Stat Comput (2007) 17:147–162
DOI 10.1007/s11222-006-9014-7

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

Agostino Nobile · Alastair T. Fearnside

Received: August 2005 / Accepted: December 2006 / Published online: 7 February 2007

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 and of multivariate normals are explicitly considered. The problem of label switching, when parameter inference is of interest, is addressed in a post-processing stage.

Keywords Classification · Galaxy data · Iris data · Label switching · Markov chain Monte Carlo · Multivariate normal mixtures · Normal mixtures · Reversible jump

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 the end of the 19-th century, 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 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 (MCMC) 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. 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; 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) have avoided placing a prior distribution on k, instead, they have estimated the marginal likelihoods of k components and used Bayes factors to test k vs. $k+1$ components. Mengersen and Robert (1996) have employed a testing approach too, but relying on the Kullback–Leibler divergence as a measure of distance between mixtures with k and $k+1$ components. Representations of the

A. Nobile (✉) · A. T. Fearnside
Department of Statistics, University of Glasgow,
Glasgow G12 8QW, U.K
e-mail: agostino@stats.gla.ac.uk

marginal likelihoods for k components have been derived by Nobile (1994, 2004) and Ishwaran et al. (2001). Much research on Bayesian mixtures has followed a different approach, accommodating an unknown number of components by means of Ferguson's Dirichlet Process. Here, we only mention two recent papers by Fearnhead (2004) and by Jain and Neal (2004) and refer to them for additional details and references.

Most authors working with Bayesian finite mixtures 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), and 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 scheme, *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, though see Zhang et al. (2004) and Dellaportas and Papageorgiou (2006). 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 one to tackle the label switching problem. For 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 $\sum_{j=1}^{k} \lambda_j = 1$. The number of components k, the mixture weights $\lambda = (\lambda_1, \ldots, \lambda_k)$ and the components' parameters $\theta = (\theta_1, \ldots, \theta_k)$ are all regarded 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. 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: $n_j = \mathrm{card}\{A_j\}$ 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 Eq. (3) with respect to the conditional distribution of g given in (2) produces the finite mixture (1).

Next we consider the prior distribution on k, λ and θ. We use as prior on k the $Poi(1)$ distribution restricted to $1 \le k \le k_{\max}$; 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 $\phi = \{\phi_1, \ldots, \phi_k\}$.

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 such 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 $A_j=\emptyset$.

Some remarks about the hyperparameters (α_j,ϕ_j), $j=1,\dots,k$ are necessary. The hyperparameters of the j-th component are the same for all values of $k\geq j$, so that the family of mixture models is nested and the corresponding marginal likelihoods are related (see Nobile, 2004). When $\alpha_j=\alpha_1$, $\phi_j=\phi_1$, $j=1,\dots,k$, the prior is symmetric with respect to a permutation of the components' labels. We call this the *symmetric case* and refer to its opposite as the *asymmetric case*. Clearly, if any information distinguishing the components is available, it should be incorporated in the prior. We only provide some suggestions for prior specification in the symmetric case. The weights' hyperparameters $\alpha_1,\dots,\alpha_k$ are treated as fixed constants, we adopt the common choice $\alpha_j=\alpha_1=1$. The posterior distribution of the number of components k may be very sensitive to changes in the hyperparameters ϕ; an example of marked sensitivity is reported in Jasra et al. (2005). Another is in Nobile (2005), who proposes an empirical Bayes approach, where some components of the ϕ_j's are fixed using basic features of the data, while others are estimated based on a preliminary MCMC run conducted with a hyperprior on ϕ. That approach is adopted and extended in the present paper; details are in Section 2.1.

The allocation sampler, discussed in Section 3, makes draws from the joint posterior 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 illustrating the sampler, we show how the model specialises to the case of (multivariate) normal components.

2.1 Mixtures of multivariate normals

If the mixture components are multivariate normals of dimension b, then $q_{g_i}(x_i|\theta_{g_i})$ in (3) is the density $N_b(x_i|m_{g_i},r_{g_i}^{-1})$ of a b-variate normal with mean vector m_{g_i} and precision (inverse covariance) matrix r_{g_i}. The priors $\pi_j(\theta_j|\phi_j)$ in (4) are the usual conjugate priors for (m_j,r_j): $r_j\sim W_b(\nu_j,\xi_j)$, a Wishart distribution with ν_j degrees of freedom and precision matrix ξ_j, and $m_j|r_j\sim N_b(\mu_j,\{\tau_j r_j\}^{-1})$, with μ_j a b-vector and τ_j a positive real number. The hyperparameters for component j are $\phi_j=\{\mu_j,\tau_j,\nu_j,\xi_j\}$. Given k and g, the marginal density of the data allocated to the j-th component is given by

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

where $\overline{x}_j=(1/n_j)\sum_{i\in A_j}x_i$ is the sample mean vector of the observations allocated to component j. We assume a symmetric prior and set the overall mean vector $\mu_1=\overline{x}$, the sample mean vector. The prior predictive distribution is multivariate t with ν_1-b+1 degrees of freedom, which we set equal to 4, to have a prior predictive with relatively thick tails, but finite second moments. This yields $\nu_1=b+3$. For the remaining hyperparameters τ_1 and ξ_1, we assume that ξ_1 is diagonal and then use independent priors. More specifically, $(1+\tau_1)^{-1}\sim Un(0,1)$ and $\xi_{1s}\sim Un(0,(\nu-2)v_s)$, $s=1,\dots,b$, where v_s is the sample variance of the s-th variable. Draws from a preliminary run of the sampler are used to make boxplots of the marginal posterior distributions of τ_1 and ξ_1 conditional on k. Estimates $\hat{\tau}_1$ and $\hat{\xi}_1$ 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 ξ_1 in subsequent runs.

When $b=1$ one obtains mixtures of univariate normals. In that case, the Wishart prior on r_j reduces to the usual Gamma prior $r_j\sim Ga(\nu_j/2,\xi_j/2)$.

3 The allocation sampler

This section discusses how to sample from the joint posterior distribution of k and g given in (8). The sampler 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) three 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, as well as g. The second type of moves consists of an absorption/ejection 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 k

The first move is a systematic sweep Gibbs sampler (GS) on the components of g, from g_1 to g_n. To sample g_i we need its full conditional distribution. This is computed by first evaluating k times the joint density $f(k, g, x|\phi)$, with $g_i = 1, \dots, k$ 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 GS move 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, there is scope for moves that attempt to change several entries of g simultaneously; we discuss three such Metropolis-Hastings moves next.

3.1.1 Metropolis-Hastings moves on g

In the first move (M1), two components, j_1 and j_2 say, are randomly selected among the k available. A draw p_1 is made from the *Beta*$(\alpha_{j_1}, \alpha_{j_2})$ distribution, and 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

$$R = \frac{f(k, g', x|\phi)}{f(k, g, x|\phi)} \frac{P(g' \to g)}{P(g \to g')} \tag{9}$$

and $P(g \to g')$ is the probability of proposing the candidate g' when the current state is g. One can show, see Appendix A.1, that $P(g' \to g)/P(g \to g') = f(g|k)/f(g'|k)$, so that R reduces to

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

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

The second move (M2) selects a group of observations currently allocated to component j_1 and attempts to re-allocate them to component j_2. In detail, j_1 and j_2 are randomly drawn from the k components. If $n_{j_1} = 0$ the move fails outright. Otherwise, m is drawn from a discrete uniform distribution on $\{1, \dots, n_{j_1}\}$. Then, m observations are randomly selected among the n_{j_1} in component j_1 and moved to component j_2. This results in a candidate g' which is accepted with probability $\min\{1, R\}$, where R is given by (9). In this move, however, the proposal ratio can be easily shown to be

$$\frac{P(g' \to g)}{P(g \to g')} = \frac{n_{j_1}}{n_{j_2} + m} \frac{n_{j_1}!\, n_{j_2}!}{(n_{j_1} - m)!\, (n_{j_2} + m)!}.$$

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

The third move (M3) is similar to M1, in that two components j_1 and j_2 are randomly selected and a candidate g' is formed by re-allocating the observations currently assigned to j_1 and j_2. However, in M3 the probabilities $p_j^{(i)}$, $j \in \{j_1, j_2\}$, of re-allocating the i-th observation to component j are not constant, as in move M1. Instead, we let the $p_j^{(i)}$'s be proportional to the probabilities that component j generated the i-th observation, conditional on its value x_i, on the previously re-allocated observations and on their new allocations. The resulting candidate allocation g' is accepted with probability $\min\{1, R\}$, where R is given in formula (9), but with the proposal ratio now equal to

$$\frac{P(g' \to g)}{P(g \to g')} = \prod_{i \in A_{j_1} \cup A_{j_2}} \frac{p_{g_i}^{(i)}}{p_{g_i'}^{(i)}}. \tag{11}$$

A more detailed description of move M3 can be found in Appendix A.2.

3.1.2 Metropolis-Hastings move on the labels

For a given k and g, let $\tilde{k}$ be the number of non-empty components in g. Then, there are $\binom{k}{\tilde{k}}\tilde{k}!$ other allocation vectors which partition the data x in the same way as g, only differing in the assignment of labels to the groups. Thus to each g with high posterior probability there correspond

$k!/(k-\tilde{k})!$ other g's also relatively likely (in the symmetric case, equally likely) a posteriori. The Gibbs and Metropolis-Hastings moves described above only move slowly from one region of high posterior probability to another. This creates a problem in the asymmetric case, since 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 move only in the asymmetric case. For parametric inference in the symmetric case, we reassign the labels in a post-processing stage, see Section 4.4.

3.2 Moves that change k

The moves that change the number of components consist of a Metropolis-Hastings pair of absorption/ejection (AE) moves: a component ejects another component and, in the reverse move, a component absorbs another component. Assume that the current state is $\{k,g\}$; an ejection move is attempted with probability p_k^e, where $p_k^e = 1/2$, $k = 2,\dots,k_{\max}-1$, $p_1^e = 1$ 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 with 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 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 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 Appendix A.3 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 reallocated to component j_1. Hence the proposal probability is

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

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

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

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 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 in (15).

.

3.3 Some examples with artificial data

To demonstrate the allocation sampler, we applied it to samples of sizes 50, 200, 500 and 2000 from a few mixtures of univariate normals, displayed in Fig. 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 ξ, consisting 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

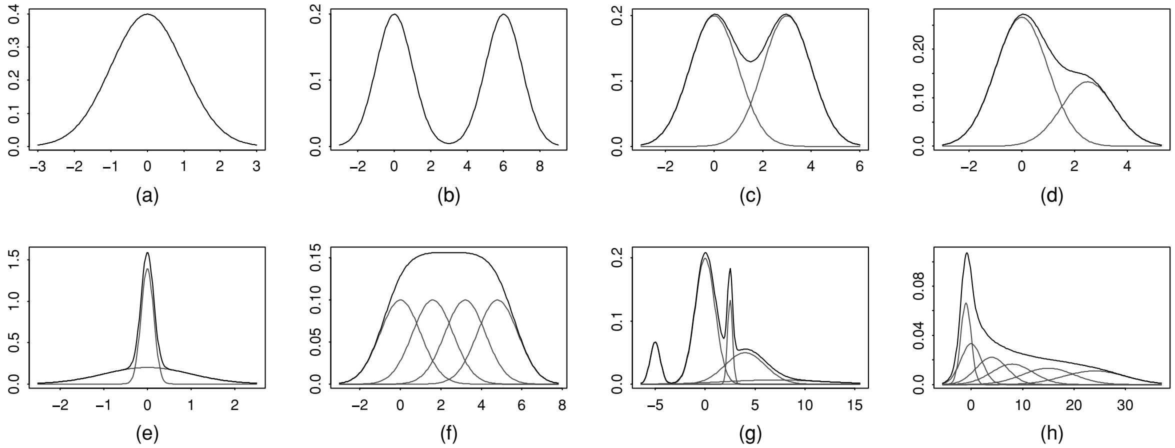

Fig. 1 Density functions of some mixtures of univariate normal distributions. (a) $N(0,1)$, (b) $\frac{1}{2}N(0,1)+\frac{1}{2}N(6,1)$, (c) $\frac{1}{2}N(0,1)+\frac{1}{2}N(3,1)$, (d) $\frac{2}{3}N(0,1)+\frac{1}{3}N(2.5,1)$, (e) $\frac{1}{2}N(0,1)+\frac{1}{2}N(0,7^{-2})$, (f) $\sum_{j=1}^{4}\frac{1}{4}N(1.6\{j-1\},1)$,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

run was used to estimate τ and ξ 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 data set; summaries of the estimated posteriors of k are reported in Table 1. For all mixtures, $\pi(k_{\text{true}}|x)$ increases with the sample size n; also, the mode of $\pi(k|x)$ tends to move towards k_{true} as n becomes larger. An idea of the mixing of the allocation sampler, using the above settings, can be obtained from Fig. 2 which displays, for mixture (f) with sample size $n=2000$, a jittered time series plot of k and a plot of the estimate of $\pi(k|x)$ across a simulation run.

As a further example, we generated 200 observations from a six component equally weighted mixture of 10-dimensional multivariate normals. The components' means were $m_{1i}=m_{2i}=m_{6i}=0$, $m_{3i}=2$, $m_{4i}=-2$, $m_{5i}=(-1)^i$, $i=1,\dots,10$. As for the covariance matrices, we used $r_1^{-1}=r_3^{-1}=r_4^{-1}=r_5^{-1}=I$, $r_2^{-1}=0.25I$ and $(r_6^{-1})_{ij}=0.9^{|i-j|}$, $i,j=1,\dots,10$, where I denotes the 10-dimensional identity matrix. The estimated posterior distribution of k is reported in Table 2. Almost all the posterior mass is concentrated on values of k between 5 and 7.

3.4 Assessing the performance of the sampler

Sections 3.1 and 3.2 describe the moves which comprise the allocation sampler. One may wonder what is gained by having, besides the AE move on k and g, four additional moves on g alone: GS, M1, M2 and M3. To address this question, we reconsidered the 10-dimensional multivariate normals example and made five short runs with samplers using different combinations of the moves, see Table 3. In all samplers except the first one, move AE had selection probability equal to 0.5, with the remaining probability equally distributed among the other moves. The allocation sampler used throughtout the paper is AE + GS + M1 + M2 + M3. Each run was $10,000\ \Delta$ sweeps long, with no burn-in, the thinning values Δ chosen to yield approximately equal run times. Figure 3 contains time series plots of the 10,000 draws of k kept in each run, along with plots of the estimated autocorrelation functions computed on the last 5,000

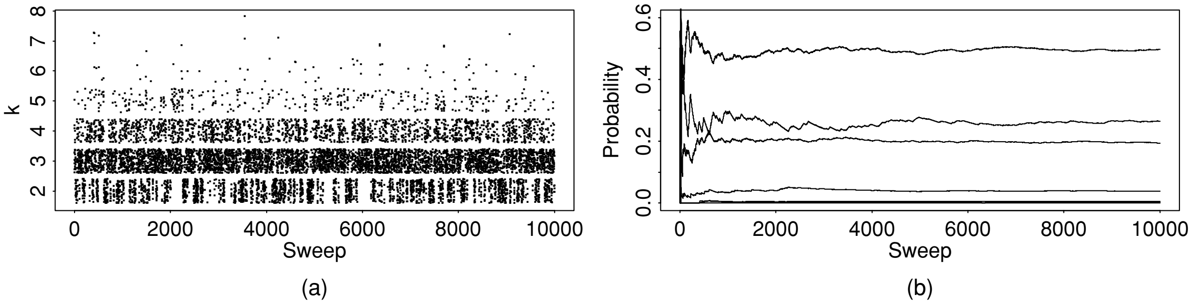

Fig. 2 Sampled values of k for mixture (f) in Fig. 1 with sample size $n=2000$, in the first of five independent runs: (a) jittered time series plot of k; (b) Estimate of $\pi(k|x)$ across the simulation run. The number of sweeps on the x-axis should be multiplied by the thinning constant, in this instance $\Delta=230$

Table 1 Posterior probabilities of selected values of k for representative samples of sizes 50, 200, 500 and 2000 from the mixtures displayed in Fig. 1

		Sample size			
		50	200	500	2000
Mixture	k	$\pi(k\|x)$	$\pi(k\|x)$	$\pi(k\|x)$	$\pi(k\|x)$
(a) k_{true} : 1	1	$\boxed{.558}$	$\boxed{.648}$	$\boxed{.742}$	$\boxed{.819}$
	2	.300	.253	.200	.150
	3	.107	.078	.048	.027
(b) k_{true} : 2	1	.000	.000	.000	.000
	2	$\boxed{.753}$	$\boxed{.820}$	$\boxed{.866}$	$\boxed{.890}$
	3	.205	.160	.122	.101
	4	.037	.018	.011	.009
(c) k_{true} : 2	1	.280	.001	.000	.000
	2	$\boxed{.450}$	$\boxed{.629}$	$\boxed{.660}$	$\boxed{.726}$
	3	.202	.277	.262	.225
	4	.055	.077	.064	.041
(d) k_{true} : 2	1	$\boxed{.388}$	.027	.000	.000
	2	.379	$\boxed{.566}$	$\boxed{.614}$	$\boxed{.653}$
	3	.170	.296	.283	.261
	4	.049	.088	.081	.070
(e) k_{true} : 2	1	.000	.000	.000	.000
	2	$\boxed{.595}$	$\boxed{.596}$	$\boxed{.622}$	$\boxed{.690}$
	3	.298	.311	.292	.249
	4	.087	.080	.072	.053
(f) k_{true} : 4	1	$\boxed{.453}$	.075	.000	.000
	2	.344	$\boxed{.561}$	$\boxed{.596}$	.279
	3	.146	.267	.295	$\boxed{.487}$
	4	.044	.078	.087	.184
	5	.010	.016	.018	.042
(g) k_{true} : 5	1	.147	.000	.000	.000
	2	$\boxed{.445}$	.001	.000	.000
	3	.272	.015	.000	.000
	4	.103	$\boxed{.411}$	.040	.000
	5	.026	.377	$\boxed{.526}$	$\boxed{.681}$
	6	.005	.149	.310	.250
	7	.001	.039	.098	.057
(h) k_{true} : 6	1	.000	.000	.000	.000
	2	$\boxed{.530}$	.035	.000	.000
	3	.331	$\boxed{.576}$	$\boxed{.452}$	.016
	4	.112	.288	.392	$\boxed{.665}$
	5	.022	.083	.124	.254
	6	.004	.015	.026	.055

Note: The probabilities $\pi(k|x)$ are averages of five estimates from independent runs of the allocation sampler. All the tabulated values have estimated standard errors, based on the five runs, below 0.01. Modal values are enclosed in a box.

Table 2 Posterior distribution of k, for a sample of 200 from a six component mixture of 10-dimensional multivariate normals

k	4	5	6	7	8
$\pi(k\|x)$	0.004	0.288	0.560	0.135	0.012
s.d.	0.002	0.016	0.015	0.003	0.001

Note: 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. Each run took about 1.5 hours, using $\Delta = 110$.

draws. From the much longer runs reported in Table 2, the region of high posterior probability corresponds to values of k between 5 and 7. All samplers were started at $k = 1$ and, after spending some time around $k = 4$, reached and stayed in that region. The exception is the sampler consisting of move AE alone, which did not converge during the run. Among the other samplers, the a.c.f. of AE + GS decays very slowly; the other three are similar, with the a.c.f. of AE + GS + M1 + M2 + M3 decaying faster. A more formal measure of mixing is the effective sample size, which, for a stationary time series of length N with a.c.f. ρ_j, satisfies $N_{\text{eff}} = N / \sum_{j=-\infty}^{\infty} \rho_j$. Estimates of N_{eff}, corresponding to $N = 10,000$, are reported in Table 3, except for the sampler consisting of AE alone. In agreement with the visual inspection of the a.c.f.'s, they indicate that, in this example, the allocation sampler is, to a varying degree, superior to the other ones.

We consider next the relationship between sample size and the moves' acceptance rates, by examining the runs for the eight mixtures of normals reported in Table 1. Each panel of Fig. 4 displays, on a doubly logarithmic scale, a plot of acceptance rates vs. sample size, for the moves AE, M1, M2, M3. All the acceptance rates tend to fall as the sample size increases, so that longer runs (i.e. higher thinning Δ) are likely needed for larger samples. However, the rates of decrease of the acceptance rates are quite variable, both between moves and between mixtures. We note that mixtures (b), (g) and (h), for which the acceptance rates of M1 and M2 fall fastest, also show relatively stable acceptance rates for M3.

Finally, we look at how sample size affects the run time of our Fortran implementation. A detailed exam, not reported here, of our simulations for the mixtures in Fig. 1, suggests two sources of computational cost. First, the average run times per sweep are roughly proportional to the sample size n. Second, the thinning values Δ also tend to increase with n, but to a varying degree, depending on the mixture: e.g., as n increases from 50 to 2000, Δ goes from 10 to 30 for mixtures (a), (b) and (e), while for mixtures (f), (g) and (h) it increases from 10 to 230, 310 and 210 respectively. Overall run times, on a PC with a 2.2 GHz 64-bit processor, range from a few seconds for all the mixtures in

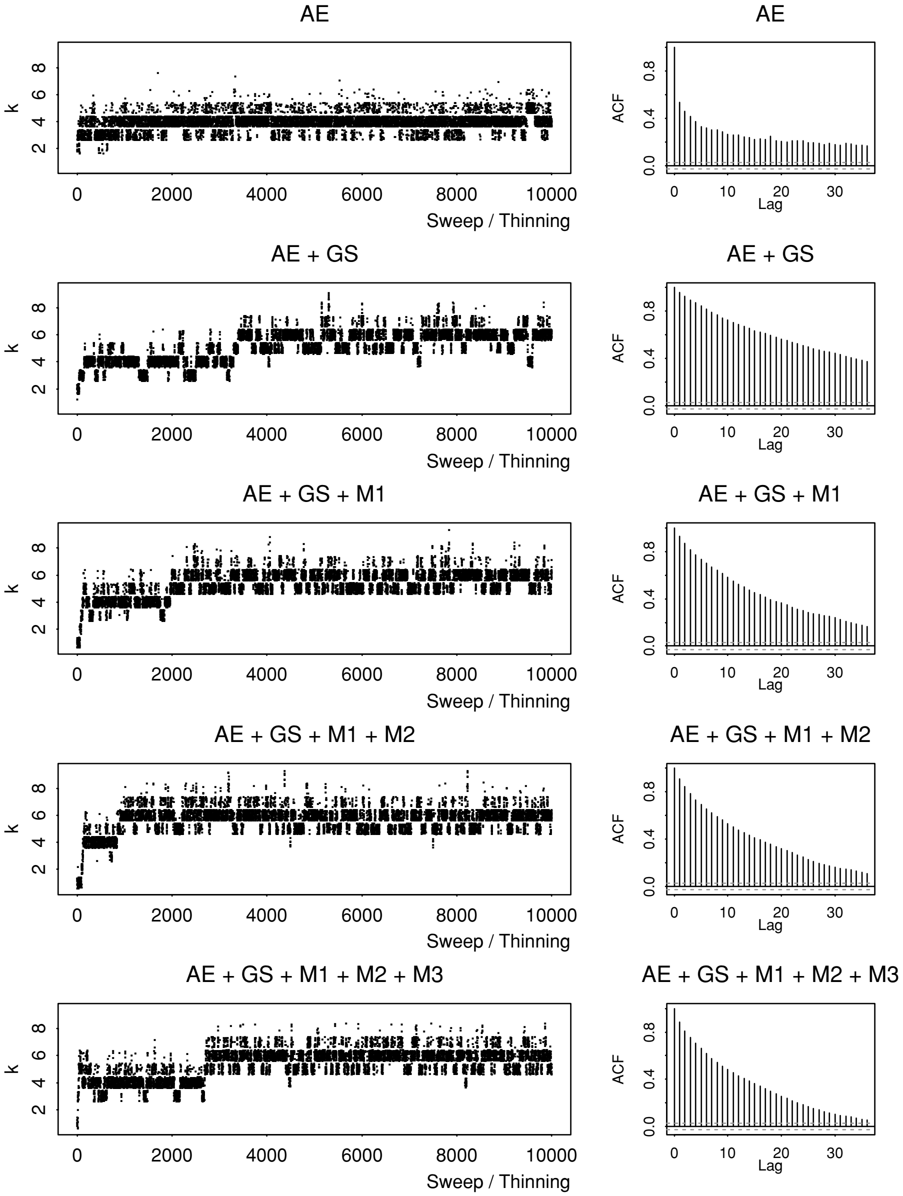

Fig. 3 Comparison of five samplers consisting of different combinations of moves AE, GS, M1, M2, M3 run on the six-component 10-dimensional multivariate normals example. Left hand side panels contain jittered time series plots of draws of k; right hand side panels display plots of corresponding estimates of the a.c.f.'s

Fig. 1 when $n = 50$, to about 4 hours for mixture (g) with $n = 2000$.

As the family of mixture components affects the sampler only through the form of the densities $p_j(x^j|\phi_j)$ in formula (6), we believe that the discussion in this section has general applicability.

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 convenient summaries. Moreover, inference about the components' weights and parameters makes little sense unless it is carried out conditionally on the number of 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:

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

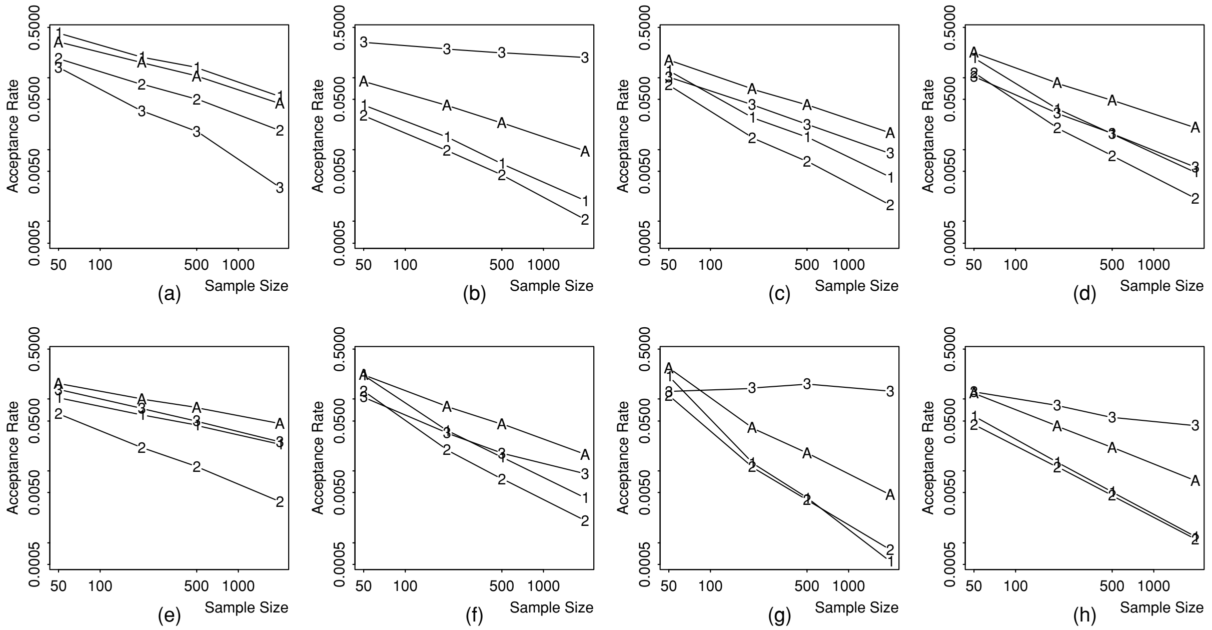

Fig. 4 Plots of acceptance rates of AE move (labelled A) and M1, M2, M3 moves (labelled 1, 2, 3) against sample size, on a doubly logarithmic scale. Each panel shows acceptance rates in runs of the allocation sampler for the mixture in the corresponding panel of Fig. 1

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.

Table 3 Comparison of five samplers run on the six-component 10-dimensional multivariate normals example

Sampler	Δ	Run Time	$\widehat{\sum_j \rho_j}$	$\widehat{N_{\text{eff}}}$
AE	1700	717.8		
AE + GS	5	726.2	73.0	137
AE + GS + M1	10	751.4	39.6	253
AE + GS + M1 + M2	14	751.7	32.8	305
AE + GS + M1 + M2 + M3	16	705.6	26.8	373

Note: Thinning values Δ were set to have approximately equal run times. Times are in seconds on a PC with a 1.8 GHz 64-bit processor. The effective sample size is $N_{\text{eff}}=N/\sum_{j=-\infty}^{\infty}\rho_j$, where ρ_j denotes the lag j autocorrelation of the thinned series of k's. The values in the last two columns use autoregressive estimates of the ρ_j's, based on the second half of the series of k's.

Marginal posterior distributions unconditional on g are obtained by averaging (16) and (17) with respect to the posterior $\pi(g|k,x,\phi)$. Of course, marginally the prior independence is lost.

4.2 Predictive distributions

Conditional on k, λ, θ, g and x, a future observation 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 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 of k

Table 4 Posterior distribution of k, galaxy data set, mixtures of normals model with $Poi(1)$ prior on k, see main text for hyperparameter values

k	1	2	3	4	5	6	7	8	9	10
$\pi(k\|x)$	0.000	0.000	0.090	0.291	0.349	0.191	0.063	0.013	0.002	0.000
s.d.	0.000	0.000	0.010	0.005	0.005	0.005	0.004	0.002	0.001	0.000

Note: 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.

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 k, the average should instead be taken with respect to $\pi(g|k,x,\phi)$.

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 multivariate normals

For multivariate normal components and the prior discussed in Section 2.1, the posteriors $\pi_j(\theta_j|x^j,\phi_j)$ in (17) are as follows: $r_j|g,x\sim W_b(\nu_j',\xi_j')$ and $m_j|r_j,g,x\sim N_b(\mu_j',\{\tau_j'r_j\}^{-1})$, where

$$\begin{aligned}\xi_j'=\xi_j&+\sum_{i\in A_j}(x_i-\overline{x}_j)(x_i-\overline{x}_j)^\top\\&+\frac{\tau_j n_j}{\tau_j+n_j}(\mu_j-\overline{x}_j)(\mu_j-\overline{x}_j)^\top,\end{aligned}$$

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

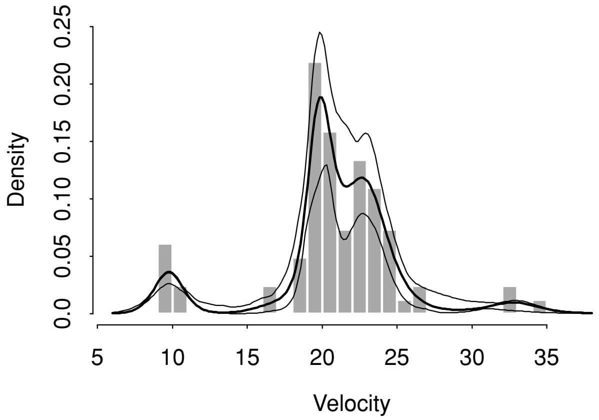

Fig. 5 Histogram and posterior predictive density for the galaxy data. The thick line is the estimate of the posterior predictive density $p(x_{n+1}|x,\phi)$, the thin lines give 0.005 and 0.995 quantiles of the simulated densities $f(x_{n+1}|k,g,x,\phi)$

The posterior predictive density $p_j(x_{n+1}|x^j,\phi_j)$ is the density of a b-variate t distribution with $\nu_j'-b+1$ degrees of freedom, location vector μ_j' and precision matrix $\{\tau_j'/(1+\tau_j')\}(\nu_j'-b+1)\xi_j'^{-1}$.

As an illustration, we fit a mixture of univariate normals to the galaxy data of Roeder (1990). These data consist of velocity measurements (1000 Km/sec) of 82 galaxies from the Corona Borealis region, a histogram is displayed in Fig. 5. Aitkin (2001) compares likelihood and Bayesian analyses of the data.

Table 4 contains an estimate of $\pi(k|x)$ obtained using the allocation sampler with hyperparameter values $\alpha=1$, $\mu=20$, $\tau=0.04$, $\nu=4$, $\xi=4$. Very little posterior mass is given to numbers of components outside the range from three to seven, in agreement with the estimate based on marginal likelihoods given in Nobile (2005). Figure 5 displays an estimate of the posterior predictive density $p(x_{n+1}|x,\phi)$. Examples of parametric inference for these data will be given in Section 4.4.

As another illustration, we model Fisher's iris data as a mixture of multivariate normals. The data consist 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 Fig. 6. Although the species of each flower is known, this information was not used in fitting the model. The hyperparameter

Table 5 Posterior distribution of k, iris data set, mixtures of multivariate normals model with $Poi(1)$ prior on k, see main text for hyperparameter values

k	1	2	3	4	5
$\pi(k\|x)$	0.000	0.002	0.718	0.267	0.013
s.d.	0.000	0.004	0.009	0.008	0.001

Note: 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.

values used were $\alpha = 1$, $\mu = (5.84, 3.06, 3.76, 1.20)^{\top}$, $\tau = 0.065$, $\nu = 7$ and $\xi = \text{diag}(0.55, 0.4, 0.35, 0.1)$. Table 5 contains an estimate of the posterior of the number of components.Most of the posterior mass is assigned to $k = 3$ and $k = 4$, with the actual number of species accounting for about 70% of the total mass. The posterior predictive density $p(x_{n+1}|x, \phi)$ of the iris data is four-dimensional, Fig. 6 displays the univariate and bivariate margins of an estimate. Within-sample classification probabilities are readily computed using the sample of g vectors. We condition on the modal number of components $k = 3$ and, after performing the post-processing discussed in Section 4.4, we plot in Fig. 7 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 five 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 readily computed from the output.

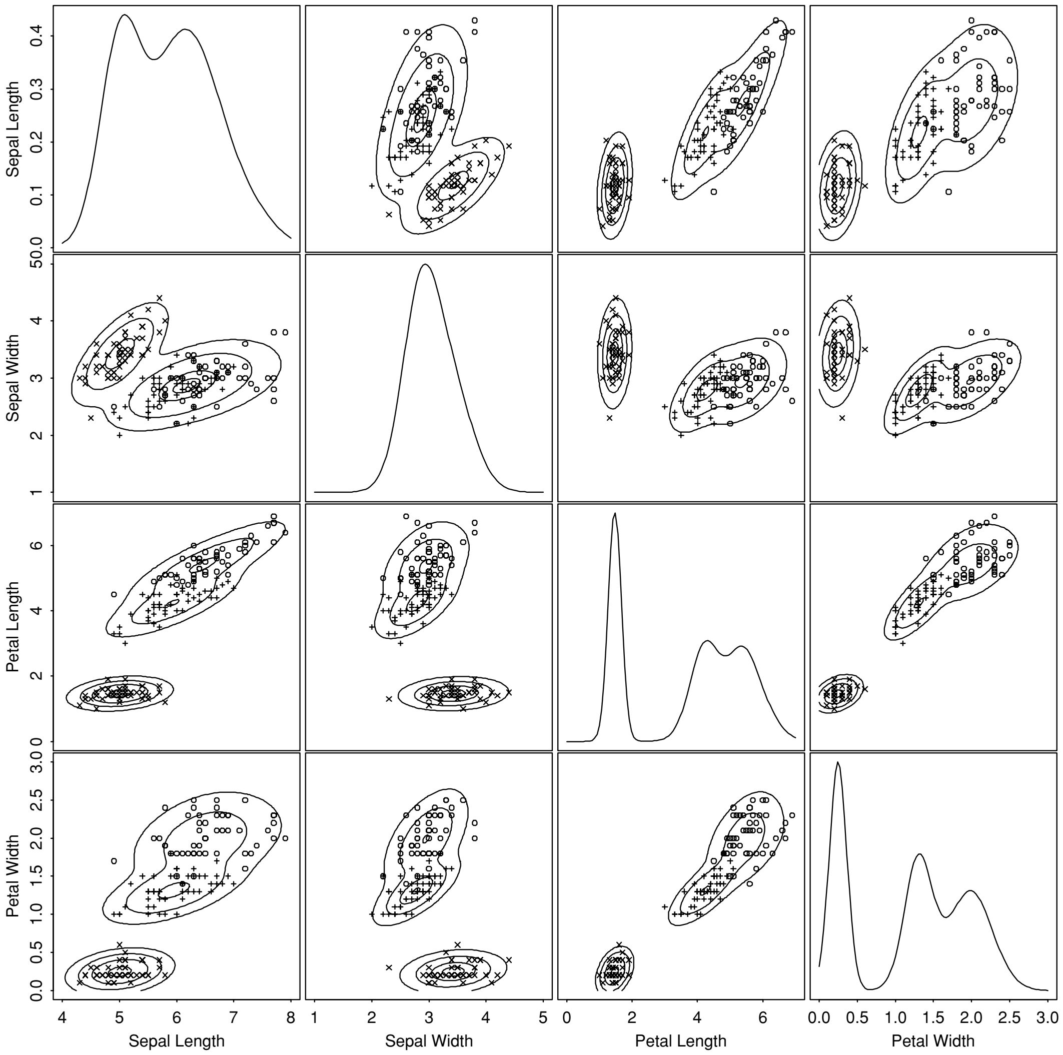

Fig. 6 Posterior predictive distribution, iris data set. Univariate margins (on main diagonal) and bivariate margins (off-diagonal) of an estimate of the four-dimensional posterior predictive density. In the bivariate plots, the contour lines are drawn at levels corresponding to 5%, 25%, 75% and 95% of the posterior probability of the displayed region. Overlaid on the contour plots are bivariate scatterplots of the data, using the symbols × = Setosa, + = Versicolor, ○ = Virginica

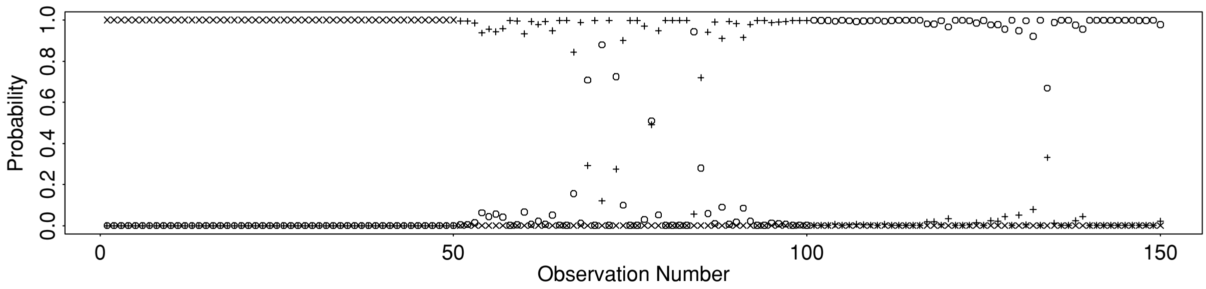

Fig. 7 Within-sample classification probabilities, iris data set

4.4 Identifiability and label switching

Finite mixture distributions are not identifiable: the likelihood (1) is invariant to permutations of the component labels. For predictive inference, lack of identifiability is of no concern, as the examples in the preceding sections show. However, classification and parametric inference require an unequivocal assignment of labels to the mixture components. To see why, let us assume that the prior is symmetric, so that the posterior is also invariant to permutations of the labels. As a consequence, the marginal posteriors 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 Fig. 8, which display the marginal posterior densities of the means in a 3-component 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 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. Nonetheless, from an inferential point of view, label switching is problematic since it precludes meaningful statements to be made about each component. Identifiability can be achieved by imposing constraints on the model parameters, either on the θ's, the λ's or both. However, this approach does not always work satisfactorily and other methods have been proposed. We refer to Richardson and Green (1997), Celeux et al. (2000), Stephens (2000b), Frühwirth-Schnatter (2001) and Jasra et al. (2005) for further discussion and additional references. The method we discuss in this section fits in the general framework of Stephens (2000b). Our setting is slightly different, 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 g.

Table 6 Estimates of some parameters in the eight mixtures of normals of Fig. 1, with a sample size $n = 2000$ and conditional on modal values of k as reported in Table 1

Mixture	Parameters	True	Raw output	Post-processed
(b)	m_1	0	2.75 ± 2.99	0.00 ± 0.02
	m_2	6	3.25 ± 2.99	6.00 ± 0.02
(c)	m_1	0	1.48 ± 1.50	0.00 ± 0.06
	m_2	3	1.51 ± 1.50	2.99 ± 0.06
(d)	m_1	0	1.24 ± 1.24	-0.02 ± 0.08
	m_2	2.5	1.18 ± 1.24	2.44 ± 0.15
(e)	m_1	0	0.00 ± 0.01	0.00 ± 0.02
	m_2	0	0.00 ± 0.01	0.00 ± 0.01
	$\sqrt{r_1}$	1	3.96 ± 3.00	1.00 ± 0.03
	$\sqrt{r_2}$	7	4.03 ± 3.00	6.98 ± 0.26
(f)	m_1		2.33 ± 1.70	0.58 ± 0.46
	m_2		2.40 ± 1.67	2.38 ± 1.23
	m_3		2.45 ± 1.70	4.22 ± 0.46
(g)	m_1	−5	2.30 ± 4.38	-5.00 ± 0.03
	m_2	0	1.77 ± 4.80	0.01 ± 0.05
	m_3	2.5	2.01 ± 4.37	2.50 ± 0.03
	m_4	4	1.79 ± 4.54	4.21 ± 0.34
	m_5	7	2.37 ± 4.67	8.52 ± 1.80
(h)	m_1		8.10 ± 8.04	-0.78 ± 0.11
	m_2		8.11 ± 7.76	2.91 ± 0.83
	m_3		7.71 ± 8.16	9.36 ± 2.11
	m_4		7.51 ± 8.01	19.93 ± 1.96

Note: The column "Raw output" contains means and standard deviations of the marginal posteriors based on the sampler's raw output. The same quantities, after re-assigning the labels, are shown in column "Post-processed". The true parameter values are reported in column "True", for the cases where modal k and true k coincide.

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

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

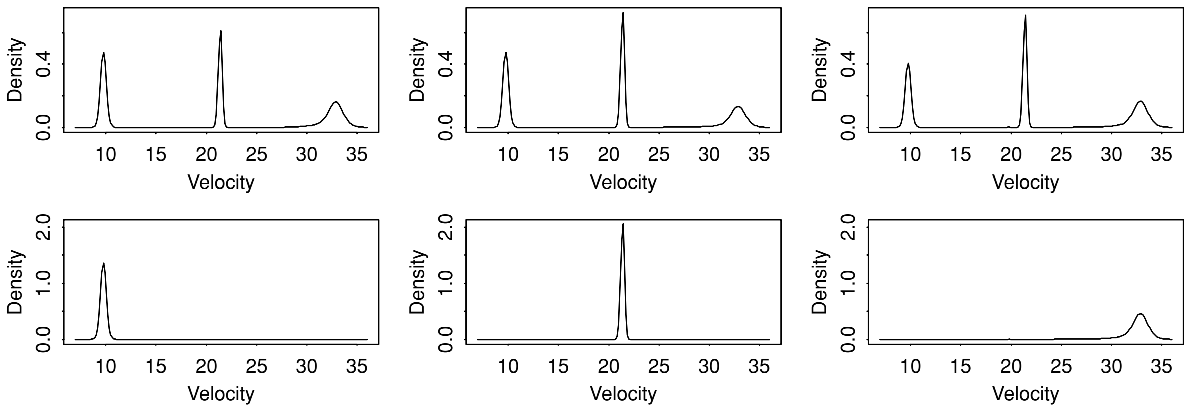

Fig. 8 Galaxy data, marginal posterior distributions of the component means m_1, m_2 and m_3 in the mixture of normals model, conditional on $k = 3$. The top row of plots displays estimates of the posteriors based on the raw sample of g vectors from the allocation sampler. Bottom row contains estimates using the allocation vectors with re-assigned labels

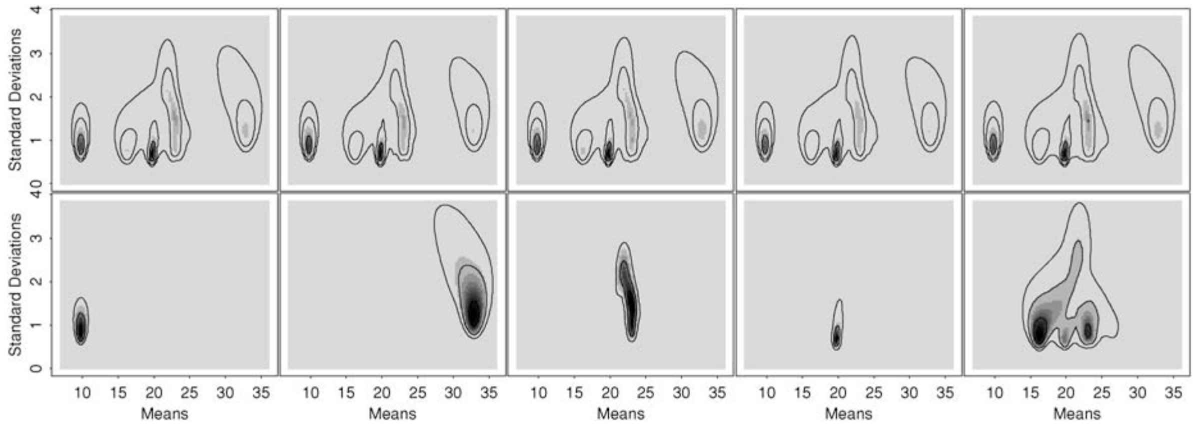

Fig. 9 Galaxy data, joint marginal posterior distributions of component means and standard deviations $(m_j, r_j^{-1/2})$, $j = 1, \ldots, 5$, in the mixture of normals model, conditional on $k = 5$. The top row of plots displays estimates of the posteriors based on the raw sample of g vectors from the 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

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

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

with respect to the sequence of permutations $\{\sigma^{(t)}, t = 1, \ldots, N\}$. An approximate solution to this problem can be obtained by replacing it with a sequence of optimisation problems, each involving a single permutation $\sigma^{(t)}$. 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 $\sigma^{(t)}$ which minimises the total cost $\sum_{j=1}^{\tilde{k}^{(t)}} C(j, \sigma_j^{(t)})$. The allocation vector with re-assigned labels is then equal to $\sigma^{(t)} g^{(t)}$. The plots in the bottom panels of Fig. 8 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 2 minutes for the actual sampling with $\Delta = 70$.

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 Fig. 9 were computed using the raw output of the 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. Our procedure manages to isolate very well the first four components, with means at about 10, 20, 23 and 33. The fifth component has a major mode with a mean of about 17, however, minor modes with means of 23 and 20 are still clearly visible.

Finally, Table 6 contains marginal posterior means and standard deviations for some of the parameters in the mixtures of normals of Fig. 1, with $n = 2000$, using the raw output and the post-processed allocations. All estimates are conditional on the modal values of k highlighted in Table 1. The "Raw output" estimates clearly display the effect of label switching. In contrast, the estimates after re-assigning the labels have much smaller variability and, in the cases where true k and modal k are equal, match very closely the true parameter values.

Appendix

A.1 Proposal ratio for move M1 in Section 3.1.1

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 (9) 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 Detailed description of move M3 in Section 3.1.1

Let j_1, j_2 be two randomly selected components and let $A = A_{j_1} \cup A_{j_2}$ be the index set of the observations currently allocated to j_1 or j_2, $A = \{i : g_i = j_1 \text{ or } g_i = j_2\}$. A candidate g' is formed as a modification of the current allocations g, by re-allocating the observations x_i, $i \in A$. The observations with index in A are processed sequentially, let $\tilde{A}$ denote the set of indexes that have already been treated, so that initially $\tilde{A} = \emptyset$. Let $\tilde{g}$ be the vector of allocations of observations with index not in A and of observations with index in A that have already been processed:

$$\tilde{g}_i = g_i,\ i \notin A; \quad \tilde{g}_i = g'_i,\ i \in \tilde{A}; \quad \tilde{g}_i \text{ undefined},\ i \in A \backslash \tilde{A}.$$

Also, let $\tilde{A}_j = \{i : \tilde{g}_i = j\}$, so that $\tilde{A} = \tilde{A}_{j_1} \cup \tilde{A}_{j_2}$, and let $\tilde{n}_j = \text{card}\{\tilde{A}_j\}$. Finally, let $\tilde{x} = \{x_i : i \notin A \backslash \tilde{A}\}$. Consider re-allocating the generic observation x_i. It is assigned to components j_1 or j_2 with respective probabilities $p_{j_1}^{(i)}, p_{j_2}^{(i)}$ satisfying $p_{j_1}^{(i)} + p_{j_2}^{(i)} = 1$ and the condition

$$\begin{aligned} \frac{p_{j_1}^{(i)}}{p_{j_2}^{(i)}} &= \frac{f(g'_i = j_1|\tilde{g}, x_i, \tilde{x}, k, \phi)}{f(g'_i = j_2|\tilde{g}, x_i, \tilde{x}, k, \phi)} \\ &= \frac{f(g'_i = j_1, \tilde{g}, x_i, \tilde{x}|k, \phi)}{f(g'_i = j_2, \tilde{g}, x_i, \tilde{x}|k, \phi)} \\ &= \frac{f(g'_i = j_1, \tilde{g}|k)}{f(g'_i = j_2, \tilde{g}|k)} \frac{f(x_i, \tilde{x}|k, g'_i = j_1, \tilde{g}, \phi)}{f(x_i, \tilde{x}|k, g'_i = j_2, \tilde{g}, \phi)} \end{aligned} \tag{22}$$

Now, from Eq. (5) the first term is easily shown to be

$$\frac{f(g'_i = j_1, \tilde{g}|k)}{f(g'_i = j_2, \tilde{g}|k)} = \frac{\alpha_{j_1} + \tilde{n}_{j_1}}{\alpha_{j_2} + \tilde{n}_{j_2}}. \tag{23}$$

Using Eq. (6), the second term in the right hand side of (22) can be rewritten as

$$\frac{f(x_i, \tilde{x}|k, g'_i = j_1, \tilde{g}, \phi)}{f(x_i, \tilde{x}|k, g'_i = j_2, \tilde{g}, \phi)} = \frac{p_{j_1}(x_i, \tilde{x}^{j_1}|\phi_{j_1}) \cdot p_{j_2}(\tilde{x}^{j_2}|\phi_{j_2})}{p_{j_1}(\tilde{x}^{j_1}|\phi_{j_1}) \cdot p_{j_2}(x_i, \tilde{x}^{j_2}|\phi_{j_2})}. \tag{24}$$

Substituting (23) and (24) into Eq. (22) one obtains

$$\frac{p_{j_1}^{(i)}}{1 - p_{j_1}^{(i)}} = \frac{\alpha_{j_1} + \tilde{n}_{j_1}}{\alpha_{j_2} + \tilde{n}_{j_2}} \cdot \frac{p_{j_1}(x_i, \tilde{x}^{j_1}|\phi_{j_1}) \cdot p_{j_2}(\tilde{x}^{j_2}|\phi_{j_2})}{p_{j_1}(\tilde{x}^{j_1}|\phi_{j_1}) \cdot p_{j_2}(x_i, \tilde{x}^{j_2}|\phi_{j_2})},$$

this is solved for $p_{j_1}^{(i)}$ and a draw of g'_i is made accordingly. Next the quantities $\tilde{g}$, $\tilde{A}$, $\tilde{A}_{j_1}$, $\tilde{A}_{j_2}$, $\tilde{x}$, $\tilde{n}_{j_1}$, $\tilde{n}_{j_2}$ are updated and another observation with index in $A \backslash \tilde{A}$ is processed. The procedure stops when $\tilde{A} = A$, resulting in a candidate allocation g' with proposal probability

$$P(g \to g') = \frac{1}{k(k-1)} \prod_{i \in A} p_{g'_i}^{(i)}$$

where the first term accounts for the random selection of the components j_1, j_2. It is straightforward to realize that the probability of proposing the reverse move from g' to g is

$$P(g' \to g) = \frac{1}{k(k-1)} \prod_{i \in A} p_{g_i}^{(i)}$$

so that the proposal ratio $P(g' \to g)/P(g \to g')$ is as given in Eq. (11). We should also remark that the order in which the observations with index in A are processed is random; this, however, does not affect the proposal ratio.

A.3 Distribution of p_E in Section 3.2

We use $p_E \sim \text{Beta}(a, a)$ 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

$$\frac{\Gamma(2a)}{\Gamma(a)} \frac{\Gamma(a+n_{j_1})}{\Gamma(2a+n_{j_1})} = \frac{p_0}{2}. \tag{25}$$

The left-hand side equals $(1/2)$ times a product of $n_{j_1} - 1$ terms, each monotonically decreasing in a, and hence it is monotonic decreasing in a. Therefore, (25) can be easily solved numerically, e.g. using bisection, except for a few small values of n_{j_1} if p_0 is too small. Since solving (25) numerically is relatively time consuming, in the simulation program we used a lookup table.

Acknowledgments We thank the Associate Editor and two referees for several comments that have led to considerable changes in the content and focus of the paper. Alastair Fearnside acknowledges financial support under a doctoral training grant from the U. K. Engineering and Physical Sciences Research Council.

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. 2006. Multivariate mixtures of normals with unknown number of components. Statistics and Computing 16: 57–68.

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.

Fearnhead P. 2004. Particle filters for mixture models with an unknown number of components. Statistics and Computing 14: 11–21.

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.

Jain S. and Neal R.M. 2004. A split-merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics 13: 158–182.

Jasra A., Holmes C.C., and Stephens D.A. 2005. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science 20: 50–67.

Marin J.-M., Mengersen K., and Robert C.P. 2005. Bayesian modelling and inference on mixtures of distributions. In: Dey D. and Rao C.R. (Eds.), Handbook of Statistics vol. 25, North-Holland.

McLachlan G. and Peel D. 2000. Finite Mixture Models, John Wiley & Sons, New York.

Mengersen K.L. and Robert C.P. 1996. Testing for Mixtures: A Bayesian entropic approach. In: Bernardo J.M. Berger J.O., Dawid A.P. and Smith A.F.M. (Eds.), Bayesian Statistics vol. 5, Oxford University Press, pp. 255–276.

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.

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

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

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, New York.

Zhang Z., Chan K.L., Wu Y., and Chen C. 2004. Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. Statistics and Computing 14: 343–355.