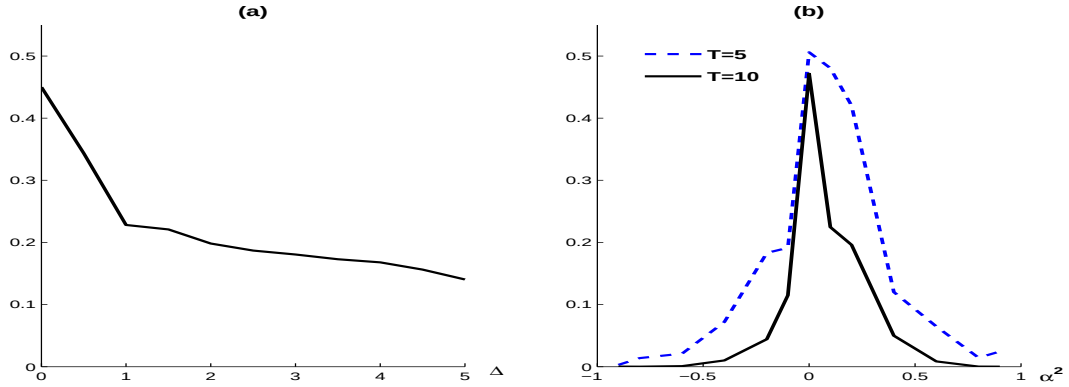

Figure 1. *Average mis-classification probabilities for the simulated data. (a) $\alpha^1 = \alpha^2$ and Δ is the normalised difference between β^1 and β^2. (b) $\alpha^1 = 0, \beta^1 = \beta^2$.*

5 Applications

Two real data sets are analysed in this section. The first contains per-capita GDP of European regions similar to that used in Canova (2004) and Frühwirth-Schnatter and Kaufmann (2008). Here we focus on annual GDP growth. 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 the prior in Subsections 2.1 and 2.2. The induced prior on each long-run growth level β^j will be $\mathrm{N}(0.05, 0.05^2)$ for the GDP data, and $\mathrm{N}(0, 0.05^2)$ for the firms example. In the case of the GDP growth data, we will use a covariate, the (standardized) level of GDP in the previous period, for which the prior of μ^j is $\mathrm{N}(0, 1)$. For the correlation parameters a in (18) and a_l in (19), we will use a uniform prior over $(-1/(K-1), 1)$ in both applications.

MCMC samplers were run for 170,000 iterations, discarding the first 20,000 and then taking every 10th draw, ending up with an effective size of 15,000. This required roughly 13 and 19.5 hrs on a single-core 3 GHz Xeon processor for each application, respectively.

5.1 Per-capita income of European regions

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/regions or convergence clubs have been found; see e.g. Canova (2004) and Quah (1997). Pesaran (2007), using data from the Penn World Tables, found evidence against convergence in levels, but in favour of convergence in growth rates.

Here we concentrate on annual per-capita GDP growth rates from 258 NUTS2 European regions, for the period 1995–2004. The NUTS Classification (Nomenclature des Unités Territoriales Statistiques) was introduced by Eurostat in order to provide a single uniform breakdown of territorial units. NUTS2 units are of intermediate size and roughly corresponds to regional level. These data cover 21 European countries and are collected by Eurostat, based on the European System of National and Regional Accounts (ESA95). We define the growth of region i from time $t-1$ to t as $y_{it} = \log(x_{it}/x_{it-1})$, where x_{it} is the per-capita GDP of region i at time t. Thus, we end up with a balanced panel of $T = 9$ and $m = 258$. As a single covariate we use the lagged level of GDP, x_{it-1} standardized to have mean zero and variance one for each region. This means that $\boldsymbol{\beta}$ now corresponds to the average long-run modal growth levels over time, whereas M can be interpreted in terms of a stabilizing temporal effect. In particular, for our situation with positive growth, negative values for μ^j would imply a decreasing trend of growth over time within cluster j.

We fit the model for $K = 1, 2, 3, 4, 5$. Estimated log Bayes factors (BF) are shown in Table 2, a positive value implying support in favour of the model in the row. For example, the model with $K = 2$ is preferred over the pooled model ($K = 1$) by a Bayes factor of $\exp(35) = 1.58 \times 10^{15}$ and by even more over the models with $K > 2$. Thus, with unitary prior odds (or any prior odds that are likely to be used in practice), the posterior probability of the two-cluster model will be virtually one. Interestingly, the simplest, completely pooled model is clearly preferred to $K = 3, 4$ and 5, but the best model by far is the one with two clusters. Since the model with $K = 5$ was the least preferred, we did not experiment with even larger values of K (which would also not be of practical interest in this

context).

Table 2. *NUTS2 GDP growth 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	-35	532	2037	2295
2		567	2071	2331
3			1504	1764
4				259

Figure 2 shows traces and scatterplots (with a smoothed density representation) of the drawn values for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, M)$ in the chain with two components using the permutation sampler (Frühwirth-Schnatter, 2001), to ensure we adequately explore the entire posterior distribution. Both the traces and the scatterplots illustrates that the dimensions in which the components are most different are the dynamics parameter $\boldsymbol{\alpha}$ and the covariate effect M. Here, we use the labelling convention according to the values of $\boldsymbol{\alpha}$, and impose that $\alpha^1 < \alpha^2$ in post-processing the data. This perfectly implements the separation between the two visually different clusters in the scatterplots. Ordering with respect to μ gives exactly the same results.

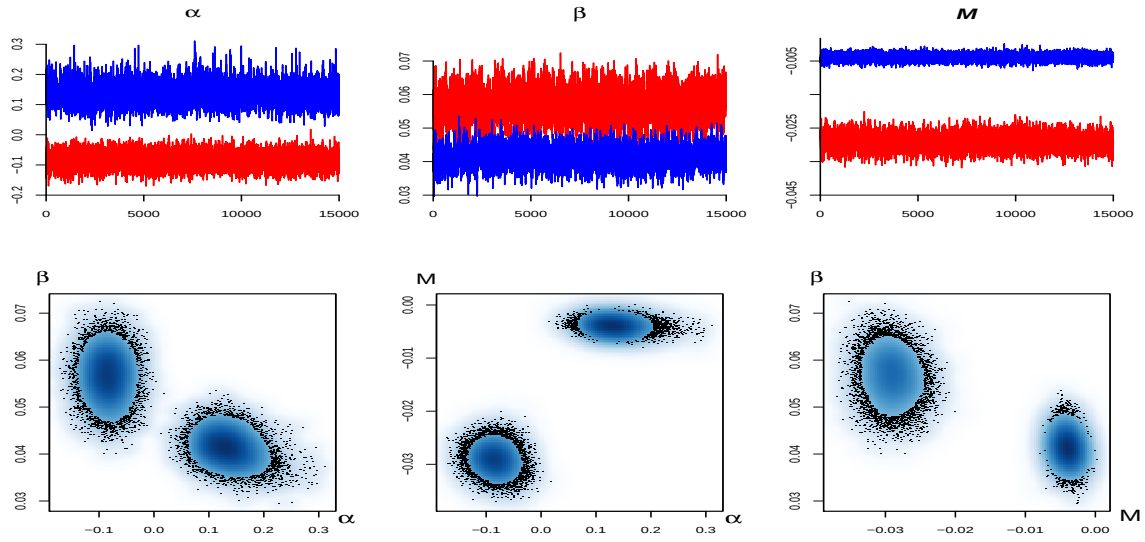

Figure 2. *NUTS2 GDP growth data. Traces and scatterplots of the sampler for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and M, using $K = 2$. Different shades indicate cluster assignment, after post-processing.*

A second important conclusion from Figure 2 is the fact that posterior dependence between the various cluster-specific parameters is quite small. This suggests that the parameterisation used clearly distinguishes between different aspects of the data and that the parameters have a well-defined role.

Figure 3 shows estimated marginal posterior densities for the model-specific parameters of the models with $K = 1, 3, 4, 5$. Throughout, we also plot the prior density in these graphs, indicated by long dashes. Estimation of the common parameters is virtually unaffected by the number of clusters. 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 illusion of precise estimation (note the different scales). Also, it is clear from the inference on $\boldsymbol{\alpha}$ with $K = 3, 4$ and 5 that these models contain more clusters than supported by the data, as there is no clear separation between the clusters with lower α^j. This lack of separation leads to markedly multimodal posteriors for μ^j and can clearly not be solved by choosing a different ordering constraint. 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.

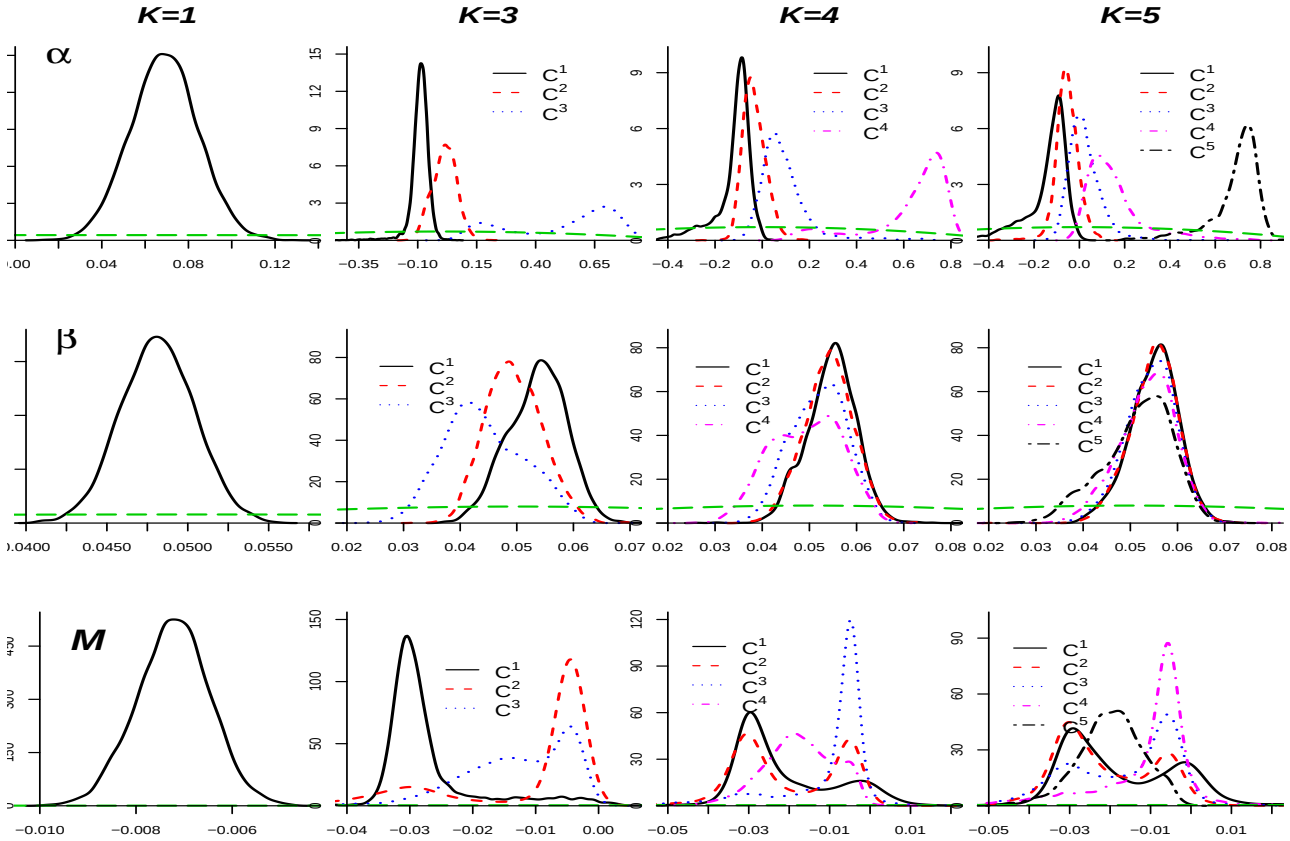

Figure 3. *NUTS2 GDP growth data. Prior (light long dashes) and posterior (as in legend) densities for the cluster-specific parameters, using $K = 1, 3, 4, 5$. Different values of K correspond to different columns. Rows relate to the densities of $\boldsymbol{\alpha}$ (top), $\boldsymbol{\beta}$ (middle) and M (bottom). In the legends C^i indicates cluster i.*

As we saw in Table 2, the two-cluster model is decisively preferred over the others. Posterior results are displayed in Figure 4. Note that the prior on λ is improper and its scaling is, therefore, arbitrary. For this best model with two clusters, convergence is fairly rapid (values of α^j are not large in absolute value), and we have a small club of regions with small negative first order growth autocorrelation (*i.e.* those with a small negative value of α) and a larger subset with small positive first order autocorrelation, as indicated in the top left graph of Figure 4. The posterior mean relative cluster sizes are $\{0.28, 0.72\}$. In addition, Figure 5 shows the individual membership probabilities with the regions ordered in ascending order according to initial GDP level. This illustrates that the first cluster tends to consist of regions with relatively low GDP in 1995. In particular, it groups emerging regions such as all of the Polish regions in the sample and most of the Czech regions, but also includes *e.g.* Inner London and Stockholm with high probability, which experience a similar, somewhat erratic, growth pattern (see Figure 6 in

the sequel).

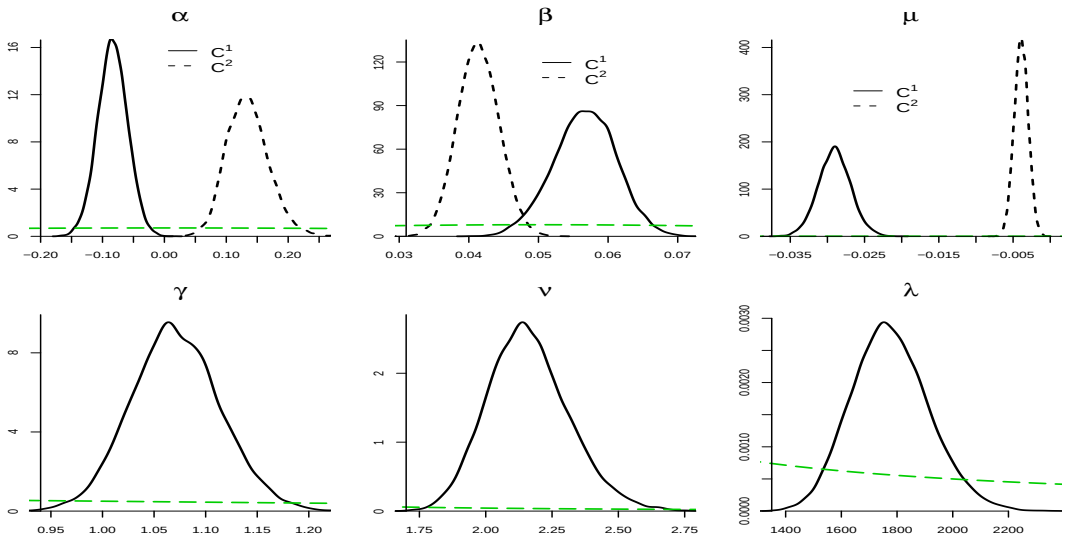

Figure 4. *NUTS2 GDP growth data. Prior (long dashes) and posterior (as in legend) densities for parameters of the model with $K = 2$. For the cluster-specific parameters C^i indicates cluster i.*

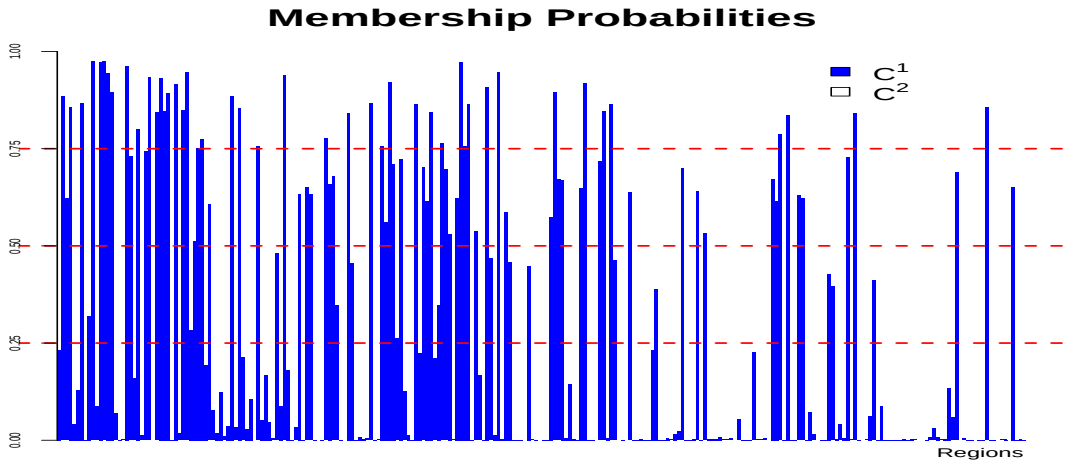

Figure 5. *NUTS2 GDP growth data. Membership probabilities for the model with $K = 2$, with the 258 units (regions) ordered according to initial GDP level. Bars indicate the posterior probability of belonging to cluster 1 for each region.*

The first club has a mean value for α^1 of -0.084, with (-0.130, -0.038) the posterior CI of probability 0.95. For the other club, α^2 has a mean of 0.135, and lies within (0.073, 0.208) with posterior probability of 0.95. Note that the posterior distribution of α for the pooled model ($K = 1$) in Figure 3 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. A summary of the marginal posterior distributions of β^j is shown in Table 3, which suggests that both clubs have different long-run average growth rates. The log Savage-Dickey density ratio in favour of $\beta^1 = \beta^2$ is -17.3, strongly supporting a different average steady-state level. The economies with alternating growth dynamics (first cluster) correspond to a higher median growth rate of around 5.9%, while the second group has a median of about 4.1%. The lower part of Table 3 presents the posterior estimates for the coefficients, $\boldsymbol{\mu}$. For the regions in cluster 1, μ^1 tends to take large negative

values, implying a fairly substantial negative trend of growth over time. For the second cluster, this effect is much smaller. Indeed, looking at Figure 6, which groups average (over regions) observed growth rates for each year, it is clear that growth rates for cluster 1 tend to go down over the sample period, while those for cluster 2 remain almost unaffected. It is also apparent that the time pattern of growth rates for cluster 2 is more stable, with the negative value of α^1 reflected in a more unstable growth pattern for cluster 1. This is in line with cluster 1 grouping mostly emerging economies, which are growing more rapidly in the beginning of the sample period. Interestingly, Figure 6 suggests convergence in growth between the two clusters by the end of the sample period.

Table 3. *NUTS2 GDP growth data. Summary statistics of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ for the skew-t model with $K = 2$.*

	Cluster	median	95% cred.interval
$\boldsymbol{\beta}$ ($\times 10^{-2}$)	1	5.85	(5.12, 6.62)
	2	4.08	(3.54, 4.61)
M ($\times 10^{-3}$)	1	-29.08	(-33.41, -24.84)
	2	-3.90	(-5.80, -2.01)

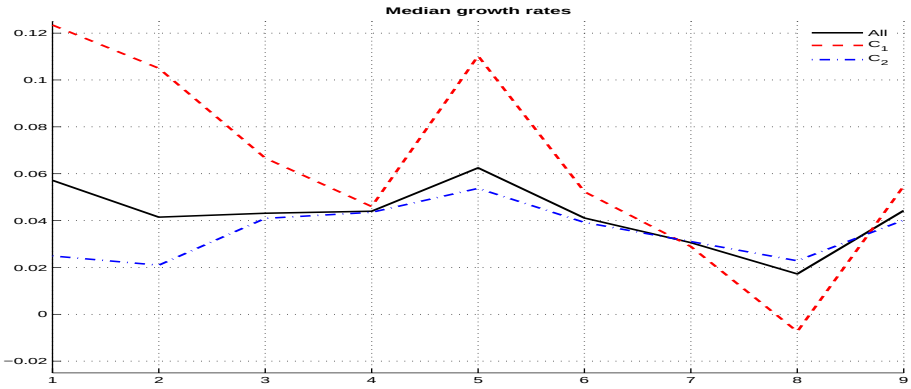

Figure 6. *NUTS2 GDP growth data. Median observed GDP growth over countries with membership according to maximum posterior probability. Solid line: full sample; dashed line: cluster 1; dot-dashed line: cluster 2.*

Figure 4 illustrates that fat tails are a very prominent feature of these data. Posterior inference on ν is quite concentrated on small values in all cases, typically $\nu \in (1.9, 2.5)$ with 0.95 posterior probability. Also, some right skewness is present in this data set. Indeed, $\gamma \in (0.99, 1.16)$ with posterior probability of 0.95. However, the log Savage-Dickey density ratio in favour of $\gamma = 1$ is 1.7, providing mild evidence in favour of the symmetric model.

Given this moderate evidence in favour of $\gamma = 1$, we estimated the symmetric t model with $K = 2$. The estimates of the common parameters and cluster membership probabilities (not shown) were very little affected by imposing symmetry. We did, however, find some small differences for the cluster specific parameters, which we present in Table 4. Clearly, the symmetric model shifts the distribution of the long-run average levels $\boldsymbol{\beta}$ to the right, in order to compensate for the right skewness in the data. Nevertheless, all estimated medians for the parameters of this model are contained in the corresponding 95% CI's of the skew version and viceversa.

Frühwirth-Schnatter and Kaufmann (2008) use a related setting to model the level of per-capita income for a similar dataset consisting of 144 regions for the 1980–1992 period. However, the income data are scaled, for each

Table 4. *NUTS2 GDP growth data. Summary statistics of the cluster-specific parameters, using the symmetric t-model with $K = 2$.*

	Cluster	median	95% cred.interval
α	1	-0.088	(-0.134, -0.041)
	2	0.125	(0.066, 0.184)
β ($\times 10^{-2}$)	1	6.22	(5.61, 6.88)
	2	4.39	(4.00 4.78)
M ($\times 10^{-3}$)	1	-29.62	(-33.62, -25.74)
	2	-4.16	(-5.98, -2.38)

time point, by the European average (see Canova, 2004). This effectively reduces the dynamic behaviour of the individual regions to movement within the European distribution of incomes. In addition, this data set differs from our growth data in that it does not include any Central European regions (or regions in Finland, Sweden and Latvia). Bearing this in mind, we analysed this data set chiefly for comparison with their results. We fitted our model to these level data, using $K = 2$ and no covariates. Like Frühwirth-Schnatter and Kaufmann (2008), we found two well separated clusters, summarized in Table 5, with estimated relative sizes $\boldsymbol{\eta} = (0.13, 0.87)'$. Thus, we have one small converged cluster with α^1 close to zero and a large group with important dynamic behaviour where α^2 is very close to one. Despite the large differences in dynamic behaviour, there is no overwhelming difference in long-term levels (as measured by $\boldsymbol{\beta}$).

Table 5. *NUTS2 income level data. Summary statistics of the cluster-specific parameters using both t models with $K = 2$. Numbers reported are median (95% CI).*

	Clus	Skew-t		t	
α	1	0.014	(-0.020, 0.045)	0.017	(-0.014, 0.049)
	2	0.993	(0.988, 0.998)	0.992	(0.988, 0.996)
β	1	-0.038	(-0.129, 0.057)	-0.086	(-0.184, 0.041)
	2	0.023	(-0.070, 0.112)	-0.102	(-0.190, -0.007)

Clustering regions according to maximum posterior probabilities, yields $m_1 = 15$ and $m_2 = 129$, not unlike the results in Frühwirth-Schnatter and Kaufmann (2008). However, our membership assignments are not quite the same as those in Frühwirth-Schnatter and Kaufmann (2008). We have to keep in mind, though, that their model uses initial income as a covariate for the membership probability.

Our model here differs from the one used in Frühwirth-Schnatter and Kaufmann (2008) in a number of respects, the following of which can have most effect on the results in this application:

i. We do not allow for unit roots, which is of some importance here as the AR(1) process on the levels is close to a unit root for one of the two clusters they find (see their Table 5 and our Table 5). Note that with these data we are modelling levels rather than growth rates.

ii. The other priors are also quite different. They use normals throughout, centred at 0 and with higher variances than ours. As we know, this will affect posterior odds between models.

iii. They do not allow for skewness. Left skewness, however, is a prominent feature of the data, as $\gamma \in (0.815, 0.892)$ with prob. 0.95. As we have seen with both the simulated and the growth data, neglecting skewness can have an important impact on the estimation of the steady-state levels.

iv. They fix the degrees of freedom for the Student-t model at 8. In contrast, we find the tails to be extremely fat, with $(1.1, 1.3)$ the 95% CI for ν, irrespective of the model used (skewed or symmetric). Of course, this will also affect the estimated observational variance and can well influence the clustering (see Subsection 4.1).

In line with the latter point, the evidence in favour of Student-t tails over normal tails is overwhelming, both for symmetric and skewed models. Once we choose a Student model, skewness is strongly preferred by the data. For normal models, however, the log Savage-Dickey density ratio in favour of $\gamma = 1$ is 3.12 (which is in line with the BF obtained from the bridge sampler). It is quite unusual to see the evidence in favour of skewness disappear when we ignore the heavy tails. Thus, if we would not consider (unknown) heavy tails, we would be led dramatically astray in the evidence regarding skewness. Of course, all other models are massively dominated by the skew-t model. Table 5 presents the estimated cluster-specific parameters for both skewed and symmetric t models. The effect of neglecting skewness is similar to that with the previous data: while the dynamics are not much influenced by the inclusion of skewness, long-run levels are affected (but due to the left skewness, now in the other direction). Note that β^2 is shifted much more than β^1.

5.2 Spanish firm employment

The data set is described in the Appendix of Alonso-Borrego and Arellano (1999) and is also used in Arellano (2003, Sec. 6.7). It consists of a balanced panel of 738 manufacturing companies, recorded 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 model described in Section 2, and letting $K = 1$, we obtain 95% CI's of (0.04, 0.08) for α and (-0.0043, 0.0030) for β. Again, inference on parameters common to models with different values of K is virtually unaffected by the choice of the number of clusters.

As shown in Table 6, $K = 1$ is strongly preferred to $K = 2$, $K = 4$ and $K = 5$. However, the model with three clusters performs considerably 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 use larger values of K.

Table 6. *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	823	-9	3074	5122
2		-831	2251	4300
3			3083	5131
4				2049

Scatterplots of the drawn values for (α, β) in the chain with $K = 3$ clearly suggest that identifying the labels through ordering the values of α^j is the natural approach, just like in the previous example.

From the posterior densities in Figure 7, it is apparent that tail behaviour is extremely heavy and very well determined by this (fairly large) data set. These data also clearly present right skewness with (1.05, 1.13) the 95% CI for γ. Both the Savage-Dickey density ratio and bridge sampling indicate massive evidence in favour of the skewed model.

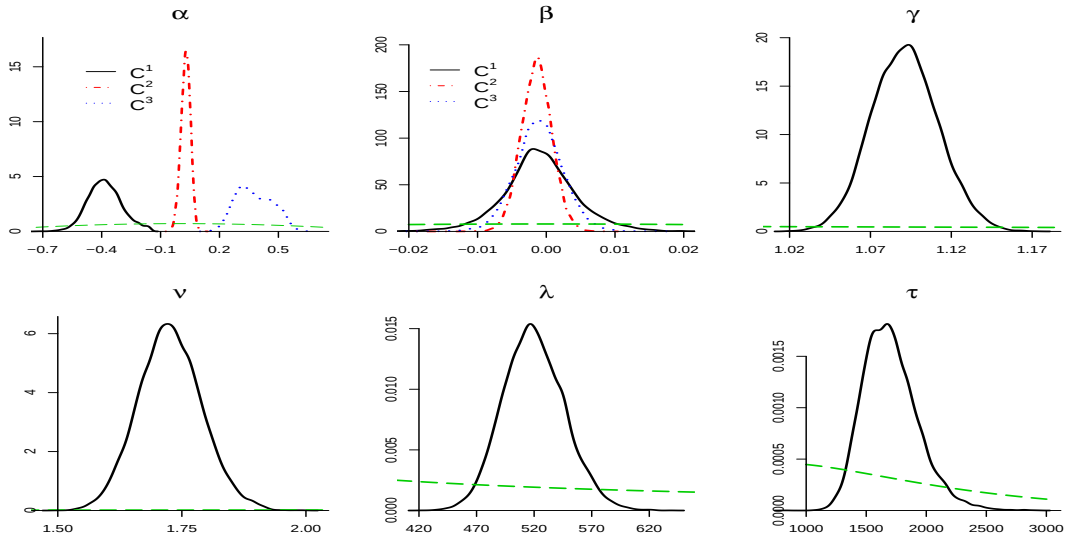

Figure 7. *Spanish firm data. Prior (long dashes) and posterior (as in legend) densities for parameters of the model with $K = 3$. For the cluster-specific parameters C^i indicates cluster i.*

The relative size of each cluster, *i.e.* the average probability of cluster membership, is $\{0.132, 0.651, 0.217\}$. From Figure 7 is is obvious that 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 Figure 7 with the 95% CI of (0.04, 0.08) for α with $K = 1$: rather than gaining strength in the process, opposites are averaged out and the spread of the dynamic behaviour is dramatically underestimated when we use only one cluster.

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 equilibrium values β^j. The posterior medians for β^j with skewness were all within (-0.0011, -0.0008), and these are now all positive, equal to $\{0.0057, 0.0051, 0.0058\}$ with the 95% CI for β^2 entirely on the positive real line. 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. Figure 8 shows that the posterior density of the correlation parameter a, as defined in (18), has a lot of mass close to one and thus strongly supports this model simplification. This is confirmed by the formal log-BF in favour of common β^j's, which is estimated at 13.4. Other parameters are virtually unaffected by this reduction of the model. The common long-run level $\beta \in (-0.005, 0.003)$ with posterior probability 0.95, very much in line with the results for cluster-specific β^j's with the skew-t model (see Figure 7), except that inference is now a bit more precise as a consequence of borrowing strength.

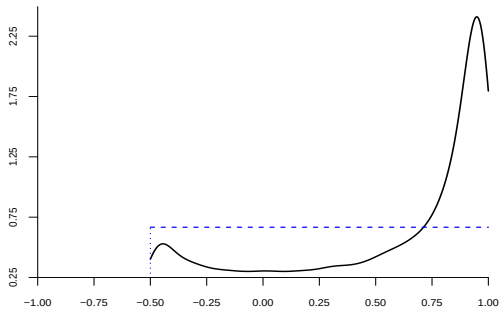

Figure 8. *Spanish firm data. Posterior (solid) and prior (dashed) densities for a in* (18) *using* $K = 3$. *Note that* $a \in (-1/(K-1), 1)$.

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 a 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 9 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 and will thus concentrate the predictive at a value which slightly extends the last observed movement. In the three-cluster model, Firm 433 (Figure 9 (a)) has most mass on the first cluster, which corresponds to large negative values for α (see Figure 7), 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 Figure 7), 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 9.

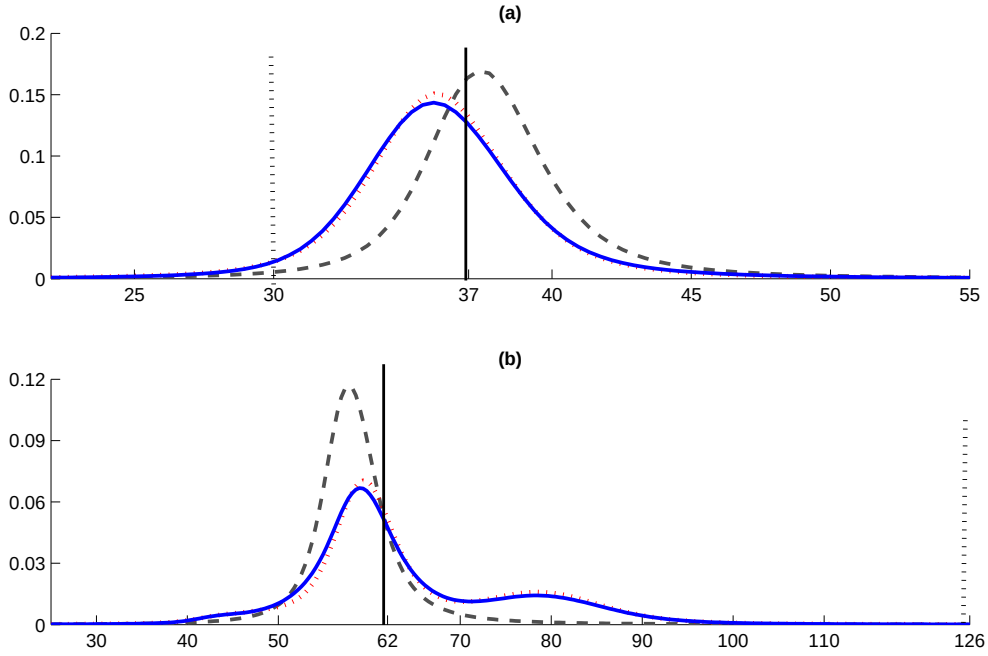

Figure 9. *Spanish firm data. Predictive distribution for 1991 for the employment of firms 433 (a) and 31 (b). Predictives are for $K = 1$ (dashed) and $K = 3$ (solid for skewed model, dotted for symmetric model). Employment numbers for 1989 and 1990 are indicated by dotted and solid vertical lines, respectively.*

6 Conclusion

This paper deals with model-based clustering of longitudinal data, where the clusters can differ in dynamic and long-run equilibrium behaviour and also according to the effect of covariates on the equilibrium levels. 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. The proposed prior structure gives the applied user the opportunity to conduct inference

with these models without spending a lot of effort on prior elicitation. A practically useful and very mild condition for the existence of the posterior distribution is provided. We use a simple scatterplot of the drawn values for the cluster-specific parameters to deal with the labelling problem.

Through simulated data we assess the ability of the model to distinguish between clusters and we find that misspecifying the error distribution (by ignoring either skewness or fat tails) can negatively affect this clustering performance.

We analyse two real (balanced) panel data sets: one on per-capita GDP growth of European regions, with 258 units and $T = 9$, and one concerning employment growth in an even larger sample of 738 manufacturing firms with $T = 7$. Both applications favour clustering, and 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. 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 both applications skewness is important; not just statistically, but also in terms of the conclusions we would draw from the data, as for instance in the firms example, where 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.

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 (as in Frühwirth-Schnatter, 2004) 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.

Other models for dealing with large numbers of time series have been proposed in the literature. For example, dynamic factor models as in Stock and Watson (2002) and Forni *et al.* (2005) are an alternative way to induce dimension reduction, especially used in the context of macroeconomic forecasting.

Acknowledgements: This research was supported by EPSRC under grant number GR/T17908/01. We gratefully acknowledge useful comments by one of the Editors, an Associate Editor, two Referees and Eduardo Ley.

References

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

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

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbations of symmetry with emphasis on a multivariate skew-*t* distribution, *Journal of the Royal Statistical Society* ***B***, **65**, pp. 367–389.

Baltagi, B. (2001). *Econometric Analysis of Panel Data*, Chichester: Wiley, 2nd ed.

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

Bauwens, L. and Rombouts, J. V. K. (2007). Bayesian clustering of many GARCH models, *Econometric Reviews*, **26**, pp. 365–386.

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

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

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

Casella, G.; Mengersen, K. L.; Robert, C. P. and Titterington, D. M. (2002). Perfect samplers for mixtures of distributions, *Journal of the Royal Statistical Society* ***B***, **64**, pp. 777–790.

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

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

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

Deschamps, P. J. (2006). A flexible prior distribution for Markov switching autoregressions with Student-*t* errors, *Journal of Econometrics*, **133**, pp. 153–190.

DiCiccio, J.; Kass, R. E.; Raftery, A. E. and Wasserman, L. (1997). Computing Bayes factors by combining simulations and asymptotic approximations, *Journal of the American Statistical Association*, **92**, pp. 903–915.

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

Diggle, P. J.; Heagerty, P.; Liand, K. Y. and Zeger, S. L. (2002). *Analysis of longitudinal data*, Oxford: University Press, 2nd ed.

Durlauf, S. N. and Johnson, P. A. (1995). Multiple regimes and cross-country growth behaviour, *Journal of Applied Econometrics*, **10**, pp. 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, *Journal of the American Statistical Association*, **93**, pp. 359–371.

Fonseca, T.; Ferreira, M. and Migon, H. (2006). Objective Bayesian analysis for the Student-t regression model, Tech. Report 187, Statistics Department, Federal University of Rio de Janeiro.

Forni, M.; Hallin, M.; Lippi, M. and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting, *Journal of the American Statistical Association*, **100**, pp. 830–840.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, pp. 611–631.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models, *Journal of the American Statistical Association*, **96**, pp. 194–209.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques, *Econometrics Journal*, **7**, pp. 143–167.

Frühwirth-Schnatter, S. (2006). *Finite Mixtures and Markov Switching Models*, New York: Springer.

Frühwirth-Schnatter, S. and Kaufmann, S. (2006). How do changes in monetary policy affect bank lending? An analysis of Austrian bank data, *Journal of Applied Econometrics*, **21**, pp. 275–305.

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series, *Journal of Business and Economic Statistics*, **26**, pp. 78–89.

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

Geweke, J. (1993). Bayesian treatment of the independent Student-*t* linear model, *Journal of Applied Econometrics*, **8**, pp. S19–S40.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works, *Computational Statistics & Data Analysis*, **51**, pp. 3529–3550.

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

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models, *Econometrica*, **70**, pp. 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**, pp. 274–283.

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

Ishwaran, H.; James, L. F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions, *Journal of the American Statistical Association*, **96**, pp. 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**, pp. 50–67.

Jones, M. C. and Faddy, M. J. (2003). A skew extension of the *t*-distribution, with applications, *Journal of the Royal Statistical Society* ***B***, **65**, pp. 159–174.

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.

Lin, C.-C. and Ng, S. (2007). Estimation of panel data models with parameter heterogeneity when group membership is unknown, Tech. report, Department of Economics, Columbia University.

Liu, M. C. and Tiao, G. C. (1980). Random coefficient first-order autoregressive models, *Journal of Econometrics*, **13**, pp. 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. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration, *Statistica Sinica*, **6**, pp. 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**, pp. 328–334.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap, *Journal of the Royal Statistical Society* ***B***, **56**, pp. 3–48.

Pesaran, M. H. (2007). A pair-wise approach to testing for output and growth convergence, *Journal of Econometrics*, **138**, pp. 312–355.

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**, pp. 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, *Journal of the Royal Statistical Society* ***B***, **59**, pp. 731–792, (with discussion).

Steele, R. J.; Raftery, A. E. and Emond, M. J. (2006). Computing normalizing constants for finite mixture models via incremental mixture important sampling, *Journal of Computational and Graphical Statistics*, **15**, pp. 712–734.

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

Stephens, M. (2000*b*). Dealing with label switching in mixture models, *Journal of the Royal Statistical Society* ***B***, **62**, pp. 795–809.

Stock, J. and Watson, M. (2002). Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association*, **97**, pp. 1167–1179.

Temple, J. (1999). The new growth evidence, *Journal of Economic Literature*, **37**, pp. 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**, pp. 614–618.

Weiss, R. E. (2005). *Modeling Longitudinal Data*, New York: Springer.