Stat Comput
DOI 10.1007/s11222-012-9351-7

Univariate Bayesian nonparametric mixture modeling with unimodal kernels

Carlos E. Rodríguez · Stephen G. Walker

Received: 17 February 2012 / Accepted: 24 August 2012

Abstract Within the context of mixture modeling, the normal distribution is typically used as the components distribution. However, if a cluster is skewed or heavy tailed, then the normal distribution will be inefficient and many may be needed to model a single cluster. In this paper, we present an attempt to solve this problem. We define a cluster, in the absence of further information, to be a group of data which can be modeled by a unimodal density function.

Hence, our intention is to use a family of univariate distribution functions, to replace the normal, for which the only constraint is unimodality. With this aim, we devise a new family of nonparametric unimodal distributions, which has large support over the space of univariate unimodal distributions.

The difficult aspect of the Bayesian model is to construct a suitable MCMC algorithm to sample from the correct posterior distribution. The key will be the introduction of strategic latent variables and the use of the Product Space view of Reversible Jump methodology.

Keywords Cluster · Dirichlet process · Mixture model · Slice sampler · Product space · Reversible jump · Label switching

We would like to thank the editor and two anonymous referees for their constructive comments which helped to improve the manuscript. The first author is at the University of Kent with grant support from CONACYT, the Mexican National Council for Science and Technology.

Electronic supplementary material The online version of this article (doi:10.1007/s11222-012-9351-7) contains supplementary material, which is available to authorized users.

C.E. Rodríguez (✉) · S.G. Walker
School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent, UK
e-mail: cerh2@kent.ac.uk

Published online: 28 September 2012

1 Introduction

There are two types of model for Bayesian mixture modeling: One assumes there is a finite integer k, which is the number of mixtures required to model the data, and k is assumed to be either known or unknown, in which case it can take any positive integer value. The other assumes an infinite mixture from the start, and hence there is no explicit parameter modeling the number of groups or clusters.

Algorithms to perform Bayesian analysis of the former mixture model first assumed k to be known and were later extended to cover the k unknown case. For a fixed k, one of the first works using MCMC is given in Diebolt and Robert (1994). To make inference for an unknown number of components, Richardson and Green (1997) used reversible jump MCMC ideas (Green 1995), and Stephens (2000a) used a birth-death process. More recent work done by Nobile and Fearnside (2007) is based on a variation of the model and the sampling technique requires that the parameters of the model can be integrated out analytically. A comprehensive review of mixture models can be found in Frühwirth-Schnatter (2006), and for trans-dimensional MCMC samplers in Green (2003) and Sisson (2005).

For the latter mixture models, when the mixtures are based specifically on the Dirichlet process, it is possible to derive a finite model by integrating out the random distribution function; see, for example, Escobar (1988, 1994) and Escobar and West (1995). More recent approaches utilize the constructive definition of the Dirichlet process Sethuraman (1994), and work directly with the infinite number of mixtures. Appropriate algorithms then need to be constructed so that the correct posterior is sampled by knowing how many of the infinite variables need to be drawn. See Papaspiliopoulos and Roberts (2008) for a retrospective sampler, and Walker (2007), and Kalli et al. (2011) for

slice samplers. The slice samplers introduce latent variables which make the infinite model finite.

A common theme in the two approaches is the use of the normal distribution as the "benchmark" components or kernel distribution. This has been mainly for two reasons; it is a well-known distribution and when used with conjugate priors the resulting MCMC algorithm simplifies. Under the finite mixture set-up, attempts to work with other parametric component distributions are few: see, for example, Stephens (2000a), who used the Student's-t distribution, and Wiper et al. (2001), who used the gamma distribution. Within the Bayesian nonparametric literature, unimodal distributions have been explored; see for example Brunner and Lo (1989) and Lo (1984), but there has been no attempt to incorporate such a nonparametric unimodal distribution as the components distribution in a mixture model.

If a density estimate is needed then the use of the normal distribution is perfectly justified. We can approximate any distribution on the real line using an infinite mixture of normals (Ferguson 1983). However, for the modeling of clusters, it does have some serious issues: if a cluster is skewed or heavy tailed then the normal distribution will be inefficient and many may be needed to model a single cluster. To motivate our proposal we can cite two important works in Bayesian mixture modeling:

Escobar and West (1995), analyzed the galaxy data using mixtures of Dirichlet Processes and obtained unrealistic high posterior values for the number of components.

> The underlying assumption is that each galactic cluster is a normal component. If the distribution of a galactic cluster is skewed or has a very light or heavy tail, then we may use two or more normal components to fit one galactic cluster component.

Richardson and Green (1997), while analyzing three data sets using normal mixtures observed the same problem as Escobar and West (1995)

> In each case, the high overall number of components can be related in part to the skewness of the data, two or three normals being sometimes needed to fit one skewed component.

Hence, to model a single cluster, two or possibly more normals are needed, so the number of mixtures does not coincide with the number of clusters; meaning that the number of normal mixtures has no real interpretation. Furthermore, we note the unsatisfactory consequence that the number of modes of the density estimate is typically less than the estimate of the number of mixtures, see Escobar and West (1995) pp. 583.

While clustering is an important part of mixture models, usually there is no corresponding definition of what a cluster actually is. Presumably, when a mixture of normals is used the assumption is that each one of the hidden groups is well represented by a normal distribution. Which is unrealistic in practice.

Our aim is to ensure the number of clusters within the data coincides with the estimate of the number of components. Our plan is to use a components distribution for which the only constraint is unimodality. We employ this in a finite mixture model context, where k is modeled explicitly. With this objective, we introduce a new family of nonparametric unimodal distributions, which has large support over the space of unimodal distributions. Hence, given k, the model is a k-mixture of unimodal densities, each of which is modeled nonparametrically. In short, therefore, we are defining a cluster as a set of observations which can be modeled by a unimodal density. In the absence of further information beyond the observations, this is the most reasonable working assumption for a cluster. The idea being that a multimodal density would reasonably be assumed to contain more than one cluster.

Much effort needs to be dedicated to the MCMC sampler. In fact, we derive a hybrid MCMC strategy. There are three key points here. First, the introduction of some latent allocation variables (as with every mixture model). Second, the slice sampling ideas of Kalli et al. (2011) to truncate the number of variables to a finite number. Third, following Godsill (2001), a trans-dimensional step is devised, for an entire stochastic process, writing a large joint distribution on the product space of candidate models and performing a Metropolis-Hastings in the usual way.

To make inference for individual components and cluster analysis the well-known problem of label switching must be addressed. We used Rodríguez and Walker (2012) post-processing algorithm, to "undo" the label switching.

The paper is structured as follows. In Sect. 2, we present the model that we use and describe some of its characteristics. The hybrid MCMC strategy is described in Sect. 3. First, we deal with the case for a fixed number of components and then we move to the trans-dimensional case. In Sect. 4 we illustrate and compare our methodology against the classic mixture of normals using simulated and real data sets. We also discuss and use a new approach to solve the label switching problem. We conclude with a discussion and an outline of future work.

2 The model

The model for the data will be a mixture model for which k is modeled explicitly:

$$f_k\left(y|w^{(k)},\lambda^{(k)},\mu^{(k)},G^{(k)}\right)=\sum_{j=1}^{k} w_j\, f(y|\mu_j,\lambda_j,G_j), \quad (1)$$

with $w^{(k)} = \{w_j\}_{j=1}^k$, $\lambda^{(k)} = \{\lambda_j\}_{j=1}^k$, $\mu^{(k)} = \{\mu_j\}_{j=1}^k$ and $G^{(k)} = \{G_j\}_{j=1}^k$. The weights, $w^{(k)}$, are non-negative and sum to one; and for each j, $f(y|\mu_j, \lambda_j, G_j)$ is a univariate unimodal distribution which can be fully characterized by a distribution function G_j, and also with the parameter λ_j, which determines the asymmetry of f, and μ_j, which determines the location of f. All the other aspects of f can be determined by the moments of the distribution G_j.

To start, we describe the family of univariate unimodal density functions which will serve as the components of the mixture model (1).

2.1 Unimodal kernels

The widely used mixture of Dirichlet process model, Lo (1984), consists in mixing a kernel $k(y|\theta)$ with respect to a random distribution function G, to define a random density;

$$f(y|G) = \int_{\Phi} k(y|\theta)G(d\theta).$$

Provided that the kernel $k(y|\theta)$ is a density for each θ, then $f(y|G)$ will be a density. The prior for G is typically taken as the Dirichlet process (Ferguson 1973): $G \sim \mathscr{P}$, where $\mathscr{P}$ has scale parameter $c > 0$ and mean distribution F_0 in $\mathbb{R}$. The Dirichlet process is almost surely a discrete measure (Blackwell 1973), and the representation of Sethuraman (1994) is useful here: $G(\cdot) = \sum_{s=1}^{\infty} w_s \delta_{\theta_s}(\cdot)$, where the (θ_s) are random variables independent and identically distributed (i.i.d.) from F_0 and

$$w_1 = v_1, \qquad w_s = v_s \prod_{l=1}^{s-1}(1 - v_l), \quad \text{for } s \geq 2,$$

with the (v_s) being i.i.d. beta$(1, c)$, for some $c > 0$. Thus we can write

$$f(y|G) = \sum_{s=1}^{\infty} w_s k(y|\theta_s). \tag{2}$$

If the kernel $k(y|\mu, \sigma^2) = N(y|\mu, \sigma^2)$ is taken, then (2) defines a multi-modal distribution, which is not what we want. On the other hand, if

$$k(y|\theta, \mu) = \mathrm{U}(y|\mu - \theta, \mu + \theta), \quad \text{with } \theta \in \mathbb{R}^+,$$

then it is well-known that scale mixtures of uniforms coincide with the class of unimodal and symmetric distributions (Feller 1971):

$$\begin{aligned} f(y|\mu, G) &= \int_0^{\infty} \mathrm{U}(y|\mu - \theta, \mu + \theta)G(d\theta) \\ &= \sum_{s=1}^{\infty} w_s \,\mathrm{U}(y|\mu - \theta_s, \mu + \theta_s) \end{aligned} \tag{3}$$

where $\mathrm{U}(y|a, b)$ is the uniform density on (a, b):

$$\mathrm{U}(y|a, b) = \frac{1}{b-a} \mathbb{1}_{(a,b)}^{(y)} \quad \text{with } b, a \in \mathbb{R} \quad \text{and} \quad b > a.$$

In this case $f(y|\mu, G)$ is a symmetric random density with mode μ, and where its variance and kurtosis are determined by G.

Brunner and Lo (1989) and Quintana et al. (2009), among others, have worked with (3). But in our case, we want to extend (3) to include asymmetry. A possible option is to use the proposal of Kottas and Gelfand (2001) where they model asymmetry using two independent Dirichlet processes G_1 and G_2, i.e.

$$\begin{aligned} &f(y|\mu, G_1, G_2) \\ &= \frac{1}{2}\left\{\int \frac{1}{\theta} \mathbb{1}_{(\mu-\theta,\mu)}^{(y)} G_1(d\theta) + \int \frac{1}{\theta} \mathbb{1}_{[\mu,\mu+\theta)}^{(y)} G_2(d\theta)\right\}. \end{aligned}$$

In this case, the tails of the distribution are being modeled independently of each other, which could lead to a large discontinuity at the mode. Also, if it is close to symmetric, then the model is inefficient. Instead, following the ideas of Fernandez and Steel (1998), we can return to (3) and incorporate an asymmetry parameter $\lambda \in \mathbb{R}$ via

$$\begin{aligned} f(y|\lambda, \mu, G) &= \int_{\mathbb{R}^+} \mathrm{U}\big(y|\mu - \theta e^{-\lambda}, \mu + \theta e^{\lambda}\big)G(d\theta) \\ &= \sum_{s=1}^{\infty} w_s \,\mathrm{U}\big(y|\mu - \theta_s e^{-\lambda}, \mu + \theta_s e^{\lambda}\big). \end{aligned} \tag{4}$$

Then (4) defines a random unimodal density determined by (λ, μ, G); μ is the location parameter; λ the asymmetry parameter and G a distribution function. Characteristics such as variance, kurtosis, tails and higher moments are determined by G.

We do not claim that the support of (4) includes all the unimodal densities on the real line. But it certainly covers all symmetric unimodal distributions and a large class, sufficiently large in our estimation, of asymmetric distributions.

Then we will use (4) as the components distribution for the mixture model (1).

2.1.1 Prior predictive

It is important to understand what kind of shapes the unimodal distribution, (4), can achieve and how the parameters λ and μ and the moments of G influence it. With this aim, setting $F_0(\theta|\alpha, \beta)$ as a gamma distribution, written as $\Gamma(\theta|\alpha, \beta)$, (parameterized such that $\mathbb{E}(\theta) = \alpha/\beta$), the prior predictive or prior guess can be calculated by integrating out G from (4) to yield

$$\mathbb{E}\big(f(y|\lambda, \mu, G)\big)$$

$$
\begin{aligned}
&= \int_{\Omega} f(y|\lambda,\mu,G)\mathscr{P}(dG) \\
&= \int_{\Omega}\left[\int_{\mathbb{R}^+} k(y|\theta,\lambda,\mu)G(d\theta)\right]\mathscr{P}(dG) \\
&= \int_{\mathbb{R}^+} k(y|\theta,\lambda,\mu)\left[\int_{\Omega} G(d\theta)\mathscr{P}(dG)\right] \\
&= \frac{\operatorname{sech}(\lambda)}{2}\int_{\mathbb{R}^+}\frac{1}{\theta}\mathbb{1}_{\left(\mu-\theta e^{-\lambda},\mu+\theta e^{\lambda}\right)}^{(y)} F_0(d\theta|\alpha,\beta) \\
&= \frac{\beta\operatorname{sech}(\lambda)}{2(\alpha-1)}\big(1-F_0\big(a(y)|\alpha-1,\beta\big)\big) \qquad (5)
\end{aligned}
$$

with $a(y) = \max\{(\mu-y)e^{\lambda}, (y-\mu)e^{-\lambda}\}$.

Note that with the choice of F_0 for $k(y|\theta,\lambda,\mu)$ to be a valid kernel it is required that $\alpha > 1$. For (5) to be a differentiable (smooth) function, $\alpha > 2$ is needed, and if $1 < \alpha \leq 2$ we will have a continuous function that is not differentiable at $y = \mu$. See Appendix A for details.

Hence, (5) is a four parameter density. But the important point here is to understand how λ, μ, α and β influence the prior predictive (5) and ultimately (4). To clarify this point some graphics have been displayed in Fig. 1. To have a measure of comparison, a plot of the standard normal distribution has been included.

We have that μ deals with the location; λ the skewness and α and β the variance and kurtosis. This can be seen for α in graphics: (a), (b) and (d) and for λ in graphics: (c), (d), (e) and (f). From these graphics is clear that α determines the degree of kurtosis. Note in graphs (e) and (f) how large values of lambda ($|\lambda| > 1$) can also impact the variance. That β influences variance can be seen from graphs (b) and (c). Finally μ only influences the location; see graphs (d) and (f).

The best approximation to the normal distribution with the prior predictive is shown in graphic (c), a zoom to the right tails, between 3 and 6, shows that the tails of the prior guess are slightly heavier than those of the standard normal distribution.

On studying the prior predictive (5) we see that contrary to the realizations of $f(\lambda,\mu,G)$ it is a continuous density. We observed how the random distribution G along with the parameters λ and μ influence the prior guess. This information is important because one of the aims is to approximate the posterior predictive $\mathbb{E}(f(y|\lambda,\mu,G)|\mathbf{y})$ once a sample from a population $\mathbf{y} = y_1,\ldots,y_n$ has been observed. Thus the study of the prior predictive gives us an insight on the representations or shapes that the posterior predictive can achieve and how (λ,μ,G) (given the observations) can influence it. Further, since we will be working with a new distribution, this knowledge can be helpful to set the values of the unspecified constants for the priors of the model.

Later on in the paper we will evaluate the performance of model (4) by fitting data sets from different unimodal distributions.

2.2 Mixtures of unimodal kernels

We now write the model (1) as

$$
\begin{aligned}
&f_k\big(y|w^{(k)},\lambda^{(k)},\mu^{(k)},G^{(k)}\big) \\
&\quad = \sum_{j=1}^{k} w_j \sum_{s=1}^{\infty} w_{js}\,\mathrm{U}\big(y|\mu_j-\theta_{js}e^{-\lambda_j},\mu_j+\theta_{js}e^{\lambda_j}\big). \qquad (6)
\end{aligned}
$$

In (6) there are no assumptions about the shape of the components, the only assumption is that of unimodality. Therefore, k means something explicit here: the number of clusters modeled by a unimodal density.

2.2.1 *Allocation variables*

For the mixture of unimodal distributions we have two types of weights and therefore we will need two sets of allocation variables. Thus a joint (z_i,d_i) of latent allocation variables needs to be defined. Then $(z_i = j, d_i = s)$ will indicate that the observation y_i has been drawn from the component j of the finite mixture ($j \in \{1,\ldots,k\}$) and from the s component of the infinite mixture for the unimodal distribution. Note that, *a priori*, (z_i,d_i) are drawn independently with distributions

$$p(z_i = j, d_i = s) = w_j w_{js}$$

for $j = 1,\ldots,k$ and $s = 1,2,\ldots$.

Hence, given the values of (z_i,d_i), the observations are sampled from their respective components;

$$
\begin{aligned}
&f\big(y_i|z_i,d_i,\lambda^{(k)},\mu^{(k)},G^{(k)}\big) \\
&\quad = \mathrm{U}\big(y_i|\mu_{z_i}-\theta_{z_i d_i}e^{-\lambda_{z_i}},\mu_{z_i}+\theta_{z_i d_i}e^{\lambda_{z_i}}\big).
\end{aligned}
$$

Summing out the (z_i,d_i) we get back to (6). Note that the likelihood is now given by

$$\prod_{i=1}^{n} f\big(y_i|z_i,d_i,\lambda^{(k)},\mu^{(k)},G^{(k)}\big). \qquad (7)$$

To sample from the full conditional of the allocation variables it will be necessary to work with

$$
\begin{aligned}
&p(z_i = j, d_i = s|\cdots) \\
&\quad \propto w_j w_{js}\,\mathrm{U}\big(y|\mu_j-\theta_{js}e^{-\lambda_j},\mu_j+\theta_{js}e^{\lambda_j}\big) \qquad (8)
\end{aligned}
$$

for $j = 1,\ldots,k$ and $s = 1,2,\ldots$. This can not be sampled directly due to the infinite choice of s, but in the next section a way to solve this problem is described.

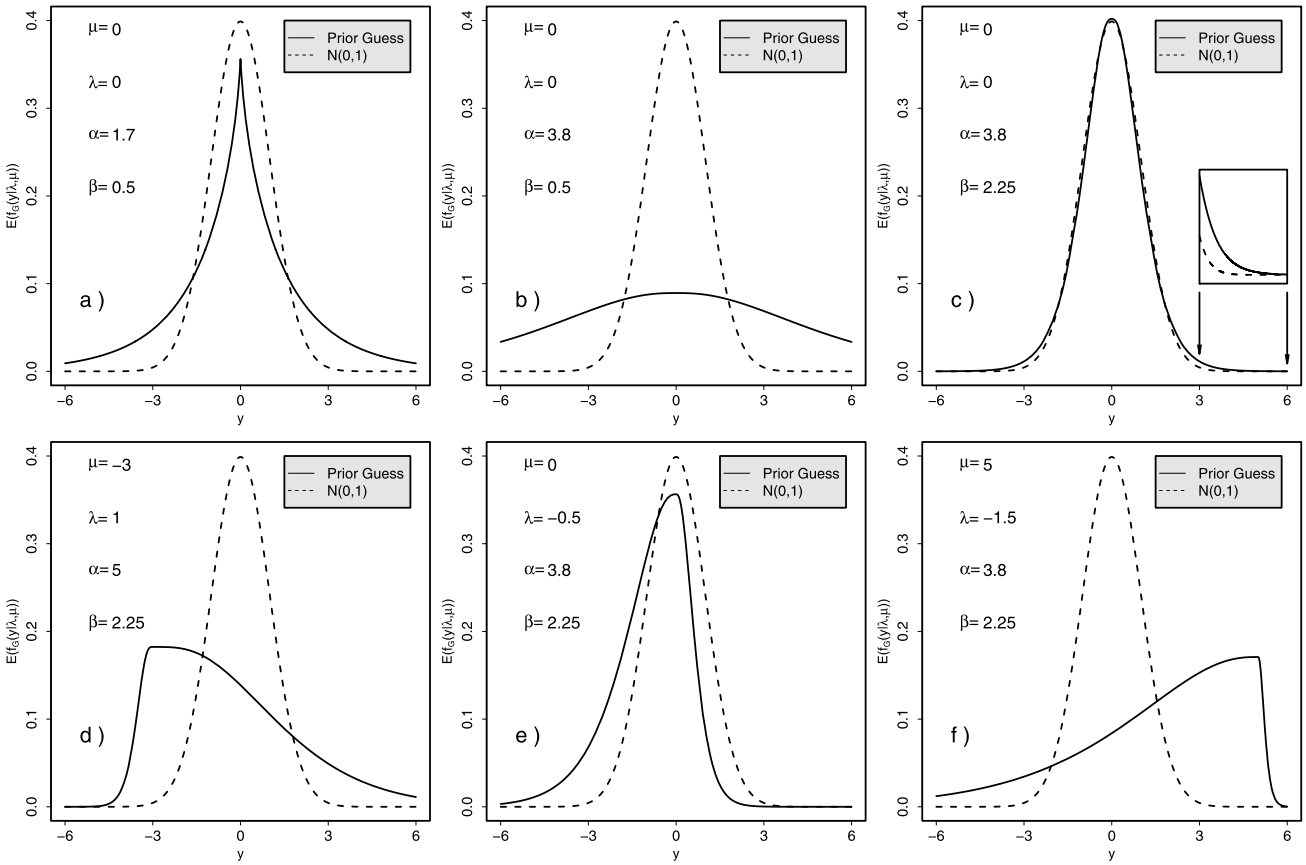

Fig. 1 Influence of λ, μ, α and β on the prior guess

3 MCMC: hybrid strategy

In this section we describe how to sample from the posterior distribution of model (6). First, in Sect. 3.1, the case for a known number of components is outlined and then the case for an unknown number of components (i.e. $k \in \{1, 2, \ldots\}$) is described in Sect. 3.2.

3.1 Known number of components

To tackle the problem for an unknown number of components with model (6) is difficult. It is better to describe how to sample the model with k fixed and then we can proceed to add the extension of a moving k. This is a standard procedure. For the normal case, Diebolt and Robert (1994) first devised the case for a fixed k, now a very well-known Gibbs sampler, and then later other authors used this as a cornerstone to derive algorithms for normal distributions with an unknown number of components; see for example Richardson and Green (1997) and Stephens (2000a).

3.1.1 Slice sampler for infinite mixtures

Walker (2007), and later Kalli et al. (2011), have devised an easy way to sample from (8). The idea is to add for every d_i a slice variable u_i such that

$$
\begin{aligned}
&p(z_i, d_i, u_i | \cdots) \\
&\quad \propto w_{z_i} w_{z_i d_i} \, \mathrm{U}(u_i | 0, \xi_{d_i}) \\
&\qquad \times \mathrm{U}\left(y_i | \mu_{z_i} - \theta_{z_i d_i} e^{-\lambda_{z_i}}, \mu_{z_i} + \theta_{z_i d_i} e^{\lambda_{z_i}}\right) \qquad (9)
\end{aligned}
$$

with ξ_l any positive and decreasing function in l. It is clear that integrating out the (u_i) we get back to the original distribution. The purpose of the (u_i) is to force each d_i to be from a finite set. This can be seen for example by setting $\xi_d = e^{-d}$, which is the form that we use in this paper.

Then, from (9),

$$
\begin{aligned}
u_i \sim \mathrm{U}\left(u_i | 0, e^{-d_i}\right) &\Rightarrow u_i < e^{-d_i} \\
&\Leftrightarrow d_i < -\log(u_i),
\end{aligned}
$$

so if $N_i = \lfloor -\log(u_i) \rfloor$ (where $\lfloor a \rfloor$ is the closest integer to a less than or equal to a) it follows that for $i = 1, 2, \ldots, n$

$$
d_i \leq N_i \leq -\log(u_i) \Rightarrow d_i \in \{1, 2, \ldots, N_i\}.
$$

Hence, to sample from $p(z_i, d_i, u_i|\cdots)$ we have the (u_i) as uniform, and

$$p(z_i = j, d_i = s|u_i, \cdots) \propto w_j w_{js} e^s \times \mathrm{U}\big(y_i|\mu_j - \theta_{js} e^{-\lambda_j}, \mu_j + \theta_{js} e^{\lambda_j}\big)$$

with $j \in \{1, \ldots, k\}$ and $s \in \{1, \ldots, N_i\}$. If we define

$$N = \max_{i=1,\ldots,n} \{N_i\} \Rightarrow \forall i,\ d_i \in \{1, \ldots, N\}. \tag{10}$$

The variables that depend on (z_i, d_i) will be matrices:

$$\big\{\{w_{js}\}_{s=1}^N\big\}_{j=1}^k \quad \text{and} \quad \big\{\{\theta_{js}\}_{s=1}^N\big\}_{j=1}^k.$$

With the inclusion of these latent variables the full posterior distribution will be proportional to

$$\prod_{i=1}^n w_{z_i} w_{z_i d_i} \,\mathrm{U}(u_i|0, \xi_{d_i}) \times \mathrm{U}\big(y_i|\mu_{z_i} - \theta_{z_i d_i} e^{-\lambda_{z_i}}, \mu_{z_i} + \theta_{z_i d_i} e^{\lambda}_{z_i}\big).$$

3.1.2 Model priors

The prior on the weights of the finite mixture will be a Dirichlet distribution, $\mathrm{Dir}(w_1, \ldots, w_k|\delta, \ldots, \delta)$. For the location parameters

$$p(\mu_1, \ldots, \mu_k) \propto k! \prod_{j=1}^k N\big(\mu_j|\mu_0, \sigma_0^2\big)\, \mathbb{1}\{\mu_1 < \cdots < \mu_k\} \tag{11}$$

the order statistics of k normal distributions.

We remark that (11) is not to provide an identifiability constraint and break the symmetry of the likelihood of (6). It has been shown (Stephens 2000b) that these constraints do not solve the label switching problem. For the fixed k case independent normals can be used, but as we are building the trans-dimensional case the purpose of this prior is to impose the order needed on the location parameters to construct the invertible transformation as in Richardson and Green (1997). Their transformation takes μ_j with $\mu_{j-1} < \mu_j < \mu_{j+2}$ and splits it into μ_{j_1} and μ_{j_2} such that $\mu_{j-1} < \mu_{j_1} < \mu_{j_2} < \mu_{j+2}$. For the inverse transformation we need to select μ_{j_1} and μ_{j_2} and combine them into μ_j, preserving the same order as in the split.

The prior for the skewness parameters are independent and uniform: $\mathrm{U}(\lambda_j| -\epsilon, \epsilon)$ for $j = 1, \ldots, k$. To model the asymmetry of each cluster in a flexible way we included a hierarchical prior for ϵ, $p(\epsilon) = \mathrm{U}(\epsilon|0, \rho)$ for some $\rho > 0$.

In the case of the weights of the infinite mixture we will use the stick-breaking construction, thus

$$p(v_{js}) = \mathrm{beta}(v_{js}|1, c) \text{ independent for all } j \text{ and } s.$$

We centered the Dirichlet process on gamma distributions, so

$$p(\theta_{js}) = \Gamma(\theta_{js}|\alpha, \beta_j) \text{ independent for all } j \text{ and } s.$$

In order not be too restrictive with the variance of each unimodal component a hierarchical prior for each β_j was included; $p(\beta_j) = \Gamma(\beta_j|a, b)$. Hence, we fix the values $(\delta, \mu_0, \sigma_0, \rho, c, \alpha, a, b)$.

The smoothing parameter c, from the stick-breaking representation of the Dirichlet process, influences the size of the truncation point N, of the infinite mixture, see (10). As a result, for small values of c small values of N are obtained, thus leading to less smooth unimodal distributions. On the other hand, larger values of c support larger values for N, thus producing smoother distributions. Initially, for the fixed k case, we followed Escobar and West (1995) and imposed a gamma prior over c but the results were very similar to those where c was fixed at the mean of the gamma distribution. So, for simplicity, we omitted this hierarchical level and set c directly.

To derive the full conditional distributions used to construct the Gibbs sampler, when k is known, is straightforward. For completeness, these are described in Appendix C.

3.2 Unknown number of components

To develop the case for an unknown number of components, we will use the product space model discussed by Godsill (2001). Godsill extended the Carlin and Chib (1995) ideas by introducing a general product space model that comprises many trans-dimensional algorithms, including the reversible jump methodology of Green (1995); see Green (2003). First, a description of the product space model is given and then the acceptance probability for the Metropolis-Hastings step in the product space, is deduced.

3.2.1 Product space model

We suppose that our observations have been generated by a model within a countable collection of candidate models

$$\{\mathcal{M}_k, k \in \mathcal{K}\}$$

where $\mathcal{K}$ is a set of candidate model indices. Model $\mathcal{M}_k$ has a vector or matrix $\phi^{(k)}$ of unknown parameters. Each parameter $\phi^{(k)}$ has support $\Phi^{(k)}$ and for models with different indices the dimension of their parameters may vary. Then, our goal would be to choose the best model for the data within all the possible models. The solution given within the Bayesian setting is to calculate $p(\phi^{(k)}, k|\mathbf{y})$ where $\mathbf{y} = \{y_i\}_{i=1}^n$. To this end one could follow the ideas described in Green (1995). In this setting, for a given index k;

$(k, \phi^{(k)}) \in (\{k\} \times \Phi^{(k)})$ and in general, for a moving model index k,

$$\left(k, \phi^{(k)}\right) \in \bigcup_{k \in \mathcal{K}} \left(\{k\} \times \Phi^{(k)}\right). \tag{12}$$

Here the idea is to devise a Markov chain with invariant distribution $p(\phi^{(k)}, k|\mathbf{y})$, that traverses the space (12), generating proposals $q(k', \phi^{(k')}|k, \phi^{(k)})$ to jump through sub-spaces of different dimensions, which are then accepted with probability $\alpha_{k,k'}$. To ensure convergence of the chain, to the correct invariant distribution, the proposals must satisfy the detailed balance condition. Once the overall detailed balance is written, the acceptance probability $\alpha_{k,k'}$ for the reversible jump methodology is worked out, and this is a complicated procedure. Hence, instead of considering the finite dimensional parameters $\phi^{(k)}$, we consider

$$\phi = \left(\phi^{(1)}, \phi^{(2)}, \phi^{(3)}, \ldots\right),$$

so we do not think of jumps between subspaces of different dimension. We define a probability distribution over the entire product space of candidate models and their parameters. That is, for

$$(k, \phi) \in \mathcal{K} \times \bigotimes_{k \in \mathcal{K}} \Phi^{(k)}. \tag{13}$$

Thus, in the product space model, we change (12) for (13). The likelihood and the prior structure are defined in a corresponding way as follows; for a particular k the likelihood depends only on the corresponding vector of parameters $\phi^{(k)}$, that is

$$p(\mathbf{y}|\phi, k) = p\left(\mathbf{y}|\phi^{(k)}, k\right).$$

The model will be completed by the prior $p(\phi|k)$ and the prior $p(k)$. Then the full posterior distribution of the product space model can be expressed as

$$\begin{aligned} p(\phi, k|\mathbf{y}) &= \frac{p(\mathbf{y}|\phi, k)p(\phi|k)p(k)}{p(\mathbf{y})} \\ &= \frac{p(\mathbf{y}|\phi^{(k)}, k)p(\phi^{(k)}|k)p(\phi^{(-k)}|\phi^{(k)}, k)p(k)}{p(\mathbf{y})} \\ &= p\left(\phi^{(k)}, k|\mathbf{y}\right)p\left(\phi^{(-k)}|\phi^{(k)}, k\right) \end{aligned} \tag{14}$$

where $\phi^{(-k)}$ denotes the model parameters in $\Phi^{(j)}$ for all $j \neq k$. The specification of the product space model for a given k is completed by the usual hierarchical structure for the models in $\Phi^{(k)}$ and by $p(\phi^{(-k)}|\phi^{(k)}, k)$ which would be the "priors" or pseudo-priors (this term was first used by Carlin and Chib 1995) for the parameters that are not used by models in $\Phi^{(k)}$. We could assign any proper distribution to $p(\phi^{(-k)}|\phi^{(k)}, k)$.

Now, if from (14), we integrate out $\phi^{(j)}$ for all $j \neq k$ we obtain $p(\phi^{(k)}, k|\mathbf{y})$, which is the target distribution.

3.2.2 *Product space Metropolis-Hastings*

Godsill (2001) showed that reversible jump is a special case of a Metropolis-Hastings in the product space. To do this, he obtained the reversible jump's acceptance probability to update a Markov chain at a state (k, ϕ) to a new state (k', ϕ') imposing the proposal (15) and using a Metropolis-Hastings step in the product space in the usual way,

$$\begin{aligned} &q\left(\phi', k'|\phi, k\right) \\ &\quad = q_1\left(k'|k\right)q_2\left(\phi'^{(k')}|\phi^{(k)}\right)p\left(\phi'^{(-k')}|\phi'^{(k')}, k'\right) \end{aligned} \tag{15}$$

where $q_1(k'|k)$ is the proposal to move the model index from k to k' and $q_2(\phi'^{(k')}|\phi^{(k)})$ is the proposal to move the parameters from the model indexed with k to one indexed with k'. Once (15) is set and following the product space model (14), it is easy to see that

$$\begin{aligned} &\alpha_{k,k'}\left(\phi^{(k)}, \phi'^{(k')}\right) \\ &\quad = \min\left(1, \frac{p(\phi', k'|\mathbf{y})q(\phi, k|\phi', k')}{p(\phi, k|\mathbf{y})q(\phi', k'|\phi, k)}\right) \\ &\quad = \min\left(1, \frac{p(\phi'^{(k')}, k'|\mathbf{y})q_1(k|k')q_2(\phi^{(k)}|\phi'^{(k')})}{p(\phi^{(k)}, k|\mathbf{y})q_1(k'|k)q_2(\phi'^{(k')}|\phi^{(k)})}\right) \end{aligned} \tag{16}$$

because the terms

$$p\left(\phi^{(-k)}|\phi^{(k)}, k\right) \quad \text{and} \quad p\left(\phi'^{(-k')}|\phi'^{(k')}, k'\right)$$

are canceled out. But this acceptance probability is rather general. To deduce the most convenient expression for the reversible jump methodology from (16), we assume that the dimension of $\phi^{(k)}$ is n_k and the dimension of $\phi'^{(k')}$ is $n_{k'}$ with $n_{k'} > n_k$. Then, the key idea is to achieve the so called "dimension matching" between $\phi^{(k)}$ and $\phi'^{(k')}$. To this end, we generate a vector or matrix $\mathbf{u}$ with distribution $q_2(\mathbf{u})$ independent of $\phi^{(k)}$ such that the dimension of $(\phi^{(k)}, \mathbf{u})$ is $n_{k'}$ (the idea of Green 1995). Then, we devise a function T such that

$$T\left(\phi^{(k)}, \mathbf{u}\right) = \phi'^{(k')} \quad \text{and} \quad T^{-1}\left(\phi'^{(k')}\right) = \left(\phi^{(k)}, \mathbf{u}\right). \tag{17}$$

Here, to apply the Change of Variable Theorem (CHVT), T must be a bijection and T and T^{-1} must be differentiable.

Then let $q_{2_{\phi^{(k)},\mathbf{u}}}(\cdot|\cdot)$, $q_{2_{\phi^{(k)}}}(\cdot|\cdot)$ denote the joint conditional distributions of $(\phi^{(k)}, \mathbf{u})$, $\phi^{(k)}$, respectively, and $q_{2_{\mathbf{u}}}(\cdot)$ the joint marginal distribution for $\mathbf{u}$. The CHVT can be used to write

$$\begin{aligned} &q_2\left(\phi'^{(k')}|\phi^{(k)}\right) \\ &\quad = q_{2_{\phi^{(k)},\mathbf{u}}}\left(T^{-1}\left(\phi'^{(k')}\right)|\phi^{(k)}\right)\left|\frac{\partial T^{-1}(\phi'^{(k')})}{\partial \phi'^{(k')}}\right| \\ &\quad = q_{2_{\phi^{(k)},\mathbf{u}}}\left(\phi^{(k)}, \mathbf{u}|\phi^{(k)}\right)\left|\frac{\partial T^{-1}(\phi'^{(k')})}{\partial \phi'^{(k')}}\right| \end{aligned}$$

$$= q_{2_{\phi^{(k)}}}\left(\phi^{(k)}|\phi^{(k)}\right) q_{2_{\mathbf{u}}}\left(\mathbf{u}|\phi^{(k)}\right) \left|\frac{\partial T^{-1}(\phi'^{(k')})}{\partial \phi'^{(k')}}\right|$$

$$= q_{2_{\mathbf{u}}}(\mathbf{u}) \left|\frac{\partial T^{-1}(\phi'^{(k')})}{\partial \phi'^{(k')}}\right| \tag{18}$$

because $\mathbf{u}$ is independent of $\phi^{(k)}$, and $q_{2_{\phi^{(k)}}}(\phi^{(k)}|\phi^{(k)}) = 1$. On the other hand $q_2(\phi^{(k)}|\phi'^{(k')})$ is always 1 because in (17) we are assuming that a function such that $\varphi(\phi'^{(k')}) = \phi^{(k)}$ exists.

Then, from (18), (16) becomes

$$\alpha_{k,k'}\left(\phi^{(k)}, \phi'^{(k')}\right) = \min\left(1, \frac{p(\phi'^{(k')}, k'|\mathbf{y}) q_1(k|k')}{p(\phi^{(k)}, k|\mathbf{y}) q_1(k'|k) q_{2_{\mathbf{u}}}(\mathbf{u})} \left|\frac{\partial T(\phi^{(k)}, \mathbf{u})}{\partial(\phi^{(k)}, \mathbf{u})}\right|\right) \tag{19}$$

because of the Inverse Function Theorem (see for example Rudin 1976).

Equation (19) is usually obtained following reversible jump ideas (see Green 1995). However, the product space formulation is a "standard" Metropolis-Hastings.

3.2.3 Product space Metropolis-Hastings for mixtures

For finite mixture models, with an unknown number of components, we need to generate a Markov chain with stationary distribution

$$p\left(\phi^{(k)}, \tau^{(k)}, k|\mathbf{y}\right) \propto p\left(\mathbf{y}|\phi^{(k)}, \tau^{(k)}, k\right) p\left(\phi^{(k)}, \tau^{(k)}|k\right) p(k) \tag{20}$$

where the $(\tau^{(k)})$ are the discrete latent allocation variables of the model with k components, and the $(\phi^{(k)})$ are the continuous parameters of the model. To obtain the acceptance probability to update the chain from state $(\phi^{(k)}, \tau^{(k)})$ to a new state $(\phi'^{(k')}, \tau'^{(k')})$, there is no need to go back to the product space model, we just rewrite the acceptance probability (16) as

$$\alpha_{k,k'}\left(\left(\phi^{(k)}, \tau^{(k)}\right), \left(\phi'^{(k')}, \tau'^{(k')}\right)\right) = \min\left(1, \frac{\pi_{k'}}{\pi_k}\right) \tag{21}$$

with

$$\frac{\pi_{k'}}{\pi_k} = \frac{p(\phi'^{(k')}, \tau'^{(k')}, k'|\mathbf{y}) q_1(k|k')}{p(\phi^{(k)}, \tau^{(k)}, k|\mathbf{y}) q_1(k'|k)} \times \frac{q_2(\phi^{(k)}, \tau^{(k)}|\phi'^{(k')}, \tau'^{(k')})}{q_2(\phi'^{(k')}, \tau'^{(k')}|\phi^{(k)}, \tau^{(k)})}.$$

In our case, the proposals for the continuous variables of the model are going to be independent of the proposed allocations, but for the allocations, the proposal does depend on the proposed continuous parameters. Thus,

$$q_2\left(\phi^{(k)}, \tau^{(k)}|\phi'^{(k')}, \tau'^{(k')}\right) = q_2\left(\phi^{(k)}|\phi'^{(k')}\right) q_2\left(\tau^{(k)}|\phi'^{(k')}, \tau'^{(k')}\right)$$

so under this formulation we can use again (18), to obtain an applied version of (21),

$$\frac{\pi_{k'}}{\pi_k} = \frac{p(\phi'^{(k')}, \tau'^{(k')}, k'|\mathbf{y}) q_1(k|k')}{p(\phi^{(k)}, \tau^{(k)}, k|\mathbf{y}) q_1(k'|k)} \times \frac{q_2(\tau^{(k)}|\phi'^{(k')}, \tau'^{(k')})}{q_2(\tau'^{(k')}|\phi^{(k)}, \tau^{(k)}) q_{2_{\mathbf{u}}}(\mathbf{u})} \left|\frac{\partial T(\phi^{(k)}, \mathbf{u})}{\partial(\phi^{(k)}, \mathbf{u})}\right|.$$

Under the product space model, (21) is a straightforward consequence of (19).

3.3 Sampling strategy

The easiest way to construct a Markov transition kernel with stationary distribution (20), using trans-dimensional methods, is to propose moves to update k one step at a time. That is, randomly choose to attempt the move $k : k+1$ or $k : k-1$ generating proposals

$$\left(\phi'^{(k+1)}, \tau'^{(k+1)}\right) \quad \text{or} \quad \left(\phi'^{(k-1)}, \tau'^{(k-1)}\right)$$

to update $(\phi^{(k)}, \tau^{(k)})$, and if the move is rejected stay at the current state. Then update $[\phi^{(k*)}, \tau^{(k*)}|k]$ via a Metropolis-Hastings or Gibbs kernel. This is a hybrid strategy as discussed in Tierney (1994).

With (21), when attempting $k : k+1$, we substitute $k' = k+1$, and to attempt $k : k-1$ we use

$$\min\left(1, \left(\frac{\pi_k}{\pi_{k-1}}\right)^{-1}\right).$$

The updating step $[\phi^{(k*)}, \tau^{(k*)}|k]$ is done with the Gibbs-Slice sampler described in Sect. 3.1 and Appendix C. We used two trans-dimensional moves: split-combine and birth-death, these are outlined in Richardson and Green (1997). A description of the transformations that we have used is given in Appendix D.

3.4 Model priors

When working with an unknown number of components, a prior for k is needed and under the assumption of no additional information, for us, the best option is to impose a discrete uniform prior

$$p(k) = \frac{1}{k_{max}} \mathbb{1}_{(1,\ldots,k_{max})}^{(k)}$$

for a preselected $k_{max} \in \mathbb{N}$. Such a truncation was also used by Richardson and Green (1997).

Some authors that have worked with normal mixtures suggest a truncated Poisson distribution, see for example Nobile and Fearnside (2007). We believe that this is done with the intention of penalizing higher values for k. But as we have seen, high overall posterior estimates for k occur due to the use of the normal distribution, so we believe we do not need to follow this idea. All the other priors are as in Sect. 3.1.2.

4 Illustrations

4.1 Setting the priors

In this section, we describe how to set the unspecified constants of the model (see Sect. 3.1.2). Let R be the range of the data ($R = y_{(n)} - y_{(1)}$). For the locations we set $\mu_0 = y_{(1)} + R/2$, $\sigma_0^2 = R^2$ and for the finite weights we take $\delta = 1$, giving a uniform prior over the space $w_1 + \cdots + w_k = 1$. For the smoothing parameter of the Dirichlet process we found that results based on $c = 2$ were good, obtaining smooth predictive densities.

The degree of asymmetry allowed is determined by ρ, which bounds the interval where λ can move. We set $\rho = 0.5$ giving room for asymmetry without supporting high variance due to large values of λ; compare the graphics (d) and (f) with graphic (e) in Fig. 1. For the kurtosis parameter, α, we chose different values depending on the data set to analyze. For example, fixing $\alpha = 2.5$ we obtained good results when estimating a normal distribution, while to estimate a Laplace distribution, $\alpha = 1.01$ was a better option. This will be shown later on in the paper. For the remaining unspecified parameters we will follow similar ideas to those of Richardson and Green (1997), pp. 747–748, to devise a "default" prior using the data.

If the asymmetry parameter is zero, the variance of the unimodal distribution is

$$\sigma^2 = \frac{1}{3}\sum_{s=1}^{\infty} w_s \theta_s^2,$$

so proceeding as in Richardson and Green (1997) we can first relate σ to the relevant parameters and then to the range of the data. To relate σ to α, a and b, note that $\mathbb{E}(\theta_s^2) = \frac{\alpha(\alpha+1)}{\beta^2}$ and $\mathbb{E}(1/\beta) = \frac{b}{a-1}$ for $a > 1$. Thus

$$\sigma \sim \frac{1}{\beta}\sqrt{\frac{\alpha(\alpha+1)}{3}} \sim \frac{b}{a-1}\sqrt{\frac{\alpha(\alpha+1)}{3}}.$$

Hence, setting $a = 2.1$ and $b = 0.134R\sqrt{\frac{\alpha(\alpha+1)}{3}}$ we are leaving the Dirichlet process to take care of the tails of each component while being weakly informative about the size of σ.

We do not claim that these choices are non-informative. It is well-known there is no way to be fully non-informative under a Bayesian mixture modeling set-up. In fact, it is well-known that the posterior for k is sensitive to the choice of the prior for the location parameters and it is clear that it is heavily influenced by the choice of prior over the variances; see for example the discussions in Richardson and Green (1997), pp. 747–749, and Jasra et al. (2005), pp. 64–65.

Another way to set the priors for the unimodal distribution is to use an interactive plot of the scaled prior predictive (5):

$$w\mathbb{E}\big(f(y|\lambda, \mu, G)\big)$$

where $0 < w \leq 1$ and in the background a histogram of the data (we generated our interactive graphic in R using the library `tcltk`, see R Development Core Team, R (2011)). The goal here is to obtain a reasonable value to set ρ, α and a sensible knowledge of the range in which β should lie. We set $k_{max} = 30$ in all examples.

4.2 Predictive density estimates

To test the unimodal distribution (4) we draw samples from six unimodal distributions, namely:

1. Normal: N(0, 1).
2. Student's-t: $t_{(2)}$.
3. Gamma: $\Gamma(2, 3)$.
4. Laplace 1: Lap(0, 20, 1).
5. Laplace 2: Lap(0, 1, 2).
6. Pearson IV: PearsonIV(3.94, 4.35, 1, 3.74).

In each case a sample of size $n = 150$ was generated. Since the objective is to assess the range of shapes the unimodal distribution can approximate, by comparing the predictive density with the true density, the samples were not randomly drawn. To improve comparability, we generated a grid of 150 equally spaced points in (0, 1) and then evaluated the quantile function, of each distribution, on every point of the grid. For the skew Laplace and Pearson IV distribution, to calculate the quantiles, the libraries `PearsonDS` and `LaplacesDemon` of R were used, see Byron (2012) and Becker (2012).

For the six unimodal data sets, we ran our fixed k sampler, setting $k = 1$, for 200,000 iterations. The first 100,000 iterations were used as a burn-in period. In each case a predictive density estimate was generated (see Appendix B), and these are displayed in Fig. 2, along with the true density. For the kurtosis parameter, we tried different values and kept the one that gave the best fit. For the remaining parameters we set the priors as described in the previous section.

From the graphics in Fig. 2 we see, first, that the predictive densities capture the correct skewness for both symmetric and asymmetric distributions. Second, in all cases,

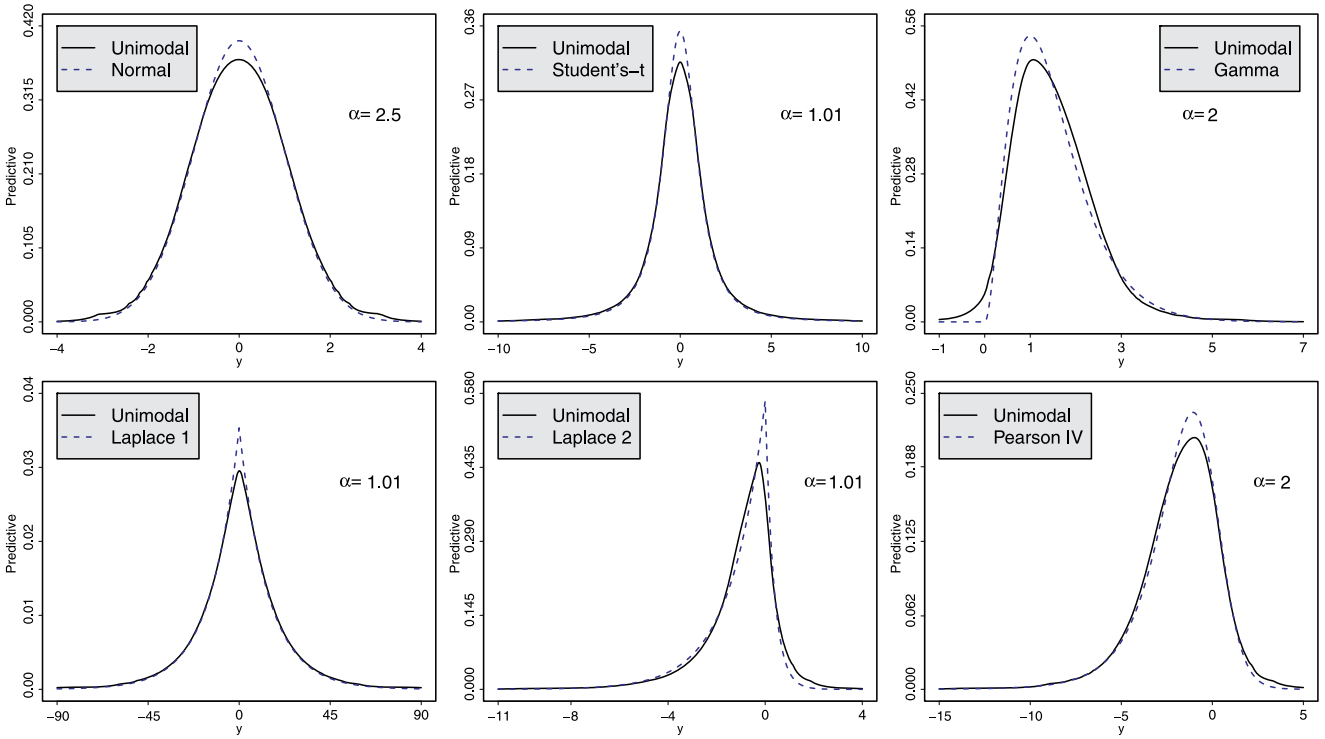

Fig. 2 Density estimates for six unimodal distributions

smooth predictive density estimates are obtained. This indicates that the choice $c = 2$, for the smoothing parameter of the Dirichlet process is a sensible choice. Third, for different values of α, different degrees of kurtosis are obtained and this agrees with what was mentioned in Sect. 2.1.1. Looking back to graphic (c) in Fig. 1, the prior predictive suggested that $\alpha = 3.8$ was a good option to generate the density estimate for the standard normal. But we present the predictive density estimate generated with $\alpha = 2.5$ instead. The reason is that the predictive generated with $\alpha = 3.8$ gave a good fit for the tails of the distribution, but produced a flat density estimate.

4.3 Examples for an unknown number of components

To test the mixture of unimodal distributions (6), simulated and real data sets were analyzed. For the simulated data, mixtures of gamma, skew Laplace and skew normal distributions were used, namely:

$$\text{Model 1: } 0.2\Gamma(40, 20) + 0.6\Gamma(6, 1) + 0.2\Gamma(200, 20)$$

$$\text{Model 2: } 0.2\text{Lap}(-5, 1, .5) + 0.4\text{Lap}(0, 1, 1) + 0.3\text{Lap}(3, 1, 1) + 0.1\text{Lap}(10, 1, 2).$$

$$\text{Model 3: } 0.1\text{SN}(-30, 3, -4) + 0.1\text{SN}(-20, 3, 0) + 0.15\text{SN}(-10, 2, 4) + 0.15\text{SN}(0, 2, -2) + 0.1\text{SN}(10, 2, 3) + 0.1\text{SN}(15, 2, 2) + 0.1\text{SN}(20, 2, 4) + 0.1\text{SN}(30, 2, 0) + 0.1\text{SN}(35, 2, 1).$$

The sample size for Models 1 and 2 was $n = 400$ observations and for Model 3 $n = 600$ observations. Model 1 was used by Wiper et al. (2001) to demonstrate the performance of their reversible jump algorithm for mixtures of gamma distributions. A plot of Models 1 to 3 is shown in Fig. 3.

For the real data sets, we use the same examples used by Richardson and Green (1997): The Galaxy data consists of the velocities (in 1000 km/sec) of 82 distant galaxies diverging from our own from six well separated conic sections of the Corona Borealis region. This data set was first studied by Postman et al. (1986) and it is widely used in the literature to illustrate methodology for mixture modeling. The enzyme data set concerns the distribution of enzymatic activity in the blood of an enzyme involved in the metabolism of a carcinogenic substance, among a group of 245 unrelated individuals. The acidity data set concerns the log scale of the acidity index measured in a sample of 155 lakes in north-central Wisconsin. The three data sets are available at the library `mixAK` of R as `Galaxy`, `Acidity` and `Enzyme`. This package contains a variety of statistical methods including MCMC methods to analyze the data using normal mixtures; see Komárek (2009).

For the multimodal data sets, we ran our hybrid sampler for 1,000,000 iterations, the first 200,000 iterations used as

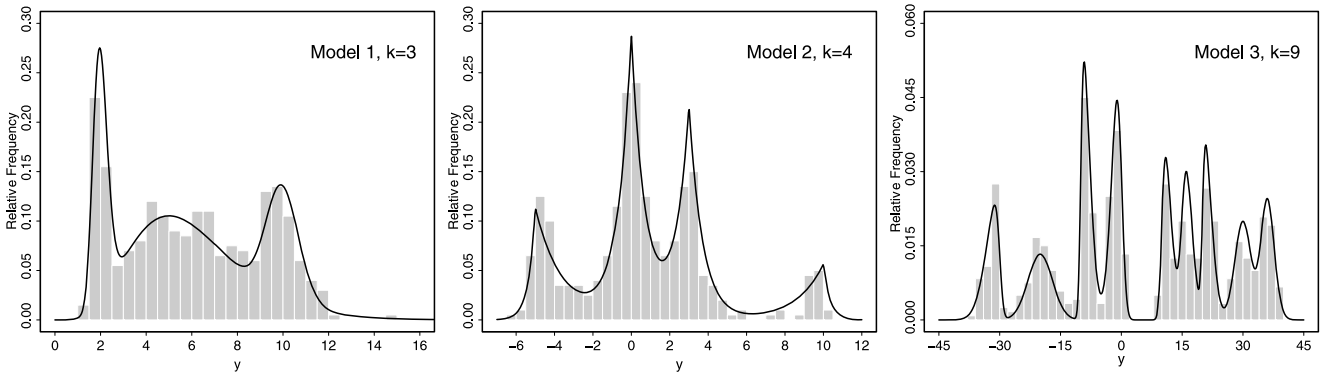

Fig. 3 Model 1; mixture of Gamma distributions, Model 2; mixture of skew Laplace distributions and Model 3 mixture of skew normal distributions

a burn-in period. Our starting point for k in all the runs was $k = 30$. With the same data sets, we also ran the algorithm described by Richardson and Green (1997), for a mixture of normals model with an unknown number of components, using their default priors. The same number of iterations and burn-in period was also used.

Posterior probabilities for k for all data sets are given in Table 1. For an easy comparison with the mixture of normals, the maximum posterior probabilities for k obtained with the mixture of normals algorithm of Richardson and Green (1997) are presented in Table 2. To show that the computational cost of our algorithm is not prohibitive, we present a comparison of CPU times also in Table 2.

In almost all experiments, the posterior distribution for k calculated with the unimodal mixture, supports lower values for k when compared to the normal mixture. There are only two cases when both models are similar in terms of k: the acidity data and the data from Model 2. Finally, we see that for Model 3, the mixture of skew normal distributions, the posterior calculated via the unimodal distribution gives support to low values for k, the maximum is achieved at $k = 9$, which is the true value. On the other hand, with the normal mixture, the maximum for $p(k|\mathbf{y})$ is attained at $k = 12$.

A comparison of predictive densities is shown in Fig. 4. These were calculated with the output of the algorithms for a moving k, see Appendix B. It is interesting to note that without giving support to unusually high number of components, in the posterior for k, the unimodal mixture gives accurate representations of each data set. This is an appealing characteristic of our model.

The acceptance rates for the Metropolis-Hastings algorithm were low but not prohibitively so. After starting the chains at $k_{max} = 30$ they all moved rapidly to a neighborhood close to the highest posterior value for k. It is worth saying that at each iteration of the algorithm the number of variables to be split or combined were $4 + 2 * N$, with N as well varying from iteration to iteration. For example, in the case of the galaxy data, the ergodic average for N stabilizes at $\bar{N} \approx 13$ (graphic not shown). Hence, we were attempting to split, on average, 30 variables.

Table 1 Unimodal mixture: posterior distribution of k for the seven data sets, default priors and taking $\alpha = 2.5$

Data set	n	$p(k\|\mathbf{y})$	
Data from Model 1	400	$p(1) = 0.000$	$p(2) = 0.049$
		$p(3) = 0.226$	p(4) =0.280
		$p(5) = 0.211$	$p(6) = 0.126$
		$\sum_{k>6} p(k) = 0.108$	
Data from Model 2	400	$\sum_{k<4} p(k) = 0.000$	$p(4) = 0.092$
		$p(5) = 0.195$	p(6) =0.227
		$p(7) = 0.196$	$p(8) = 0.133$
		$\sum_{k>8} p(k) = 0.156$	
Data from Model 3	600	$\sum_{k<7} p(k) = 0.000$	$p(7) = 0.122$
		$p(8) = 0.259$	p(9) =0.281
		$p(10) = 0.190$	$\sum_{k>10} p(k) = 0.149$
Enzyme	245	$p(1) = 0.000$	$p(2) = 0.257$
		p(3) =0.325	$p(4) = 0.222$
		$p(5) = 0.112$	$p(6) = 0.052$
		$\sum_{k>6} p(k) = 0.032$	
Acidity	155	$p(1) = 0.000$	$p(2) = 0.149$
		p(3) =0.238	$p(4) = 0.235$
		$p(5) = 0.173$	$p(6) = 0.105$
		$\sum_{k>7} p(k) = 0.1$	
Galaxy	82	$p(1) = 0.004$	$p(2) = 0.057$
		$p(3) = 0.154$	p(4) =0.217
		$p(5) = 0.212$	$p(6) = 0.161$
		$\sum_{k>6} p(k) = 0.195$	

4.4 Examples for a known number of components

The likelihood of a mixture model is invariant under permutation of their parameters. When symmetric priors are used, the posterior will inherit the invariance of the likelihood. As

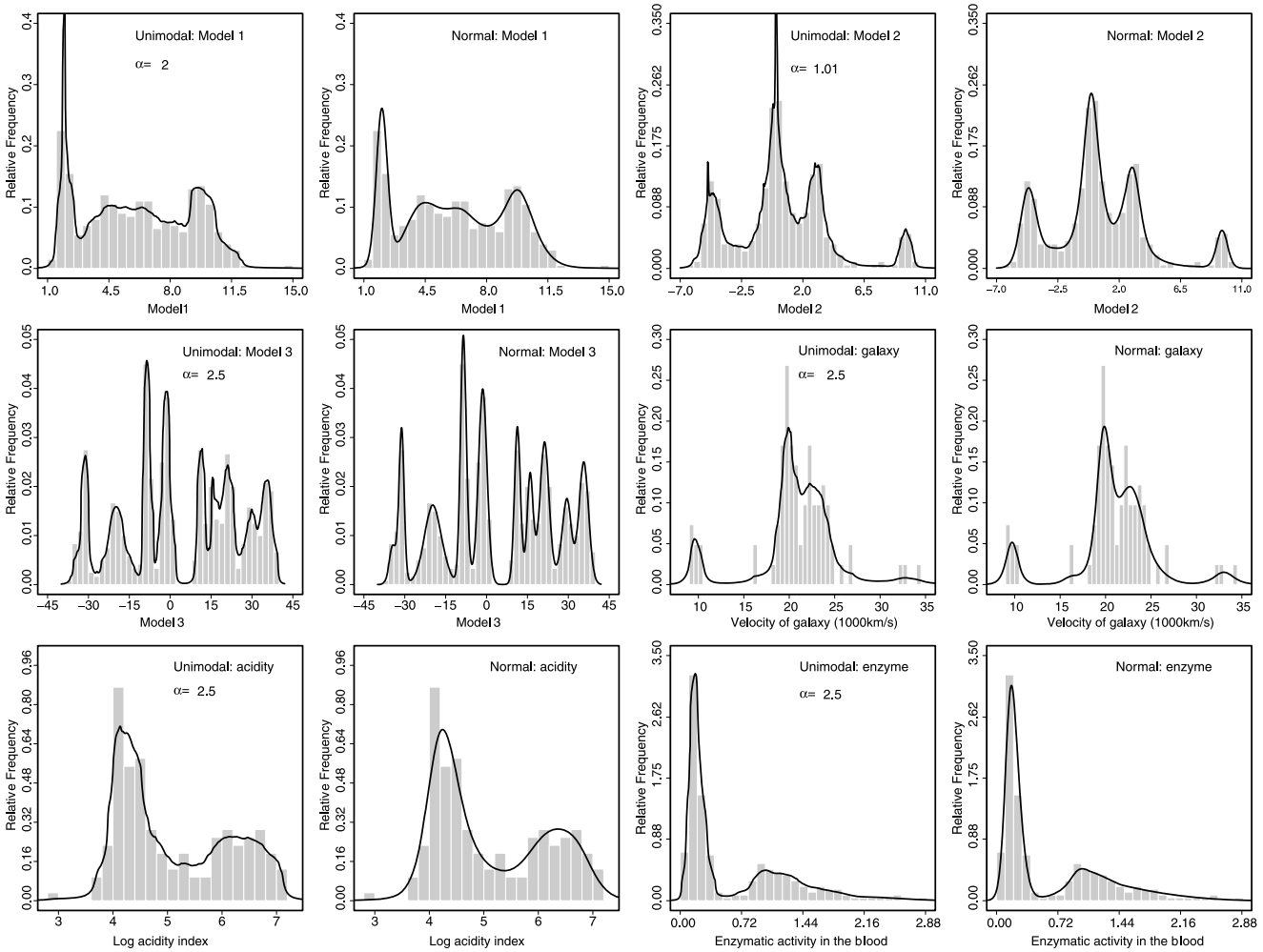

Fig. 4 Comparison of Predictive densities: unimodal vs normal mixture models, default priors

Table 2 True k (if known), normal mixture: maximum posterior distribution of k (default priors) and time comparison between unimodal and normal mixture algorithms

Data set	True k	normal mixture $\max\{p(k\|\mathbf{y})\}$	approx time (minutes)	
			normal	unimodal
Model 1	3	$p(5) = 0.260$	8	15
Model 2	4	$p(6) = 0.191$	10	20
Model 3	9	$p(12) = 0.192$	27	29
Enzyme	–	$p(4) = 0.339$	4	9
Acidity	–	$p(3) = 0.258$	3	8
Galaxy	–	$p(6) = 0.189$	3	7

a result, in an MCMC algorithm, the permutation can change multiple times between the iterations of the sampler. Hence, we will not be able to identify the hidden groups that we are looking for. This makes ergodic averages to estimate characteristics of the components useless: this is known as the label switching problem.

Thus, to make inference for individual components and classification, we need to "undo the label switching". With this aim we use the deterministic relabeling algorithm of Rodríguez and Walker (2012). Their idea is based on the meaning of the relationship between the allocation variables and the observations: the latent allocation z_i indicates the component or cluster from which y_i has been drawn, see Sect. 2.2.1. From iteration to iteration of an MCMC algorithm the labels of the clusters may change. But if the sampler has converged, the clusters must remain roughly the same. Using this fundamental idea we can devise a k-means type of diverging measure to keep track of each cluster at each iteration of the sampler, where all that is needed are the observations and the latent allocations. Rodríguez and Walker (2012) compared their algorithm against the relabeling algorithms of Stephens (2000b) and Papastamoulis and Iliopoulos (2010), obtaining favorable results.

We ran the fixed k algorithms, with default priors, for the unimodal and normal mixture. We used the galaxy, acidity and enzyme data assuming, for the unimodal mixture,

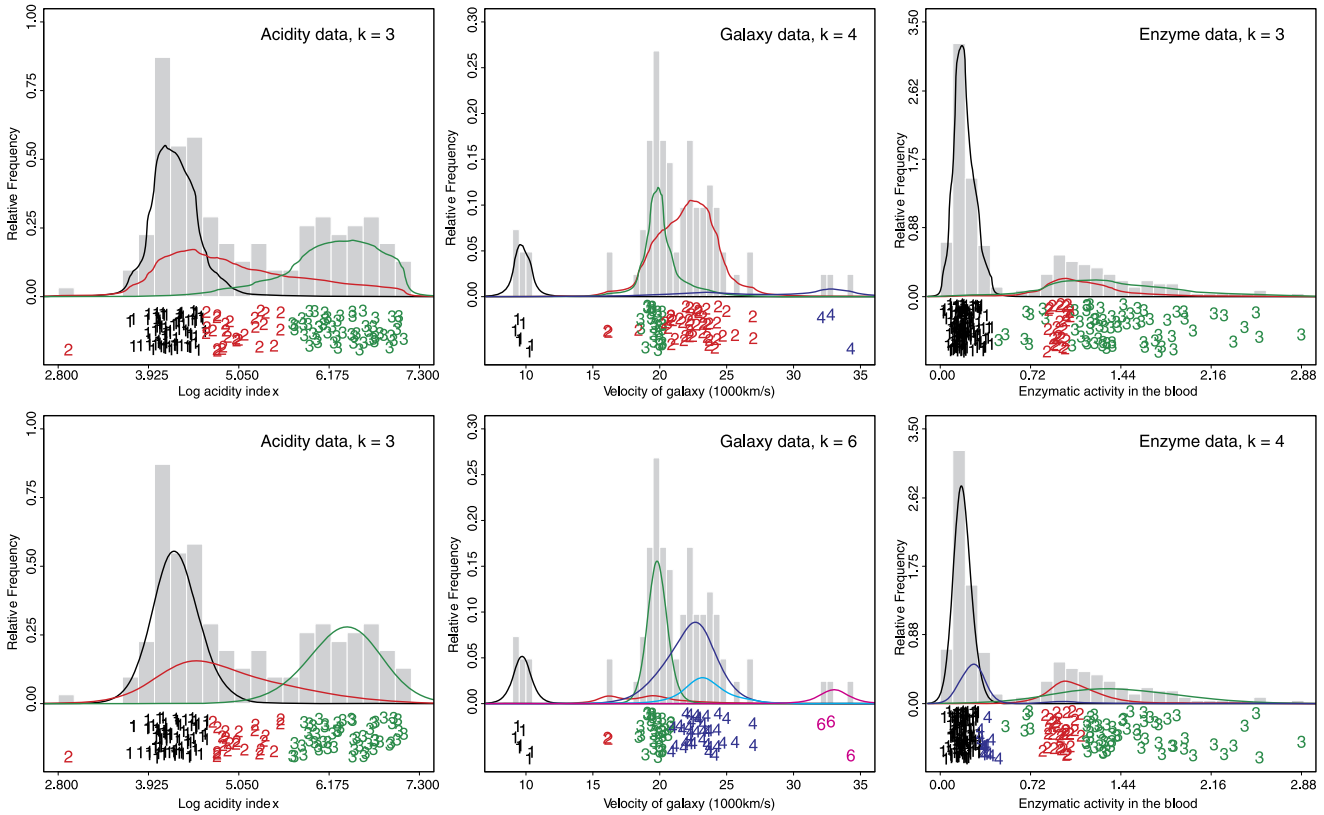

Fig. 5 Unimodal (*top row*) and normal mixture for a known k: estimated scaled densities and single best clustering

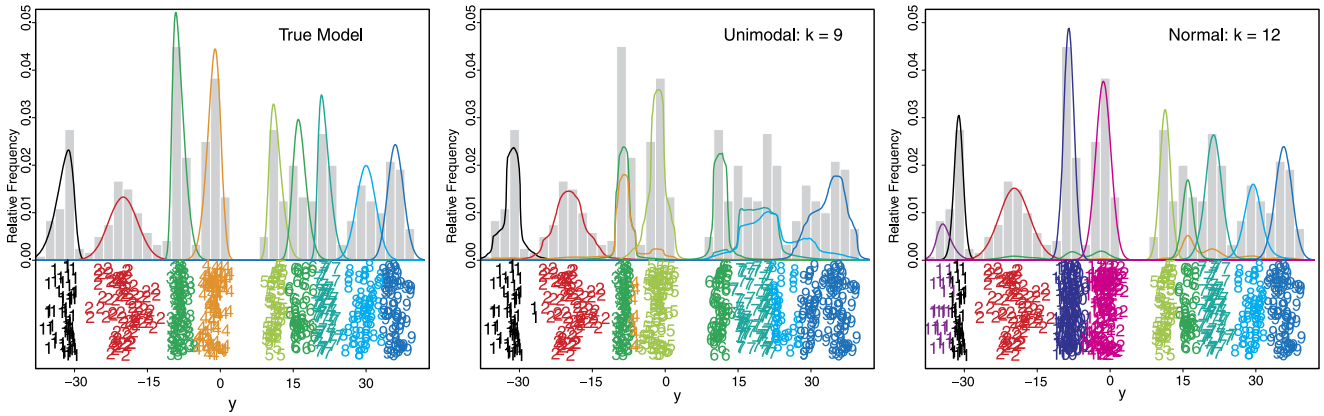

Fig. 6 True, estimated scaled densities and single best clustering for Model 3

$k = 4$, $k = 3$ and $k = 3$, respectively, and $k = 6$, $k = 3$ and $k = 4$, for the normal mixture. Note that these are the values for which the posterior distributions attained its maximum for each algorithm. To test our model when the number of underlying components is large, we performed the same comparison using Model 3. In all cases, we generated 200,000 iterations, throwing the first 100,000 iterations as a burn-in period. Then we post-processed the MCMC output using Rodríguez and Walker (2012) ideas to undo the label-switching and estimated the scaled densities (see Appendix B) and single best clustering. The results are displayed in Fig. 5. For Model 3 the results are displayed in Fig. 6, here a plot of the scaled densities and single best clustering calculated via the true model was included.

For the acidity data, the estimated scaled densities and single best clustering obtained with the unimodal and normal mixtures are similar. However, from the normal perspective, where a cluster should be a set of observations adequately modeled by a normal distribution, we observe skew scaled densities which makes no sense. For the galaxy data,

in the case of the unimodal mixture, we see four unimodal clusters where each cluster has at least one observation assigned to it. Instead, in the normal case, cluster five has no observations assigned. But beyond this, and more importantly, we observe again a small degree of skewness in clusters three and four of the mixture of normals. For the enzyme data, with the unimodal mixture, from the single best clustering we are able to identify three skew clusters. Instead, to model skewness, the normal mixture needs two normals to model a single cluster (see the cluster labeled as cluster one). For the data set from Model 3, the single best clustering calculated with the unimodal mixture allocates observations to eight clusters, missing group six. There are problems when the components are overlapped, and this can be seen from components three to five. With the normal mixture, the observations are allocated to ten clusters instead of twelve. Here again it is clear that in presence of skew components two normals are needed to model a single skew component, this can be seen in clusters one and eleven for example.

4.5 Computational issues

All the experiments were performed in a Dell Precision M4-400 (processor Intel core 2 Duo at 2.26 GHz) running Linux openSUSE 12.1. We coded our approach in **C** and used the `.C` Interface to R to have a friendly data input interface, see Peng and Leeuw (2002) for a good introduction. To manage the random seed and generate the random variates from the common distributions we used the GSL-GNU Scientific Library, see Galassi et al. (2009). The analysis and graphics were done in R. All the reported CPU times were measured using the function `system.time` of R.

5 Discussion

We have extended the mixture model of Richardson and Green (1997) to allow for clusters to be modeled by unimodal distributions. In all the examples considered, we obtained accurate representations of the data, without giving support to unusually high number of components. This is an appealing feature of the model and provides k with an explicit interpretation: the number of clusters modeled by unimodal densities.

In the absence of further information, it is natural to associate the number of clusters with the number of unimodal distributions. Hence, if we assume unimodality instead of normality for the components distribution of a mixture, we are giving a proper meaning to k. For this, obviously, $f(y|\lambda, \mu, G)$ must be unimodal, which is based on Feller's representation of unimodal and symmetric distributions.

It is fair to say that replacing the nonparametric density (4) with a flexible parametric family, which includes skewness and kurtosis parameters, would result in a simpler model. However, we believe that this would bring problems for modeling tails. A parametric model carries a certain type of heavy tail and given the nature of the problem we are tackling, tails will be playing an important role. It is imperative not to get them wrong, and parametric models offer this opportunity. Nonparametric models do not.

We believe that the ideas of Godsill (2001) nicely complement the paper of Green (1995) and allow us to think clearly about how to deal with a trans-dimensional problems. We also believe that we would not have been able to implement an MCMC strategy without the use of the slice variables needed to sample each $(f(y|\lambda_j, \mu_j, G_j))$. These two concepts together give us the ability to move a stochastic process across sub-spaces of different dimensions.

One of the aims is to extend our ideas to regression models and also a multivariate model. For the latter we have to construct a nonparametric unimodal and multivariate distribution, which is not so straightforward. The initial idea is to describe a multivariate unimodal distributions via

$$Y_j = U_j Z_j \quad \text{for } j = 1, \ldots, p,$$

where the $U = (U_1, \ldots, U_p)$ are i.i.d standard normal and the $Z = (Z_1, \ldots, Z_p)$ have any joint distribution. We would be trying to extend Khinchin's characterization (Khinchin 1938) of univariate unimodal distributions to the multivariate setting, see Devroye (1997) and Tao (1989).

From the four parameter densities: the Pearson IV family of densities and the densities summarized in Rigby and Stasinopoulos (2005), pp. 516–517, which includes the Box-Cox power exponential among others, we believe that the prior predictive (5) deserves further study. Its closed form is easy to write in full and its parameters can be easily understood.

6 Supplemental material

The supplemental material is available on-line and it includes all the appendices mentioned in the paper (Appendix.pdf).

References

Becker, M.: PearsonDS: Pearson Distribution System, R Package Version 0.93, http://CRAN.R-project.org/package=PearsonDS (2012)

Blackwell, D.: The discreteness of Ferguson selections. Ann. Stat. **1**, 356–358 (1973)

Brunner, J.L., Lo, A.Y.: Bayes methods for symmetric unimodal density and its mode. Ann. Stat. **17**, 1550–1566 (1989)

Byron, H.: LaplacesDemon: Software for Bayesian Inference, R Package Version 12.01.02, http://cran.r-project.org/web/packages/LaplacesDemon/index.html (2012)

Carlin, B.P., Chib, S.: Bayesian model choice via Markov chain Monte Carlo methods. J. R. Stat. Soc. B **57**, 473–484 (1995)

Damien, P., Walker, S.G.: Sampling truncated normal, beta and gamma densities. J. Comput. Graph. Stat. **10**, 206–215 (2001)

Devroye, L.: Random variate generation for multivariate unimodal densities. ACM Trans. Model. Comput. Simul. **7**, 447–477 (1997)

Diebolt, J., Robert, C.: Estimation of finite mixture distributions through Bayesian sampling. J. R. Stat. Soc. B **56**, 363–375 (1994)

Escobar, M.D.: Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University (1988)

Escobar, M.D.: Estimating normal means with a Dirichlet process prior. J. Am. Stat. Assoc. **89**, 268–277 (1994)

Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**, 577–588 (1995)

Feller, W.: An Introduction to Probability Theory and Its Applications, pp. 157–158. Wiley, New York (1971)

Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)

Ferguson, T.S.: Bayesian density estimation by mixtures of normal distributions. In: Chernoff, H., Rustagi, J.S., Rizvi, M.H., Siegmund, D. (eds.) Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday, pp. 287–302. Academic Press, New York (1983)

Fernandez, C., Steel, M.F.J.: On Bayesian modeling of fat tails and skewness. J. Am. Stat. Assoc. **93**, 359–371 (1998)

Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer Series in Statistics. Springer, New York (2006)

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F.: GNU Scientific Library Reference Manual, Network Theory Limited. http://www.gnu.org/software/gsl/ (2009)

Godsill, S.J.: On the relationship between Markov chain Monte Carlo methods for model uncertainty. J. Comput. Graph. Stat. **10**, 230–248 (2001)

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

Green, P.J.: Trans-dimensional Markov chain Monte Carlo. In: Green, P.J., Hjort, N.L., Richardson, S. (eds.) Highly Structured Stochastic Systems, pp. 179–198. Oxford University Press, Oxford (2003)

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

Kalli, M., Griffin, J.E., Walker, S.G.: Slice sampling mixture models. Stat. Comput. **21**, 93–105 (2011)

Khinchin, A.Y.: On unimodal distributions. Trams. Res. Inst. Math. Mech. (University of Tomsk) **2**, 1–7 (1938) (in Russian)

Komárek, A.: A new R Package for Bayesian estimation of multivariate normal mixtures allowing for selection of number of components and interval-censored data. Comput. Stat. Data Anal. **53**, 3932–3947 (2009)

Kottas, A., Gelfand, A.E.: Bayesian semiparametric median regression modeling. J. Am. Stat. Assoc. **96**, 1456–1468 (2001)

Lo, A.Y.: On a class of Bayesian nonparametric estimates: I. Density estimates. Ann. Stat. **12**, 351–357 (1984)

Nobile, A., Fearnside, A.T.: Bayesian finite mixtures with an unknown number of components: the allocation sampler. Stat. Comput. **17**, 147–162 (2007)

Papaspiliopoulos, O., Roberts, G.O.: Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika **95**, 169–186 (2008)

Papastamoulis, P., Iliopoulos, G.: An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. J. Comput. Graph. Stat. **19**, 313–331 (2010)

Peng, R.D., Leeuw, J.: An Introduction to the `.C` Interface to R. UCLA: Academic Technology Services, Statistical Consulting Group, http://www.ats.ucla.edu/stat/r/library/interface.pdf (2002)

Postman, M., Huchra, J.P., Geller, M.J.: Probes of large-scale structures in the Corona Borealis region. Astrophys. J. **92**, 1238–1247 (1986)

Quintana, F.A., Steel, M.F.J., Ferreira, J.T.A.S.: Flexible univariate continuous distributions. Bayesian Anal. **4**, 497–522 (2009)

R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. http://www.R-project.org/ (2011)

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

Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. J. R. Stat. Soc. C **54**, 507–554 (2005)

Rodríguez, C.E., Walker, S.G.: Label switching in Bayesian mixture models: deterministic relabeling strategies (2012, submitted manuscript)

Rudin, W.: Principles of Mathematical Analysis, pp. 221–223. McGraw-Hill, New York (1976)

Sethuraman, J.: A constructive definition of Dirichlet priors. Stat. Sin. **4**, 639–650 (1994)

Sisson, S.A.: Transdimensional Markov chains: a decade of progress and future perspectives. J. Am. Stat. Assoc. **100**, 1077–1089 (2005)

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

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

Tao, D.: On multivariate unimodal distributions. University of British Columbia, MSc Thesis, https://circle.ubc.ca/handle/2429/27411 (1989)

Tierney, L.: Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762 (1994)

Walker, S.G.: Sampling the Dirichlet mixture model with slices. Commun. Stat. **36**, 45–54 (2007)

Wiper, M., Insua, R.D., Ruggeri, F.: Mixtures of gamma distributions with applications. J. Comput. Graph. Stat. **10**, 440–454 (2001)