Bayesian finite mixtures with an unknown number of components: the allocation sampler

Agostino Nobile* and Alastair Fearnside

University of Glasgow, U.K.

2nd June 2005

Abstract

A new Markov chain Monte Carlo method for the Bayesian analysis of finite mixture distributions with an unknown number of components is presented. The sampler is characterized by a state space consisting only of the number of components and the latent allocation variables. Its main advantage is that it can be used, with minimal changes, for mixtures of components from any parametric family, under the assumption that the component parameters can be integrated out of the model analytically. Artificial and real data sets are used to illustrate the method and mixtures of univariate normals, of multivariate normals and of uniform distributions are explicitly considered. Issues of label switching, when parameter inference is of interest, are addressed in a post-processing stage.

Keywords: Classification; Label switching; Markov chain Monte Carlo; Multivariate normal mixtures; Mixtures of uniforms; Normal mixtures.

[*] *Address for correspondence*: Agostino Nobile, Department of Statistics, University of Glasgow, Glasgow G12 8QW, U.K.

Email: agostino@stats.gla.ac.uk

1 Introduction

Finite mixture distributions are receiving increasing interest as a way of modelling population heterogeneity and, to a larger extent, as a means of relaxing distributional assumptions. Monographs on finite mixtures include, among others, Titterington *et al.* (1985) and McLachlan and Peel (2000). Böhning and Seidel (2003) is a recent review with emphasis on nonparametric maximum likelihood, while Marin *et al.* (2005) is an introduction from a Bayesian perspective. Although statistical finite mixtures date as far back as Newcomb (1886) and Pearson (1894), their widespread use was for long prevented by the difficulties associated with their estimation. A major breakthrough occurred with the appearance of the EM algorithm of Dempster *et al.* (1977), and the associated idea of explicitly representing, by means of latent allocation variables, the mixture components generating each observation. The very same idea plays a central role in the Bayesian approach using Markov chain Monte Carlo methods, such as the Gibbs sampler of Diebolt and Robert (1994). Inference about the number of components in the mixture has been more difficult, see Böhning and Seidel (2003) for a brief summary and pointers to the non-Bayesian literature. Within the Bayesian approach, a definite advance has been the application by Richardson and Green (1997) of the reversible jump MCMC method of Green (1995), which allowed one to sample from the joint posterior distribution of all the parameters, including the number k of components. Beside Richardson and Green (1997), other researchers have studied methods to estimate the posterior distribution of k. Some of them (Nobile 1994 and 2005, Roeder and Wasserman 1997) have provided estimates of the marginal likelihoods of k components, then used Bayes theorem to obtain the posterior of k. Others (Phillips and Smith 1996, Stephens 2000a) have derived MCMC methods that share with Richardson and Green's the idea of running a sampler on a composite model, so that the posterior of k can be estimated by the relative frequency with which each model is visited during the simulation. Some other authors (Carlin and Chib 1995, Chib 1995, Raftery 1996) preferred to avoid placing a prior distribution on k, instead, they estimated the marginal likelihoods of k components and possibly used Bayes factors to test k vs. $k+1$ components. Mengersen and Robert (1996) too employed a testing approach, however, they relied on the Kullback–Leibler divergence as a measure of distance between mixtures with k and $k+1$ components. Representations of the marginal likelihoods for k components have been derived by Nobile (1994, 2004) and Ishwaran *et al.* (2001).

Most authors within the Bayesian approach have devised posterior sampling schemes to draw from the joint distribution of mixture parameters and allocations. Only a few, among which Nobile (1994), Casella *et al.* (2000), Steele *et al.* (2003), have preferred to work in terms of the allocation variables only, after analytically integrating the parameters out of the model. The present paper belongs to this thread and presents a new MCMC sampler, which we call *the allocation sampler*, that makes draws from the joint posterior distribution of the number of components and the allocation variables. The main advantage of this approach is that the sampler remains essentially the same, irrespective of the data dimensionality and of the family of mixture components. In contrast, the reversible jump method of Richardson and Green (1997) requires the invention of "good" jumping moves, to apply it to a new family of mixtures; this has slowed its application to mixtures of multivariate normal distributions, however, see Dellaportas and Papageorgiou (2003).

The allocation sampler consists of several moves, some of which change the number of components k. We illustrate its performance with real and artificial data, reporting examples of posterior inference for k, for the mixture parameters and for future observables. Meaningful parametric inference in mixture models requires to tackle the label switching problem, see Celeux *et al.* (2000), Stephens (2000b), Frühwirth-Schnatter (2001). To this purpose, we adapt a proposal of Stephens (2000b) to the situation where only a sample of the allocations is available.

2 The model

We assume that random variables $x_1, \ldots, x_n$ are independent and identically distributed with density (with respect to some underlying measure)

$$f(x|k, \lambda, \theta) = \sum_{j=1}^{k} \lambda_j q_j(x|\theta_j), \tag{1}$$

where $\lambda_j > 0, j = 1, \ldots, k$ and $\sum_{j=1}^{k} \lambda_j = 1$. Several parameters enter in the finite mixture (1): the number of components k, the mixture weights $\lambda = (\lambda_1, \ldots, \lambda_k)$ and the components' parameters $\theta = (\theta_1, \ldots, \theta_k)$. In this paper we regard all these parameters as unknowns. We also assume that the mixture components q_j belong to the same parametric family. As an aside on notation, we will use q for mixture component densities, π for priors and posteriors, p for (prior and posterior) predictives and f for all other densities.

A useful way of thinking of model (1) is as follows: each observation has probability λ_j of arising from component j in the mixture. More precisely, let g_i be the index or label of the component that generated x_i; the latent vector $g = (g_1, \ldots, g_n)^\top$ is called the allocation vector. We assume that the g_i's are conditionally independent given k and λ with $\Pr[g_i = j|k, \lambda] = \lambda_j$, so that

$$f(g|k, \lambda) = \prod_{j=1}^{k} \lambda_j^{n_j} \tag{2}$$

where n_j is the number of observations allocated by g to component j: $n_j = \text{card}\{A_j\}$ and $A_j = \{i : g_i = j\}$. Conditional on g, the density of x_i is q_{g_i} and

$$f(x|k, \lambda, \theta, g) = \prod_{i=1}^{n} q_{g_i}(x_i|\theta_{g_i}). \tag{3}$$

Integrating the density in equation (3) with respect to the conditional distribution of g given in (2) produces the finite mixture (1).

The specification of a Bayesian finite mixture model requires prior distributions on k, λ and θ. We use as prior on k the $Poi(1)$ distribution restricted to $1 < k \leq k_{\max}$, in the examples in this paper we used $k_{\max} = 50$. Other authors have used priors on k proportional to Poisson distributions: Phillips and Smith (1996) with mean 3, Stephens (2000a) with means 1, 3 and 6. For a justification of the $Poi(1)$ prior on k, see Nobile (2005). Conditional on k, the weights λ are assumed to have a $Dir(\alpha_1, \ldots, \alpha_k)$ distribution, where the α's are positive constants. Independent priors are assigned to the parameters in θ:

$$\pi(\theta|k, \phi) = \prod_{j=1}^{k} \pi_j(\theta_j|\phi_j), \tag{4}$$

where ϕ_j is a possibly empty set of hyperparameters and $\phi = \{\phi_1, \ldots, \phi_k\}$.

As already mentioned, the distinguishing feature of our approach is that it uses a model where the mixture weights λ and the parameters θ have been integrated out. Integrating the density (2) with respect to the $Dir(\alpha_1, \ldots, \alpha_k)$ distribution of the weights gives

$$f(g|k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \prod_{j=1}^{k} \frac{\Gamma(\alpha_j + n_j)}{\Gamma(\alpha_j)}, \tag{5}$$

where $\alpha_0 = \sum_{j=1}^{k} \alpha_j$.

We assume that the independent priors on the θ_j's are chosen in a way that these parameters can also be integrated out analytically from expression (3), as will be the case if priors $\pi_j(\theta_j|\phi_j)$ which are conjugate to the distributions $q_j(x|\theta_j)$ are employed. Multiplying (3) by (4) and integrating with respect to θ yields

$$f(x|k, g, \phi) = \prod_{j=1}^{k} p_j(x^j|\phi_j) \tag{6}$$

where

$$p_j(x^j|\phi_j) = \int \prod_{i \in A_j} q_j(x_i|\theta_j)\pi_j(\theta_j|\phi_j)\, d\theta_j \tag{7}$$

is the marginal density of the observations $x^j = \{x_i : i \in A_j\}$ allocated to component j, after integrating with respect to the prior of θ_j, and $p_j(x^j|\phi_j) = 1$ if $A_j = \emptyset$.

Before proceeding we make some remarks about the hyperparameters (α_j, ϕ_j), $j = 1, \ldots, k$. We assume that the hyperparameters of the j-th component are the same for all values of $k \geq j$. This assumption ensures that the family of mixture models we entertain is nested and the corresponding marginal likelihoods are related, see Nobile (2004). One important special case occurs when $\alpha_j = \alpha_1$, $\phi_j = \phi_1$, $j = 1, \ldots, k$, so that the prior is symmetric with respect to a permutation of the components' labels. We refer to this situation as the *symmetric case* and on occasion, to emphasize that it is not assumed, we will use the term *asymmetric case*. Clearly, if any information distinguishing the components is available, it should be incorporated in the prior distribution of the θ's and λ's. We only provide some suggestions for prior specification in the symmetric case. The weights hyperparameters $\alpha_1, \ldots, \alpha_k$ are treated as fixed constants; $\alpha_j = \alpha_1 = 1$ is the most common choice, corresponding to a uniform distribution on the simplex of weights, and we have adopted it here. Experimentation has shown that not only parametric inference, but also inference about the number of components may be very sensitive to changes in some of the components' hyperparameters ϕ_j. This is the reason why we have listed ϕ to the right of the conditioning bar in formulae (4) and (6). Nobile (2005) contains an example of marked sensitivity of the posterior of k to changes in ϕ, for mixtures of univariate normal distributions. He proposes to adopt an empirical Bayes approach, where some components of the ϕ_j's are fixed using basic features of the data, while others are estimated using a posterior sample from a preliminary MCMC run conducted with a hyperprior on ϕ. That approach is adopted and extended in the present paper, details are in Sections 2.1 and 2.2.

The allocation sampler, to be discussed in Section 3, aims at sampling from the joint posterior distribution of k and g:

$$\pi(k, g|x, \phi) \propto f(k, g, x|\phi) = \pi(k)f(g|k)f(x|k, g, \phi) \tag{8}$$

where $f(g|k)$ and $f(x|k, g, \phi)$ are given by formulae (5) and (6). If a hyperprior is placed on ϕ, an additional Metropolis-Hastings move is used to update ϕ given k, g and x. Before discussing the allocation sampler, we illustrate how the model specialises to the cases of univariate and multivariate normal components. More details on these, as well as on the cases of Poisson and Exponential components, can be found in Nobile (1994, Chapter 2). We also consider mixtures of uniforms, as an example of a situation where conjugate priors $\pi_j(\theta_j|\phi_j)$ are not available.

2.1 Mixtures of univariate normals

If the mixture components are univariate normals, then $q_{g_i}(x_i|\theta_{g_i})$ in (3) is the density $N(x_i|m_{g_i}, r_{g_i}^{-1})$ of a normal distribution with mean m_{g_i} and variance $r_{g_i}^{-1}$. The priors $\pi_j(\theta_j|\phi_j)$ in (4) are the usual conjugate priors for (m_j, r_j): $r_j \sim Ga(\gamma_j, \delta_j)$ and $m_j|r_j \sim N(\mu_j, \{\tau_j r_j\}^{-1})$, so that $\phi_j = (\mu_j, \tau_j, \gamma_j, \delta_j)$. Given k and g, the marginal distribution of the data allocated to the j-th component is given by

$$\begin{aligned} p_j(x^j|\phi_j) \quad &= \quad \pi^{-n_j/2}\left[\frac{\tau_j}{\tau_j + n_j}\right]^{1/2} \frac{\Gamma(\gamma_j + \{n_j/2\})}{\Gamma(\gamma_j)} \quad \cdot \\ & \qquad (2\delta_j)^{\gamma_j}\left\{2\delta_j + \sum_{i \in A_j}(x_i - \overline{x}_j)^2 + \frac{\tau_j n_j}{\tau_j + n_j}(\overline{x}_j - \mu_j)^2\right\}^{-(\gamma_j + \{n_j/2\})} \end{aligned}$$

where $\overline{x}_j = (1/n_j)\sum_{i \in A_j} x_i$. As for the hyperparameters, we assume a symmetric prior and follow the approach of Nobile (2005). The overall mean μ_1 is fixed to a round value close to the sample mean $\overline{x}$. The prior predictive distribution is t with $2\gamma_1$ degrees of freedom, we use $\gamma_1 = 2$ to have a prior predictive with relatively thick tails, but finite second moments. Independent priors are placed on the other two hyperparameters, with $(1 + \tau_1)^{-1} \sim Un(0, 1)$ and $\delta_1 \sim Un(0, \delta_U)$ where $\delta_U = (\gamma_1 - 1)s_x^2$ and s_x^2 is the sample variance. Draws from a preliminary run of the sampler are used to make boxplots of the marginal posterior distributions of τ_1 and δ_1 conditional on k. Estimates $\hat{\tau}_1$ and $\hat{\delta}_1$ are computed using medians of the posterior draws, but keeping only the draws that correspond to values of k after a rough leveling off of the medians has occurred. These estimates are then used as the values of τ_1 and δ_1 in subsequent runs.

2.2 Mixtures of multivariate normals

For b-variate normal components, the densities $q_{g_i}(x_i|\theta_{g_i})$ in (3) are $N_b(x_i|m_{g_i}, r_{g_i}^{-1})$, where now m_{g_i} and r_{g_i} denote the mean vector and precision matrix (inverse covariance) of the multivariate normal. Conjugate priors are placed on the pairs (m_j, r_j), with $r_j \sim W_b(\nu_j, \xi_j)$, a Wishart distribution with ν_j degrees of freedom and precision matrix ξ_j, while $m_j|r_j \sim N_b(\mu_j, \{\tau_j r_j\}^{-1})$, with μ_j a b-vector and τ_j a positive real number. The hyperparameters for component j are $\phi_j = \{\mu_j, \tau_j, \nu_j, \xi_j\}$. The marginal density of the data allocated to component j, given k and g, is

$$\begin{aligned} p_j(x^j|\phi_j) &= \pi^{-bn_j/2} \left[\frac{\tau_j}{\tau_j + n_j}\right]^{b/2} \prod_{s=1}^{b} \frac{\Gamma(\{\nu_j + n_j + 1 - s\}/2)}{\Gamma(\{\nu_j + 1 - s\}/2)} \; |\xi_j|^{\nu_j/2} \; \cdot \\ & \left| \xi_j + \sum_{i \in A_j} (x_i - \overline{x}_j)(x_i - \overline{x}_j)^\top + \frac{\tau_j n_j}{\tau_j + n_j} (\overline{x}_j - \mu_j)(\overline{x}_j - \mu_j)^\top \right|^{-(\nu_j + n_j)/2} \end{aligned}$$

where $\overline{x}_j$ is the sample mean vector of the observations allocated to component j. A symmetric prior is assumed and we set $\mu_1 = \overline{x}$, the sample mean vector. The prior predictive distribution is multivariate t with $\nu_1 - b + 1$ degrees of freedom, which when set equal to 4 yields $\nu_1 = b + 3$. For the remaining hyperparameters τ_1 and ξ_1, we assume that ξ_1 is diagonal, then apply a procedure similar to the one employed for univariate normal components.

2.3 Mixtures of uniforms

Mixtures of uniform distributions $Un(a, b)$ are not identifiable, not even up to a permutation of the components' labels (Titterington *et al.*, 1985, page 36). However, this does not prevent their use for density estimation, only inference about the components' parameters. Assume that the densities $q_{g_i}(x_i|\theta_{g_i})$ in (3) are $Un(x_i|a_{g_i}, b_{g_i})$ and let (a_j, b_j), $j = 1, \ldots, k$, be a priori independent with densities $\pi_j(a_j, b_j) = 1/(2\phi_j^2)$, $-\phi_j < a_j < b_j < \phi_j$, where $\phi_j = \phi$ is a positive constant. Also assume, for simplicity, that the data contains no ties. Then, a straightforward but tedious computation shows that, conditional on k and g, the marginal density of the data allocated to component j is

$$p_j(x^j|\phi) = \begin{cases} \frac{1}{2\phi^2} \frac{\left[(x^j_{(n_j)}-x^j_{(1)})^{-n_j+2}-(x^j_{(n_j)}+\phi)^{-n_j+2}-(\phi-x^j_{(1)})^{-n_j+2}+(2\phi)^{-n_j+2}\right]}{(n_j-1)(n_j-2)} & n_j > 2 \\ \frac{1}{2\phi^2} \left[\log \frac{(\phi-x^j_{(1)})(\phi+x^j_{(2)})}{2\phi(x^j_{(2)}-x^j_{(1)})}\right] & n_j = 2 \\ \frac{1}{2\phi^2} \left[x^j_1 \log \frac{\phi-x^j_1}{\phi+x^j_1} + \phi \log \frac{(2\phi)^2}{(\phi-x^j_1)(\phi+x^j_1)}\right] & n_j = 1 \end{cases} \tag{9}$$

where $x^j_{(i)}$ denotes the i-th order statistic of x^j. The constant ϕ is determined using a variant of the procedure used for univariate normal components, with prior distribution $\phi^{-1} \sim Un(0, 1/\max_i |x_i|)$.

3 The allocation sampler

This section discusses how to sample from the joint posterior distribution of k and g given in (8). If a hyperprior distribution is placed on ϕ, an additional Metropolis-Hastings move is needed to update ϕ. Otherwise, the sampler we use runs on k and g only and comprises two types of moves: moves that do not affect the number of components and moves that change it; each sweep of the allocation sampler begins with a random selection of the move to be performed. The first type of moves consists of (i) Gibbs sampling on the components of g, (ii) two Metropolis-Hastings moves to simultaneously change several allocations and (iii) a Metropolis-Hastings move on the component labels. Moves akin to (i) and (iii) were used by Nobile (1994). A relabelling move was also employed by Frühwirth-Schnatter (2001) in her permutation sampler, but it affected λ and θ, beside g. The second type of moves consists of an ejection/absorption Metropolis-Hastings move: either a mixture component is absorbed into another one or, the reverse move, a component ejects another component. These moves are similar in spirit to the reversible jump moves of Richardson and Green (1997). In fact one can think of the allocation sampler as a version of reversible jump on a state space consisting of k and g only, so that the transitions occur between discrete spaces with different number of elements, rather than spaces of varying dimensions. We prefer to use the terms *ejection* and *absorption*, rather than *split* and *combine*, since they convey asymmetric roles for the components involved, we found this helpful in devising the moves.

3.1 Moves that do not change the number of components

These moves keep k at its current value and update only g.

3.1.1 Gibbs sampling of g

The first move is a systematic sweep Gibbs sampler on the components of g, from g_1 to g_n. To sample g_i we need its full conditional distribution. This can be easily computed by first evaluating k times the joint density $f(k, g, x|\phi)$, with $g_i = 1, \ldots, k$ and all the other quantities at their current values, then renormalizing the resulting values. In fact, only $k-1$ evaluations are necessary, since $f(k, g, x|\phi)$ at the current g is already known. Also, changing the current g_i to the other $k-1$ possible values only affects two terms in (5) and (6), so the computation time increases linearly with k.

This Gibbs sampler changes only one entry of g at a time, so one can expect very strong serial dependence of the sampled g's, especially for moderate to large sample sizes n. Therefore, it makes sense to also have moves that attempt to change several entries of g simultaneously. This is easily accomplished by means of the Metropolis-Hastings algorithm, the details are in Section 3.1.2.

3.1.2 Metropolis-Hastings moves on g

In the first move, two components, say j_1 and j_2, are randomly selected among the k available. A draw p_1 is made from the $Beta(\alpha_{j_1}, \alpha_{j_2})$ distribution, then each observation in components j_1 and j_2 is re-allocated to component j_1 with probability p_1 or to component j_2 with probability $1-p_1$. The result is a candidate allocation g' which is accepted with probability $\min\{1, R\}$, where

$$R = \frac{f(k, g', x|\phi)}{f(k, g, x|\phi)} \frac{P(g' \to g)}{P(g \to g')} \tag{10}$$

and $P(g \to g')$ is the probability of proposing the candidate g' when the current state is g. One can show, see the Appendix, that $P(g' \to g)/P(g \to g') = f(g|k)/f(g'|k)$, so that the Metropolis-Hastings acceptance probability ratio in (10) reduces to

$$R = \frac{f(x|k, g', \phi)}{f(x|k, g, \phi)}. \tag{11}$$

The computation of $f(x|k, g', \phi)$ only involves a change in two terms in the product (6).

The second move selects a group of observations currently allocated to some component, say j, and attempts to re-allocate them to another component, say j'. In detail, j and j' are randomly drawn from the k components. If $n_j = 0$ the move fails outright. Otherwise, m is drawn from a discrete uniform distribution on $\{1, \ldots, n_j\}$. Then, m observations are randomly selected among the n_j in component j and moved to component j'. This results in a candidate g' which is accepted with probability $\min\{1, R\}$, where R is given by (10). In this move, however, the proposal ratio can be easily shown to be

$$\frac{P(g' \rightarrow g)}{P(g \rightarrow g')} = \frac{n_j}{n_{j'} + m} \frac{n_j! \, n_{j'}!}{(n_j - m)! \, (n_{j'} + m)!}.$$

Again, computing $f(k, g', x|\phi)$ requires only a change of two terms in (5) and (6).

3.1.3 Metropolis-Hastings move on the labels

Conditional on k, there are $k!$ allocation vectors g which partition the data x into k non-empty groups in the same way, only differing in the assignment of labels to the groups (for simplicity we disregard the possibility of empty components). Thus to each g with high posterior probability there correspond other $k! - 1$ g's, obtained by permuting the component labels, also relatively likely a posteriori. In fact, in the symmetric case, all the $k!$ g's have the same posterior probability. The Gibbs and Metropolis-Hastings moves described above, will only move slowly from one region of high posterior probability to another, let alone visit all the $k!$ regions in a typical length simulation run. This may or may not be a serious problem. Obviously, a problem arises in the asymmetric case, since there the labelling currently visited by the sampler may not be the one that best matches the prior to the groups in the data. For these reasons, we also employ a Metropolis-Hastings relabelling move. Given the current state g, a candidate state g' is generated by randomly selecting two components and exchanging their labels. As the proposal kernel is symmetric, the candidate g' is accepted with probability $\min\{1, f(k, g', x|\phi)/f(k, g, x|\phi)\}$. We perform this relabelling move only in the asymmetric case, where information distinguishing the components is incorporated in the prior. If parameter inference is of interest, labels are re-assigned for the symmetric case in a post-processing stage, after the sample has been collected, see Section 4.6.

3.2 Moves that change the number of components

This section describes the moves that change the number of components in the mixture. They consist of a Metropolis-Hastings pair of moves: a component ejects another component and, in the reverse move, a component absorbs another component.

Assume that the current state is $\{k, g\}$ and let $k_{\max}$ be the maximum allowed number of mixture components. An ejection move is attempted with probability p_k^e, where $p_k^e = 1/2$, $k = 2, \ldots, k_{\max} - 1$, $p_1^e = 1$ and $p_{k_{\max}}^e = 0$, otherwise an absorption move is attempted. Suppose that an ejection is attempted and denote the candidate state by $\{k', g'\}$ with $k' = k + 1$. The candidate is accepted as the next state according to the usual Metropolis-Hastings acceptance probability $\min\{1, R\}$, where

$$R = \frac{f(k', g', x|\phi)}{f(k, g, x|\phi)} \frac{P(\{k', g'\} \to \{k, g\})}{P(\{k, g\} \to \{k', g'\})}. \tag{12}$$

In the reverse move, an attempted absorption from $\{k', g'\}$ to $\{k, g\}$ is accepted with probability $\min\{1, 1/R\}$, with R as given in (12).

We have yet to describe how g' is proposed and how the proposal probabilities in (12) are computed. The procedure to form a proposal is slightly different between the asymmetric and symmetric cases, although the proposal probabilities do not change. We discuss the asymmetric case first. In an ejection move, with current state $\{k, g\}$, one of the k mixture components, say j_1, is randomly selected as the ejecting component, while the ejected component is assigned label $j_2 = k + 1$. A draw p_E is made from a $Beta(a, a)$ distribution and each observation currently allocated to the ejecting component is randomly re-allocated with probability p_E to component j_2 and with probability $1 - p_E$ to component j_1. The parameter a can have a critical effect on the performance of the sampler. We choose it to ensure that empty components j_2 are proposed relatively often (see the Appendix for the details). This provided reasonably good results in our experiments. If $\tilde{n}_{j_1}$ and $\tilde{n}_{j_2}$ are the numbers of observations re-allocated to components j_1 and j_2, then the probability of the resulting allocation, after integrating with respect to the distribution of p_E, is $\Gamma(2a)\Gamma(a + \tilde{n}_{j_1})\Gamma(a + \tilde{n}_{j_2}) / \{\Gamma(a)\Gamma(a)\Gamma(2a + n_{j_1})\}$. Therefore, when at $\{k, g\}$, the candidate $\{k', g'\}$ is proposed with probability

$$P(\{k, g\} \to \{k', g'\}) = p_k^e \frac{1}{k} \frac{\Gamma(2a)}{\Gamma(a)\Gamma(a)} \frac{\Gamma(a + \tilde{n}_{j_1})\Gamma(a + \tilde{n}_{j_2})}{\Gamma(2a + n_{j_1})}. \tag{13}$$

In the reverse absorption move, the absorbed component has label $j_2 = k' = k+1$, while the absorbing component is randomly selected from the remaining k components. All the observations currently allocated to component j_2 are re-allocated to component j_1. Hence the proposal probability is

$$P(\{k', g'\} \rightarrow \{k, g\}) = (1 - p_k^e)\frac{1}{k}. \tag{14}$$

Therefore, the ratio of proposal probabilities in (12) is

$$\frac{P(\{k', g'\} \rightarrow \{k, g\})}{P(\{k, g\} \rightarrow \{k', g'\})} = \frac{1 - p_k^e}{p_k^e} \frac{\Gamma(a)\Gamma(a)}{\Gamma(2a)} \frac{\Gamma(2a + n_{j_1})}{\Gamma(a + \tilde{n}_{j_1})\Gamma(a + \tilde{n}_{j_2})}. \tag{15}$$

The computation of $f(k', g', x|\phi)$ in (12) again requires the change of only two terms in (5) and (6).

In the symmetric case we can improve mixing by randomly selecting both the absorbing and the absorbed components. Reversibility then requires that the ejected component in an ejection should be any of the resulting $k+1$ components. This is easily achieved by including in the ejection move a swap between the label $j_2 = k+1$ of the ejected component and the label of a randomly selected component, including the ejected itself. As a result, the proposal probabilities in (13) and (14) are both multiplied by $1/(k+1)$ and their ratio remains as given in (15).

We conclude this discussion with a remark about implementation. If the absorbed component is randomly selected, a successful absorption can create a gap in the sequence of components. The gap could be easily filled, by moving into it the last component or by decreasing by one all the labels of the components greater than the absorbed one. These adjustments, however, are unnecessary, since in the symmetric case the labels are just place-holders. Therefore, we prefer to save some computation time, by allowing gaps in the sequence of labels to arise, while storing in an additional vector the information about which components are currently in the mixture.

3.3 Some examples with artificial data

As an illustration of the performance of the allocation sampler, we applied it to samples of size 50, 200, 500 and 2000 from a few mixtures of univariate normals, displayed in Figure 1. In order to improve comparability, the samples were not randomly drawn, instead they were obtained by evaluating the quantile function of the mixture on a grid in $(0, 1)$. The model and prior were as detailed in Section 2.1. For each sample a preliminary run was made with random τ and δ, consisting

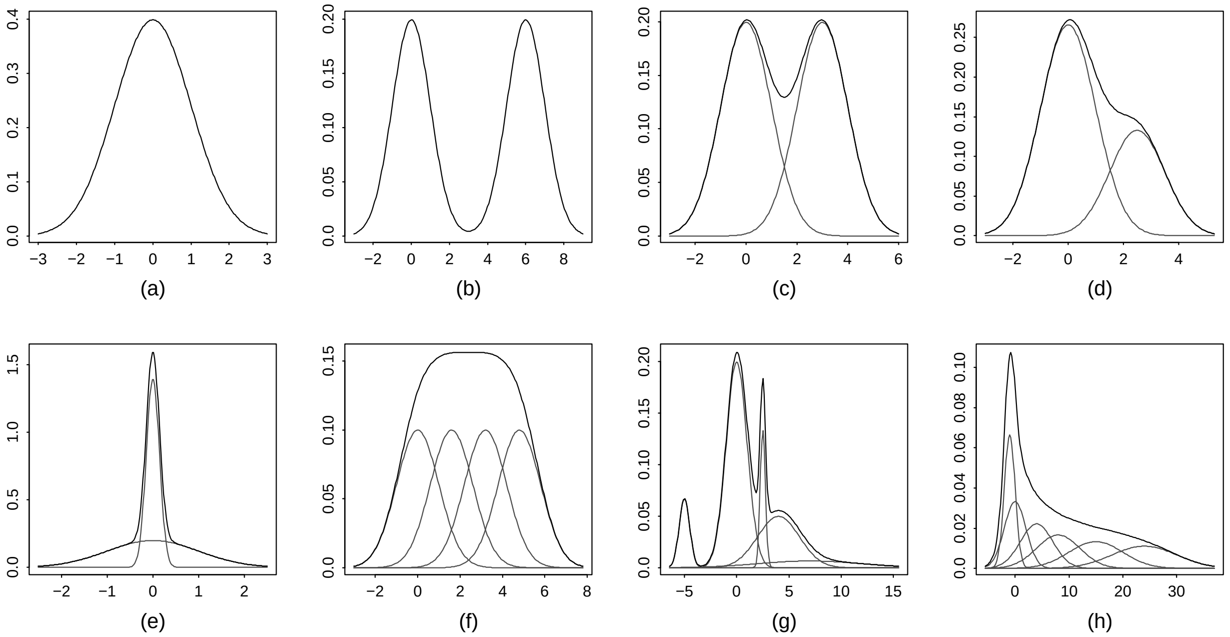

FIGURE 1: Density functions of some mixtures of univariate normal distributions. (a) $N(0,1)$, (b) $\frac{1}{2}N(0,1)+\frac{1}{2}N(6,1)$, (c) $\frac{1}{2}N(0,1)+\frac{1}{2}N(3,1)$, (d) $\frac{2}{3}N(0,1)+\frac{1}{3}N(2.5,1)$, (e) $\frac{1}{2}N(0,1)+\frac{1}{2}N(0,7^{-2})$, (f) $\sum_{j=1}^{4}\frac{1}{4}N(1.6\{j-1\},1)$, (g) $\frac{6}{12}N(0,1)+\frac{3}{12}N(4,2^2)+\frac{1}{12}N(-5,2^{-2})+\frac{1}{12}N(2.5,4^{-2})+\frac{1}{12}N(7,5^2)$, (h) $\sum_{j=1}^{6}\frac{1}{6}N(j(j-2),j^2)$.

of 500,000 sweeps, plus 10,000 sweeps of burn-in, with a thinning of $\Delta = 10$, i.e. only 1 draw every 10 was kept. The preliminary run was used to estimate τ and δ as detailed in Section 2.1 and also to select a thinning value Δ likely to achieve a lag 1 autocorrelation of about 0.7 in the sampled k's. The following runs comprised $10,000\,\Delta$ sweeps, plus $1,000\,\Delta$ sweeps of burn-in. Five independent runs of the sampler were made for each sample. Summaries of the estimated posteriors of k with relative standard deviations are reported in Table 1. For all mixtures, the posterior probability of the true number of components increases with the size n of the sample. For mixtures (a)-(e) the value of k with highest posterior mass is always equal to the true k, except for mixture (d) with $n = 50$. Mixtures (f), (g) and (h) are trickier, however, as n increases, the mode of $\pi(k|x)$ tends to move towards the true k. In the case of mixture (g) the modal and true k coincide by $n = 500$, while for mixtures (f) and (h) they still differ even with $n = 2000$.

The sampler was coded in Fortran, when possible the programs of Nobile (1994) were adapted. Simulation times on a PC with a 2.6 GHz 32-bit processor ranged between a few seconds for

	Sample size											
Mixture	50			200			500			2000		
	k	$\pi(k\mid x)$	s.d.	k	$\pi(k\mid x)$	s.d.	k	$\pi(k\mid x)$	s.d.	k	$\pi(k\mid x)$	s.d.
(a)	1	$\boxed{.557}$	.015	1	$\boxed{.644}$	.009	1	$\boxed{.726}$	.015	1	$\boxed{.805}$	.010
$k_{\text{true}} : 1$	2	.296	.004	2	.250	.007	2	.207	.006	2	.164	.007
	3	.110	.008	3	.080	.008	3	.053	.009	3	.027	.005
	1	.000	.000	1	.000	.000	1	.000	.000	1	.000	.000
(b)	2	$\boxed{.740}$	.010	2	$\boxed{.785}$	.009	2	$\boxed{.842}$	.008	2	$\boxed{.905}$	.005
$k_{\text{true}} : 2$	3	.214	.007	3	.184	.009	3	.139	.010	3	.087	.003
	4	.041	.005	4	.028	.001	4	.017	.002	4	.007	.002
	1	.198	.012	1	.001	.002	1	.000	.000	1	.000	.000
(c)	2	$\boxed{.473}$	.008	2	$\boxed{.607}$	.011	2	$\boxed{.645}$	.011	2	$\boxed{.774}$	.009
$k_{\text{true}} : 2$	3	.235	.005	3	.292	.010	3	.270	.007	3	.193	.007
	4	.073	.004	4	.080	.006	4	.071	.006	4	.030	.003
	1	$\boxed{.386}$	.008	1	.022	.007	1	.000	.000	1	.000	.000
(d)	2	.379	.007	2	$\boxed{.576}$	.009	2	$\boxed{.596}$	.008	2	$\boxed{.666}$	.010
$k_{\text{true}} : 2$	3	.169	.005	3	.287	.006	3	.290	.008	3	.260	.009
	4	.050	.004	4	.090	.005	4	.088	.003	4	.063	.004
	1	.000	.000	1	.000	.000	1	.000	.000	1	.000	.000
(e)	2	$\boxed{.600}$	.005	2	$\boxed{.614}$	.008	2	$\boxed{.615}$	.006	2	$\boxed{.628}$	.008
$k_{\text{true}} : 2$	3	.292	.005	3	.294	.008	3	.297	.005	3	.289	.007
	4	.086	.003	4	.076	.005	4	.075	.003	4	.071	.006
	1	$\boxed{.449}$	.011	1	.075	.015	1	.000	.000	1	.000	.000
(f)	2	.348	.011	2	$\boxed{.557}$	.008	2	$\boxed{.600}$	.008	2	.255	.018
$k_{\text{true}} : 4$	3	.148	.001	3	.271	.005	3	.297	.004	3	$\boxed{.488}$	.012
	4	.044	.003	4	.076	.005	4	.083	.005	4	.199	.008
	5	.010	.002	5	.017	.002	5	.017	.002	5	.048	.003
	1	.144	.004	1	.000	.000	1	.000	.000	1	.000	.000
	2	$\boxed{.439}$	.005	2	.000	.000	2	.000	.000	2	.000	.000
(g)	3	.283	.002	3	.015	.003	3	.000	.000	3	.000	.000
$k_{\text{true}} : 5$	4	.103	.006	4	$\boxed{.416}$	.007	4	.038	.006	4	.000	.000
	5	.025	.002	5	.377	.006	5	$\boxed{.508}$	.011	5	$\boxed{.697}$	.012
	6	.005	.001	6	.146	.004	6	.321	.009	6	.240	.009
	7	.001	.000	7	.037	.002	7	.106	.005	7	.054	.003
	1	.001	.000	1	.000	.000	1	.000	.000	1	.000	.000
	2	$\boxed{.546}$	.011	2	.037	.004	2	.000	.000	2	.000	.000
(h)	3	.330	.009	3	$\boxed{.565}$	.009	3	$\boxed{.441}$	.014	3	.012	.007
$k_{\text{true}} : 6$	4	.099	.002	4	.299	.005	4	.397	.009	4	$\boxed{.649}$	.010
	5	.020	.002	5	.082	.002	5	.130	.004	5	.269	.008
	6	.003	.001	6	.015	.002	6	.028	.004	6	.059	.006

Table 1: Posterior probabilities of selected values of k for representative samples of sizes 50, 200, 500 and 2000 from the mixtures displayed in Figure 1. Probability $\pi(k|x)$ is the average of five estimates from independent runs of the allocation sampler, s.d. is the standard deviation of the five estimates. Modal values are enclosed in a box.

mixture (a) with $n = 50$ and about 7 hours for mixture (g) with $n = 2000$. For each mixture, average time per sweep increased roughly linearly with n. However, samples of larger size usually required larger thinning values Δ (which varied between 10 and 400) and this affected computation times considerably. Figure 2 provides an idea of the mixing of the allocation sampler, using the settings discussed above.

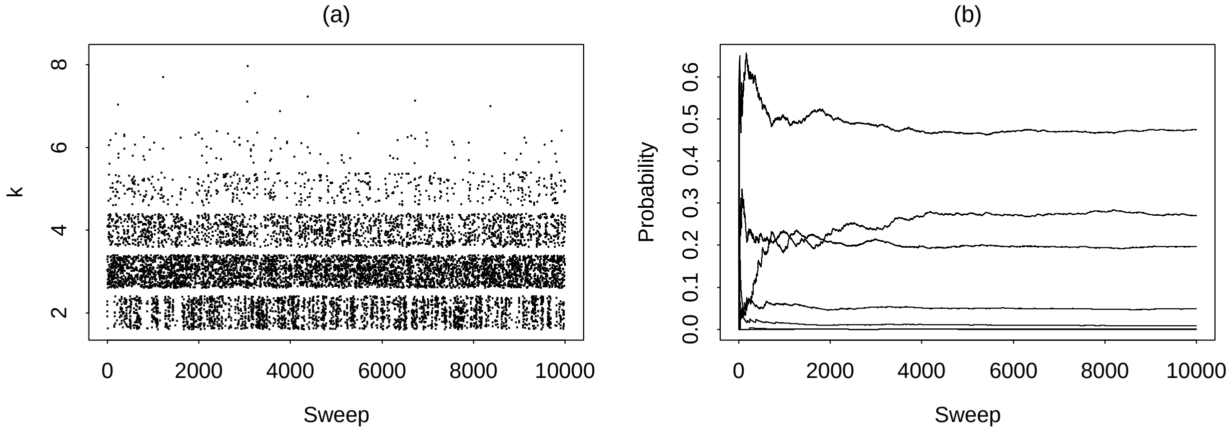

FIGURE 2: Sampled values of k for mixture (f) in Figure 1 with sample size $n = 2000$, in the first of five independent runs: (a) jittered time series plot of k; (b) Estimate of $\pi(k|x)$ across the simulation run. The number of sweeps on the x-axis should be multiplied by the thinning constant, in this instance $\Delta = 320$.

Although these examples concerned mixtures of univariate normals only, we believe the results to be indicative of the performance of the sampler for mixtures of other distributions too. As we mentioned earlier, the family of mixture components affects the sampler only through the form of the densities $p_j(x^j|\phi_j)$ in formula (6).

4 Parameter and predictive posterior distributions

Parameter inference in finite mixture models is somewhat problematic. To begin with, the number of parameters can be very large, so they lose their appeal as a convenient summary of the features of a model. Moreover, inference about the components' weights and parameters makes little sense unless it is carried out conditionally on the number of mixture components. A further complication is that the mixture likelihood is invariant to permutations of the component labels. Nonetheless, posterior distributions of the parameters still play an important role, at the very least as a route to the computation of posterior predictive distributions of future observables.

4.1 Posterior distributions

Given k and g, the prior independence between λ and θ, as well as between the θ_j's, is preserved a posteriori. Conditional on k and g, the posterior distribution of the weights is Dirichlet:

$$\lambda|k, g, x \sim Dir(\alpha_1 + n_1, \dots, \alpha_k + n_k). \tag{16}$$

For the components parameters θ one has:

$$\pi(\theta|k, g, x, \phi) = \prod_{j=1}^{k} \pi_j(\theta_j|x^j, \phi_j), \tag{17}$$

where $\pi_j(\theta_j|x^j, \phi_j)$ denotes the posterior of θ_j given that g allocates x^j to component j. In general, this distribution can be written as

$$\pi_j(\theta_j|x^j, \phi_j) = \frac{\pi_j(\theta_j|\phi_j) \prod_{i \in A_j} q_j(x_i|\theta_j)}{p_j(x^j|\phi_j)} \tag{18}$$

where the normalizing constants $p_j(x^j|\phi_j)$ were defined in (7). If conjugate priors $\pi_j(\theta_j|\phi_j)$ are used, the factors in (17) take the simple form $\pi_j(\theta_j|x^j, \phi_j) = \pi_j(\theta_j|\phi_j')$, where ϕ_j' is the updated value of the hyperparameter, according to the relevant rule for the family of distributions in question.

Marginal posterior distributions unconditional on g are obtained by averaging (16) and (17) with respect to the posterior of g:

$$\begin{aligned} \lambda|k, x &\sim \sum_g \pi(g|k, x) Dir(\alpha_1 + n_1, \dots, \alpha_k + n_k), \\ \pi(\theta|k, x, \phi) &= \sum_g \pi(g|k, x) \prod_{j=1}^{k} \pi_j(\theta_j|x^j, \phi_j). \end{aligned}$$

Of course, marginally the prior independence is lost. Estimates of $\pi(\lambda|k, x)$ and $\pi(\theta|k, x, \phi)$ can be produced using the output from the allocation sampler, by averaging (16) and (17) over the sampled g's, corresponding to a given value of k.

In the symmetric case, no information distinguishing the components is present in the prior. Since the likelihood (1) is also invariant with respect to a permutation of the components labels, so is the posterior. Thus, there is one common marginal posterior distribution for all the λ_j's and one common marginal posterior for all the θ_j's. In this case, meaningful parametric inference requires the adoption of a particular ordering of the labels. This must be enforced either during the simulation or in a post-processing stage, since MCMC samplers typically wander between label orderings, the so-called label switching phenomenon. We return to this topic and our proposed solution in Section 4.6.

4.2 Predictive distributions

The posterior predictive distribution of a future observation x_{n+1} can be easily computed as follows. Conditional on k, λ, θ, g and x the future observable x_{n+1} is independent of the past data x and has density as in (1):

$$f(x_{n+1}|k,\lambda,\theta,g,x,\phi) = \sum_{j=1}^{k} \lambda_j q_j(x_{n+1}|\theta_j).$$

Integrating this density with respect to the joint distribution of λ and θ given k, g and x (see Nobile 1994, page 28, for the details) yields

$$f(x_{n+1}|k,g,x,\phi) = \sum_{j=1}^{k} \frac{\alpha_j + n_j}{\alpha_0 + n}\, p_j(x_{n+1}|x^j,\phi_j) \tag{19}$$

where

$$p_j(x_{n+1}|x^j,\phi_j) = \int q_j(x_{n+1}|\theta_j)\pi_j(\theta_j|x^j,\phi_j)d\theta_j \tag{20}$$

is the posterior predictive density of x_{n+1} according to component j. Finally, the posterior predictive of x_{n+1} is obtained by averaging (19) with respect to the joint posterior distribution of k and g:

$$p(x_{n+1}|x,\phi) = \sum_{k,g} \pi(k,g|x,\phi) \sum_{j=1}^{k} \frac{\alpha_j + n_j}{\alpha_0 + n}\, p_j(x_{n+1}|x^j,\phi_j).$$

To condition on a certain number of components k, the average should instead be taken with respect to $\pi(g|k,x,\phi)$. An estimate of $p(x_{n+1}|x,\phi)$ can be easily produced from the output of the allocation sampler, by taking the average of the right hand side of formula (19) over the sampled pairs $\{k,g\}$.

A simpler expression is available for the posterior predictives $p_j(x_{n+1}|x^j,\phi_j)$. Substituting (18) in (20) gives

$$\begin{aligned} p_j(x_{n+1}|x^j,\phi_j) &= \frac{1}{p_j(x^j|\phi_j)} \int \prod_{i\in A_j\cup\{n+1\}} q_j(x_i|\theta_j)\pi_j(\theta_j|\phi_j)\, d\theta_j \\ &= \frac{p_j(\tilde{x}^j|\phi_j)}{p_j(x^j|\phi_j)} \end{aligned} \tag{21}$$

where $\tilde{x}^j$ denotes the vector x^j augmented with x_{n+1}: $\tilde{x}^j = \{x_i : i \in A_j \cup \{n+1\}\}$ and we used formula (7). If conjugate priors are used, then $\pi_j(\theta_j|x^j, \phi_j) = \pi_j(\theta_j|\phi'_j)$ and substituting this in (20) yields

$$\begin{aligned} p_j(x_{n+1}|x^j, \phi_j) &= \int q_j(x_{n+1}|\theta_j)\pi_j(\theta_j|\phi'_j)d\,\theta_j \\ &= p_j(x_{n+1}|\phi'_j) \end{aligned}$$

where, again, we used (7).

4.3 Mixtures of univariate normals

For univariate normal components and the prior described in Section 2.1, the posteriors $\pi_j(\theta_j|x^j, \phi_j)$ in (17) are as follows: $r_j|g, x \sim Ga(\gamma'_j, \delta'_j)$ and $m_j|r_j, g, x \sim N(\mu'_j, \{\tau'_j r_j\}^{-1})$, where

$$\gamma'_j = \gamma_j + \frac{n_j}{2}, \qquad \delta'_j = \delta_j + \frac{1}{2}\sum_{i \in A_j}(x_i - \overline{x}_j)^2 + \frac{\tau_j n_j}{2(\tau_j + n_j)}(\overline{x}_j - \mu_j)^2,$$

$$\tau'_j = \tau_j + n_j, \qquad \mu'_j = \frac{\tau_j\mu_j + n_j\overline{x}_j}{\tau_j + n_j}.$$

The posterior predictive density $p_j(x_{n+1}|x^j, \phi_j)$ is the density of a univariate t distribution with $2\gamma'_j$ degrees of freedom, location μ'_j and precision $\{\gamma'_j/\delta'_j\}\{\tau'_j/(1+\tau'_j)\}$.

We provide an illustration using the galaxy data of Roeder (1990). These data consist of velocity measurements (1000 Km/sec) of 82 galaxies from the Corona Borealis region. Aitkin (2001) compares likelihood and Bayesian analyses of the data. Two histograms, with different bin widths, are displayed in Figure 3. Table 2 contains an estimate of $\pi(k|x)$ obtained using the allocation sampler. The hyperparameters were determined as discussed in Section 2.1 and had values $\alpha = 1$, $\mu = 20$, $\tau = 0.04$, $\gamma = 2$, $\delta = 2$. Very little posterior mass is given to numbers of components

k	1	2	3	4	5	6	7	8	9	10
$\pi(k\|x)$	0.000	0.000	0.086	0.293	0.343	0.200	0.064	0.013	0.002	0.000
s.d.	0.000	0.000	0.007	0.006	0.005	0.006	0.003	0.001	0.000	0.000

TABLE 2: Posterior distribution of k, galaxy data set, mixtures of normals model with $Poi(1)$ prior on k, see main text for hyperparameter values. The probability $\pi(k|x)$ is the average of five estimates from independent runs of the allocation sampler, s.d. is the standard deviation of the five estimates.

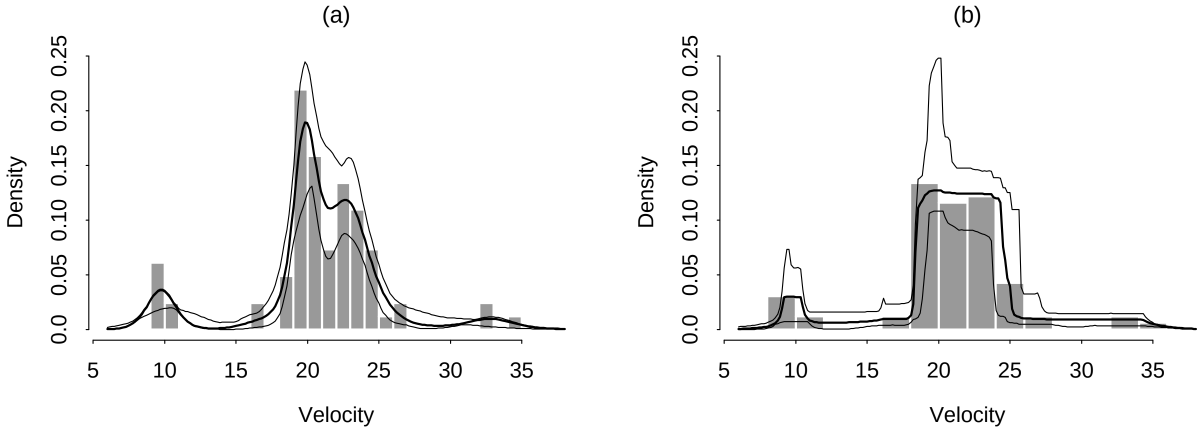

FIGURE 3: Histograms and posterior predictive densities for the galaxy data. Panel (a): mixture of normals model, the thick line is the estimate of the posterior predictive density, the thin lines give 0.005 and 0.995 quantiles of the simulated densities; the histogram bin width is 1. Panel (b): mixture of uniforms model, the lines have same meaning as in panel (a), the histogram bin width is 2.

outside the range from three to seven, in agreement with the estimate based on marginal likelihoods given in Nobile (2005). Figure 3(a) displays, as a thick line overlaid on a histogram, an estimate of the posterior predictive density $p(x_{n+1}|x,\phi)$. Also displayed, as thin lines, are the point-wise 0.005 and 0.995 quantiles of the simulated densities $f(x_{n+1}|k,g,x,\phi)$. Examples of parametric inference for these data will be given in Section 4.6.

4.4 Mixtures of multivariate normals

For multivariate normal components with the prior discussed in Section 2.2, the posteriors $\pi_j(\theta_j|x^j,\phi_j)$ are: $r_j|g,x \sim W_b(\nu'_j,\xi'_j)$ and $m_j|r_j,g,x \sim N_b(\mu'_j,\{\tau'_j r_j\}^{-1})$, where

$$\nu'_j = \nu_j + n_j, \qquad \xi'_j = \xi_j + \sum_{i\in A_j}(x_i-\overline{x}_j)(x_i-\overline{x}_j)^\top + \frac{\tau_j n_j}{\tau_j+n_j}(\mu_j-\overline{x}_j)(\mu_j-\overline{x}_j)^\top,$$

$$\tau'_j = \tau_j + n_j, \qquad \mu'_j = \frac{\tau_j\mu_j + n_j\overline{x}_j}{\tau_j+n_j}.$$

The posterior predictives $p_j(x_{n+1}|x^j,\phi_j)$ are b-variate t distributions with $\nu'_j - b + 1$ degrees of freedom, location vectors μ'_j and precision matrix $\{\tau'_j/(1+\tau'_j)\}(\nu'_j - b + 1)\xi'^{-1}_j$.

As an illustration, we fit a mixture of multivariate normals to Fisher's iris data. The data consists of measurements in centimetres on four variables (sepal length and width, petal length and width) for 50 flowers from each of three species of iris (I. Setosa, I. Versicolor and I. Virginica). Bivariate scatterplots of the data are displayed in Figure 4. Although the species of each flower

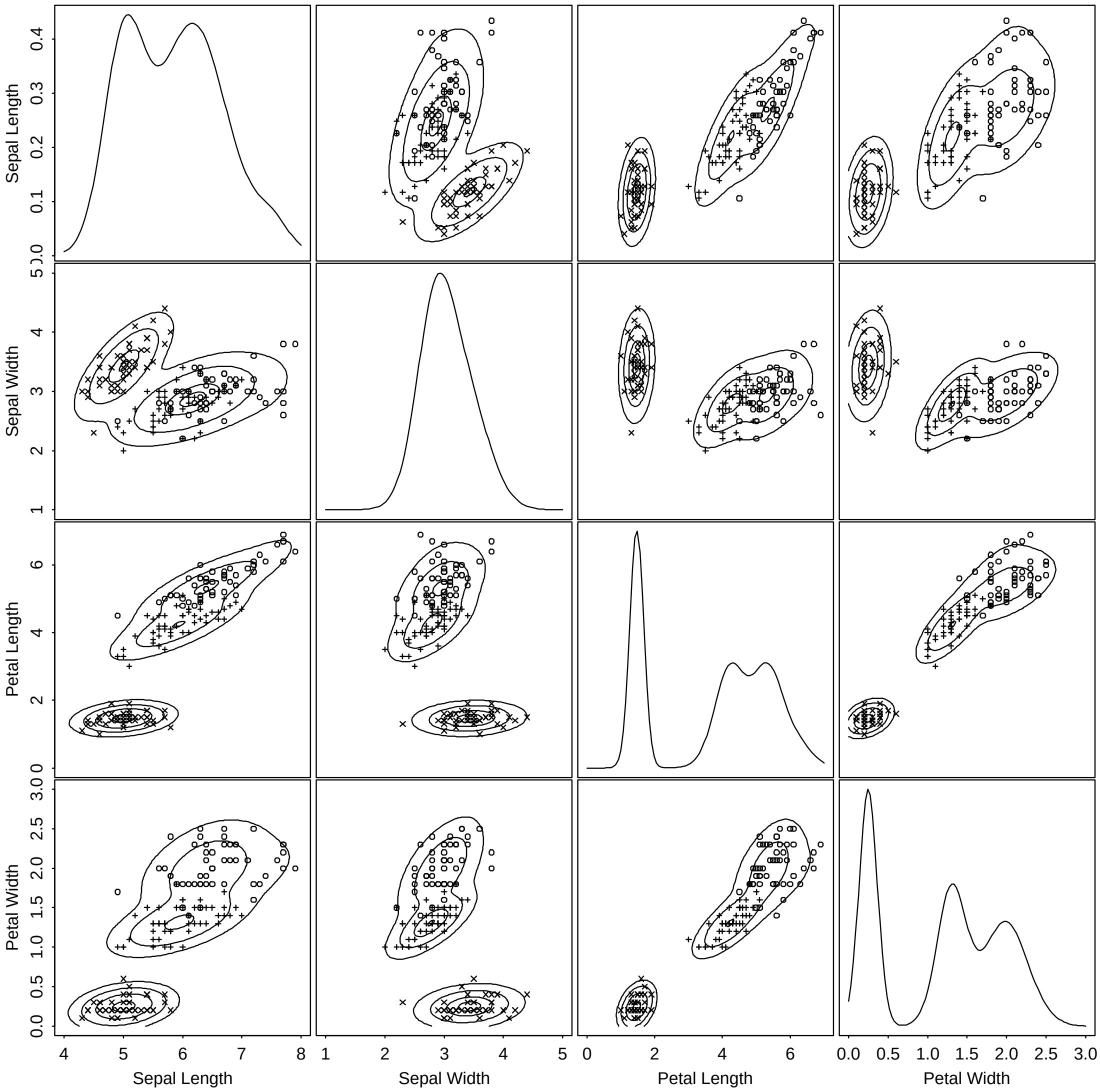

FIGURE 4: Posterior predictive distribution, iris data set. Univariate margins (on main diagonal) and bivariate margins (off-diagonal) of an estimate of the four-dimensional posterior predictive density. In the bivariate plots, the contour lines are drawn at levels corresponding to 5%, 25%, 75% and 95% of the posterior probability of the displayed region. Overlaid on the contour plots are bivariate scatterplots of the data, using the symbols $\times$ = Setosa, $+$ = Versicolor, $\circ$ = Virginica.

is known, this information was not used in fitting the model. Table 3 contains an estimate of the posterior of the number components. The hyperparameter values used were $\alpha = 1$, $\mu = (5.84, 3.06, 3.76, 1.20)^\top$, $\tau = 0.065$, $\nu = 7$ and $\xi = \mathrm{diag}(0.55, 0.4, 0.35, 0.1)$. Most of the posterior

k	1	2	3	4	5	6
$\pi(k\|x)$	0.000	0.000	0.598	0.380	0.022	0.001
s.d.	0.000	0.000	0.031	0.030	0.002	0.000

TABLE 3: Posterior distribution of k, iris data set, mixtures of multivariate normals model with $Poi(1)$ prior on k, see main text for hyperparameter values. The probability $\pi(k|x)$ is the average of five estimates from independent runs of the allocation sampler, s.d. is the standard deviation of the five estimates.

mass is assigned to $k = 3$ and $k = 4$, with the actual number of species accounting for about 60% of the total mass. The posterior predictive density $p(x_{n+1}|x, \phi)$ of the iris data is four-dimensional, Figure 4 displays the univariate and bivariate margins of an estimate computed as detailed in Section 4.2. Within-sample classification probabilities are readily computed using the sample of g vectors from the allocation sampler. We condition on the modal number of components $k = 3$ and, after performing the label permutation procedure discussed in Section 4.6, we plot in Figure 5 the relative frequency with which each observation was allocated to components 1, 2 and 3 in the course of the simulation run. From the plot it is apparent that all but six observations were most often allocated to their correct class. Other summaries, such as the posterior probability that any given group of observations are allocated to the same component, can also be easily computed from the output of the allocation sampler.

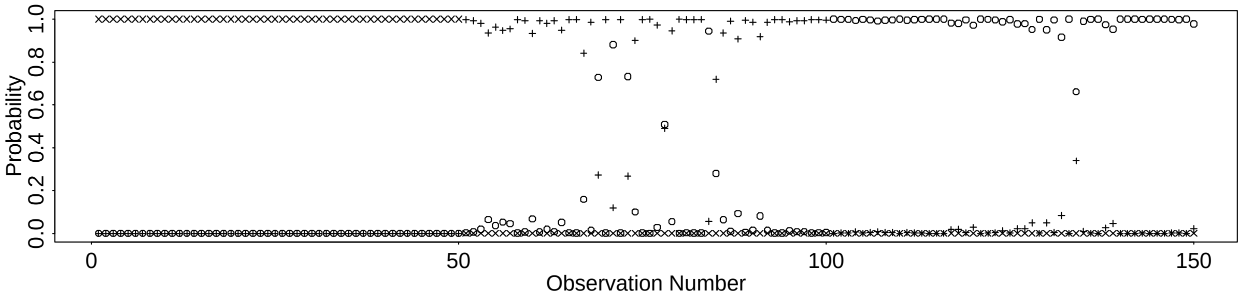

FIGURE 5: Within-sample classification probabilities, iris data set.

4.5 Mixtures of uniforms

Because of the lack of identifiability mentioned at the beginning of Section 2.3, parametric inference in a mixture of uniforms is, *per se*, groundless. We derive the posterior distributions of the parameters only as a means to produce the posterior predictive distribution of a future observation. With the prior and notation as in Section 2.3, the posteriors $\pi_j(\theta_j|x^j, \phi_j)$ are obtained by substituting in (18) the relevant quantities:

$$\pi_j(\theta_j|x^j, \phi_j) = \frac{1}{p_j(x^j|\phi)} \frac{1}{2\phi^2(b_j - a_j)^{n_j}} \qquad -\phi < a_j < x^j_{(1)}, \quad x^j_{(n_j)} < b_j < \phi$$

where $p_j(x^j|\phi)$ is given in equation (9). The posterior predictives $p_j(x_{n+1}|x^j, \phi_j)$ are given by (21), with both the numerator and denominator evaluated according to (9).

To provide an example, we return to the galaxy data and now fit a mixture of uniforms model, with $\phi = 40$. An estimate of the posterior predictive density $p(x_{n+1}|x, \phi)$ is reported in panel (b) of Figure 3, overlaid on a histogram of the data. Comparing the two posterior predictive densities in panels (a) and (b) of Figure 3, one can see that the mixture of uniforms model gives a coarser representation of the data than the normal mixture. At least in part, this is due to the fact that, in the mixture of uniforms, the posterior of the number of components favours smaller values of k, with $\pi(k = 2 \text{ or } 3|x) \approx 0.85$. Of course, the histograms in Figure 3 were matched to the two predictive densities on purpose. We should also remark that histogram representations of even major features of the data, such as the presence of one or two modes in the large central cluster, are affected not only by the bin width, but also by a simple shift of the bins by a quarter or half of the width.

4.6 Identifiability and label switching

Finite mixture distributions are not identifiable: the likelihood (1) is invariant to permutations of the component labels. This lack of identifiability is so basic that a less stringent definition, usually termed "mixture identifiability", has become necessary to denote mixtures that are uniquely decomposable in terms of their components, up to a permutation of the labels, see Titterington *et al.* (1985, page 36). If one is only interested in predictive inference, lack of identifiability is of no concern, as one can evince from the examples in the preceding sections. However, classification and parametric inference require an unequivocal assignment of labels to the mixture components. To show why, let us assume that the prior distribution does not contain any information distinguishing the components, so that the posterior is also invariant to permutations of the labels. As a consequence, the marginal posterior distributions of, say, the component means in a mixture of normals are all equal. An instance of this can be observed in the plots of the top row of Figure 6, which display the marginal posterior densities, conditional on $k = 3$, of the means m_1, m_2 and m_3 in a normal mixture for the galaxy data. These plots were produced by applying the formulae of Section 4.3 to the raw output of the allocation sampler. Since we know the three densities to be equal, it is reassuring to observe that the three plots display similar features. This occurs because, throughout the simulation run, mixture components swap labels relatively often, the label switching phenomenon. Thus, label switching is simply a consequence of the symmetry or near-symmetry of the mixture posterior, coupled with good mixing behaviour of the MCMC sampler. Nonetheless, from an inferential point of view, label switching is problematic since it precludes meaningful statements to be made about each component. One way to achieve identifiability consists in imposing

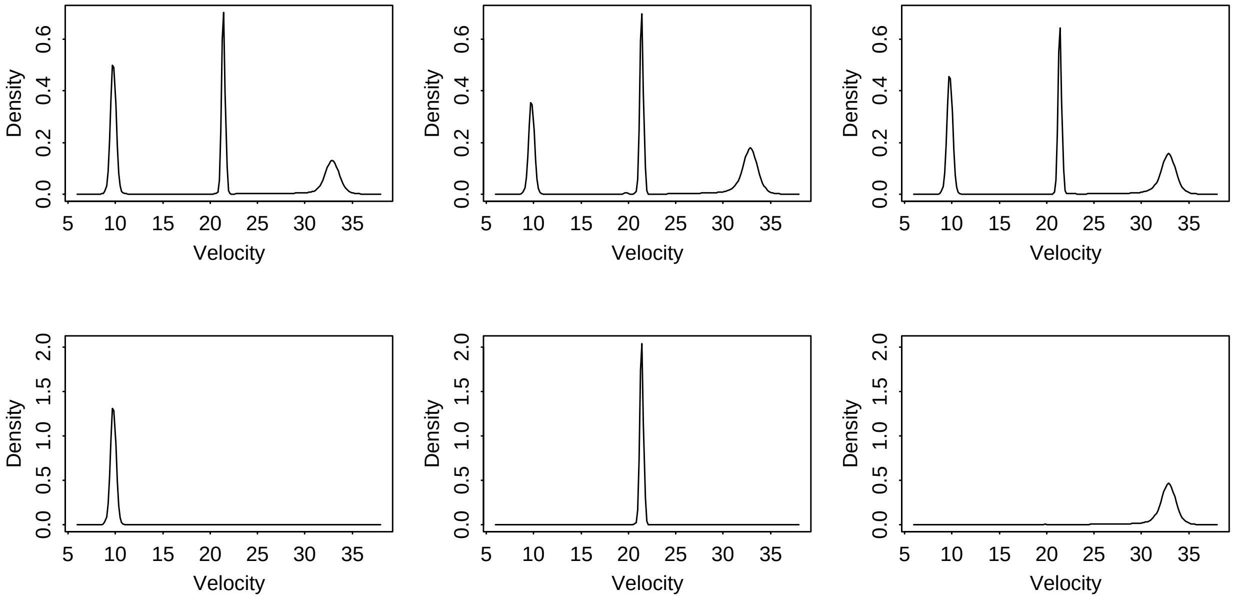

FIGURE 6: Galaxy data, marginal posterior distributions of the component means m_1, m_2 and m_3 in the mixture of normals model, conditional on $k = 3$. The top row of plots displays estimates of the posteriors based on the raw sample of g vectors from the allocation sampler. Bottom row contains estimates using the allocation vectors with re-assigned labels.

constraints on the model parameters, either on the θ's, the λ's or both. However, this approach does not always works satisfactorily and the results may depend on the type of constraint chosen, so other methods have been proposed. We refer to Richardson and Green (1997), Celeux *et al.* (2000), Stephens (2000b) and Frühwirth-Schnatter (2001) for further discussion and additional references.

The method we propose in this section fits in the general framework provided by Stephens (2000b). Our setting is slightly different from the ones examined in the papers mentioned above, since the parameters are not part of the state space of the allocation sampler. Nevertheless, lack of identifiability and associated label switching persist: a prior invariant to permutations of the labels, coupled with the likelihood (1), yields densities $f(g|k)$ in (5) and $f(x|k,g,\phi)$ in (6) that are invariant to permutations of the labels in the allocation vector g.

Let π denote a permutation of the integers $1,\ldots,k$ and let πg be the allocation vector obtained by applying the permutation π to the labels of g: $\pi g = (\pi_{g_1}, \pi_{g_2}, \ldots, \pi_{g_n})$. Define a distance between two allocations g and g' as the number of coordinates where they differ:

$$D(g,g') = \sum_{i=1}^{n} I\left\{g_i \neq g'_i\right\}.$$

Let $S = \{g^{(t)}, t = 1,\ldots,N\}$ be the sequence of sampled allocation vectors. The aim is to minimise the sum of all distances between allocations

$$\sum_{t=1}^{N-1} \sum_{s=1}^{N} D(\pi^{(t)} g^{(t)}, \pi^{(s)} g^{(s)})$$

with respect to the sequence of permutations $\{\pi^{(t)}, t = 1,\ldots,N\}$. An approximate solution to this problem can be obtained by replacing it with a sequence of optimisation problems, each involving a single permutation $\pi^{(t)}$. These simpler problems are instances of the square assignment problem, for which we used the algorithm and publicly available code of Carpaneto and Toth (1980). Minor implementation details apart, the allocations $g^{(t)}$ in S are processed in the order of increasing number $\tilde{k}^{(t)}$ of non-empty components. When re-assigning the labels of $g^{(t)}$, the vector is compared to the set $B^{(t)}$ of allocations $g \in S$ that have already been processed and that have $\tilde{k}^{(t)}$ or $\tilde{k}^{(t)} - 1$ non-empty components. A cost matrix is computed with generic element

$$C(j_1, j_2) = \sum_{g \in B^{(t)}} \sum_{i=1}^{n} I\{g_i \neq j_1, g_i^{(t)} = j_2\}$$

and the square assignment algorithm then returns the permutation $\pi^{(t)}$ which minimises the total cost $\sum_{j=1}^{\tilde{k}^{(t)}} C(j, \pi_j^{(t)})$. The allocation vector with re-assigned labels is then equal to $\pi^{(t)} g^{(t)}$. The plots in the bottom panels of Figure 6 were produced using the allocations with re-assigned labels, in that example our procedure manages to isolate the three components very well. For the galaxy data, post-processing 10,000 allocation vectors took about 9 minutes, compared to about 4 minutes for the actual sampling, using a thinning constant $\Delta = 100$.

As another example, we consider the joint marginal posterior distributions, conditional on $k = 5$, of means and standard deviations $(m_j, r_j^{-1/2})$, $j = 1, \ldots, 5$, in a normal mixture for the galaxy data. The panels in the top row of Figure 7 contain the densities computed using the raw output of the allocations sampler. Here too, the plots are very similar, due to label switching, and show five modes, some more defined than others. The plots in the bottom row use instead the post-processed allocations. In this example our procedure manages to isolate very well the first four components, with means at about 10, 20, 23 and 33. The fifth component, which according to the plots in the top row is much less defined, is displayed in the plot in the bottom right corned. This posterior has a major mode corresponding to a mean of about 17, however, minor modes with means of 23

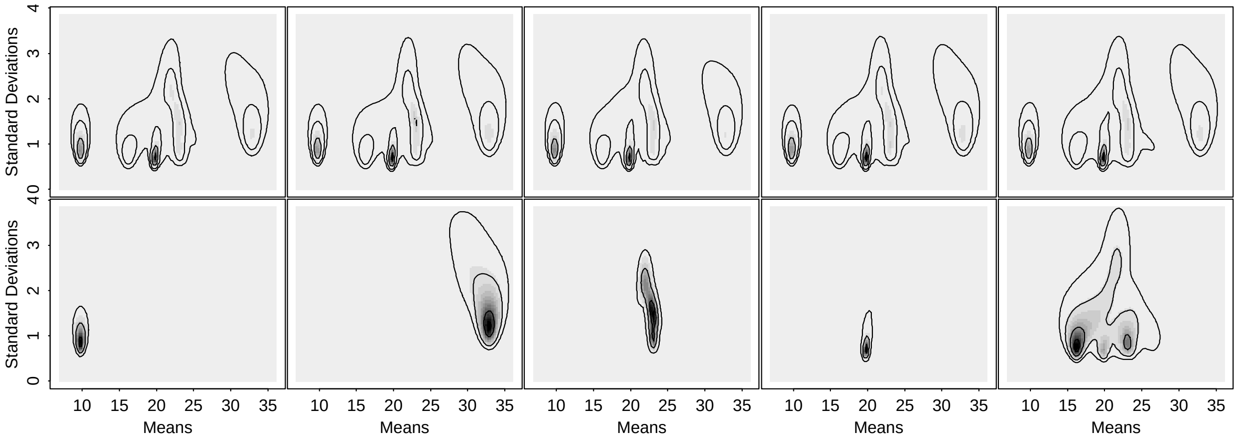

FIGURE 7: Galaxy data, joint marginal posterior distributions of component means and standard deviations $(m_j, r_j^{-1/2})$, $j = 1, \ldots, 5$, in the mixture of normals model, conditional on $k = 5$. The top row of plots displays estimates of the posteriors based on the raw sample of g vectors from the allocation sampler. Bottom row contains estimates using the allocation vectors with re-assigned labels. The contour lines are drawn at levels corresponding to 5%, 25%, 75% and 95% of the posterior probability of the displayed region.

and 20 are still clearly visible.

5 Conclusions

We have proposed a Markov chain Monte Carlo method for the analysis of finite mixtures with an unknown number of components. The method can be applied, essentially unchanged, to any parametric family of mixture components, provided that the mixture weights and component parameters can be integrated out of the model in closed form. This condition is of course satisfied if natural conjugate priors are used, but it also holds in other cases. The resulting sampler has a state space which comprises, besides the number of components k, only the allocation vector g, for this reason we have called it the "allocation sampler". The sampler contains moves that keep k fixed and moves that change it, and it can be thought of as a version of the reversible jump algorithm. We have explicitly considered normal mixtures, mixtures of multivariate normals and mixtures of uniform distributions and have illustrated the performance of the allocation sampler using artificial data and two real data sets. Posterior distributions of the parameters and of a future observation have been derived, estimates based on the sampled allocations are easily computable. We have also discussed a post-processing technique that re-assigns the labels in the allocations, to overcome a basic lack of identifiability in the finite mixture model.

Acknowledgements

Alastair Fearnside acknowledges financial support under a doctoral training grant from the Engineering and Physical Sciences Research Council.

Appendix

A.1 Proposal ratio for first M-H move in Section 3.1.2

The probability of selecting the candidate allocation g', after integrating with respect to the distribution of p_1, is

$$P(g \to g') = \frac{\Gamma(\alpha_{j_1} + \alpha_{j_2})}{\Gamma(\alpha_{j_1})\Gamma(\alpha_{j_2})} \frac{\Gamma(\alpha_{j_1} + \tilde{n}_{j_1})\Gamma(\alpha_{j_2} + \tilde{n}_{j_2})}{\Gamma(\alpha_{j_1} + \alpha_{j_2} + n_{j_1} + n_{j_2})}$$

where $\tilde{n}_{j_1}$, $\tilde{n}_{j_2}$ are the numbers of observations re-allocated to components j_1 and j_2. Therefore, the ratio of proposal probabilities in (10) is

$$\frac{P(g' \to g)}{P(g \to g')} = \frac{\Gamma(\alpha_{j_1} + n_{j_1})\Gamma(\alpha_{j_2} + n_{j_2})}{\Gamma(\alpha_{j_1} + \tilde{n}_{j_1})\Gamma(\alpha_{j_2} + \tilde{n}_{j_2})} = \frac{f(g|k)}{f(g'|k)}.$$

A.2 Distribution of p_E in Section 3.2

We use $p_E \sim Beta(a, a)$ and require a to be such that $\Pr[\tilde{n}_{j_2} = 0] = p_0/2$. Thus, p_0 is the probability of proposing to eject either an empty component or a component that contains all the observations currently in the ejecting component. In our experiments, setting $p_0 = 0.2$ worked well. Then

$$\begin{aligned} \frac{p_0}{2} &= \int_0^1 (1 - p_E)^{n_{j_1}} \frac{\Gamma(2a)}{\Gamma(a)\Gamma(a)} p_E^{a-1} (1 - p_E)^{a-1}\, dp_E \\ &= \frac{\Gamma(2a)}{\Gamma(a)\Gamma(a)} \frac{\Gamma(a)\Gamma(a + n_{j_1})}{\Gamma(2a + n_{j_1})}. \end{aligned}$$

resulting in the equation

$$\frac{\Gamma(2a)}{\Gamma(a)} \frac{\Gamma(a + n_{j_1})}{\Gamma(2a + n_{j_1})} = \frac{p_0}{2}. \tag{22}$$

The left hand side equals (1/2) times a product of $n_{j_1} - 1$ terms, each monotonically decreasing in a, hence it is monotonic decreasing in a. Therefore (22) can be easily solved numerically, e.g. using bisection, apart for few small values of n_{j_1} if p_0 is too small. Since solving (22) numerically is relatively time consuming, the equation was solved only for n_{j_1} in a grid of values, equispaced on the log-scale. These solutions were then stored in the simulation program and, at each stage, the appropriate value of $\log a$ for the current value of $\log n_{j_1}$ was determined using linear interpolation of the solutions for the closest n_{j_1}'s in the grid.

References

Aitkin, M. (2001). Likelihood and Bayesian analysis of mixtures. *Statistical Modelling*, **1**, 287–304.

Böhning, D. and Seidel, W. (2003). Editorial: recent developments in mixture models. *Computational Statistics and Data Analysis*, **41**, 349–357.

Casella, G., Robert, C. P. and Wells, M. T. (2000). Mixture Models, Latent Variables and Partitioned Importance Sampling. Tech Report 2000-03, CREST, INSEE, Paris.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, **57**, 473–484.

Carpaneto, G. and Toth, P. (1980). Algorithm 548: Solution of the Assignment Problem [H]. *ACM Transactions on Mathematical Software*, **6**, 104–111.

Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association*, **95**, 957–970.

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

Dellaportas, P. and Papageorgiou I. (2003). Multivariate mixtures of normals with unknown number of components. Tech Report, Dept. of Statistics, Athens U. of Economics and Business.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1–38.

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

Frühwirth-Schnatter, S. (2001). Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association*, **96**, 194–209.

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

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**, 1316–1332.

Marin, J.-M., Mengersen, K. and Robert, C.P. (2005). Bayesian Modelling and Inference on Mixtures of Distributions. *Handbook of Statistics*, **25**, (eds D. Dey and C.R. Rao). Elsevier-Sciences (to appear)

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, John Wiley & Sons.

Mengersen, K. L. and Robert, C. P. (1996). Testing for Mixtures: a Bayesian Entropic Approach. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith), 255–276, Oxford University Press.

Newcomb, S. (1886). A Generalized Theory of the Combination of Observations so as to Obtain the Best Result. *American Journal of Mathematics*, **8**, 343–366.

Nobile, A. (1994). Bayesian Analysis of Finite Mixture Distributions, Ph.D. dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh. Available at `http://www.stats.gla.ac.uk/~agostino`

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics*, **32**, 2044–2073.

Nobile, A. (2005). Bayesian finite mixtures: a note on prior specification and posterior computation. Technical Report 05-3, Department of Statistics, University of Glasgow.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. A*, **185**, 71–110.

Phillips, D. B. and Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 215–239, Chapman & Hall.

Raftery, A. E. (1996). Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 163–187, Chapman & Hall.

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

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association*, **85**, 617–624.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Steele, R. J., Raftery, A. E. and Emond, M. J. (2003). Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS). Tech Report 436, Dept of Statistics, U. of Washington.

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

Stephens, M. (2000b). Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society B*, **62**, 795–809.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons.