Statistical Science
2005, Vol. 20, No. 1, 50–67
DOI 10.1214/088342305000000016
© Institute of Mathematical Statistics, 2005

Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling

A. Jasra, C. C. Holmes and D. A. Stephens

Abstract. In the past ten years there has been a dramatic increase of interest in the Bayesian analysis of finite mixture models. This is primarily because of the emergence of Markov chain Monte Carlo (MCMC) methods. While MCMC provides a convenient way to draw inference from complicated statistical models, there are many, perhaps underappreciated, problems associated with the MCMC analysis of mixtures. The problems are mainly caused by the nonidentifiability of the components under symmetric priors, which leads to so-called *label switching* in the MCMC output. This means that ergodic averages of component specific quantities will be identical and thus useless for inference. We review the solutions to the label switching problem, such as artificial identifiability constraints, relabelling algorithms and label invariant loss functions. We also review various MCMC sampling schemes that have been suggested for mixture models and discuss posterior sensitivity to prior specification.

Key words and phrases: Bayesian statistics, mixture modeling, MCMC, label switching, identifiability, sensitivity analysis.

1. INTRODUCTION

In their intrinsic form, mixture models provide a flexible way to model heterogeneous data. That is, if data are thought to belong to one of k classes (or components), but whose individual class memberships are unavailable, then mixture models provide a natural framework for statistical modeling. Moreover, due to the large class of functions that can be approximated by a mixture model, they are attractive for describing nonstandard distributions. For a comprehensive list of the applications of mixture models see Titterington, Smith and Makov (1985), and for a recent overview see McLachlan and Peel (2000).

As a result of the early work of Newcomb (1886) and Pearson (1894) mixture models were established as a useful statistical tool. In addition, methodological advances in computational methods for frequentist mixture models, including the maximum likelihood approach of Baum, Petrie, Soules and Weiss (1970) and more generally the expectation–maximization (EM) algorithm (Dempster, Laird and Rubin, 1977), added to their popularity. However, difficulties often arise in the application of mixture models. For example, in the context of frequentist mixtures with location–scale component distributions, the likelihood can become unbounded (see Aitkin, 2001, for further details).

From the Bayesian perspective, before Markov chain Monte Carlo (MCMC) methods (Hastings, 1970; Green, 1995; for a general introduction, see Robert and Casella, 2004, or Liu, 2001), mixture models were restricted to a few specialized cases (e.g., Bernardo and Giròn, 1988). Following the work of Diebolt and

A. Jasra is a Ph.D. student, Department of Mathematics, Imperial College London, London, SW7 2AZ, UK (e-mail: ajay.jasra@imperial.ac.uk). C. C. Holmes is Lecturer in Statistics, Oxford Centre for Gene Function, Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK, and Mammalian Genetics Unit, MRC Harwell, UK (e-mail: cholmes@stats.ox.ac.uk). D. A. Stephens is Lecturer in Statistics, Department of Mathematics, Imperial College London, London, SW7 2AZ, UK (e-mail: d.stephens@imperial.ac.uk).

Robert (1994, data augmentation Gibbs sampler applied to mixtures), Bayesian mixture models could be applied routinely when the number of components is assumed known. Bayesian analysis via mixture models with an unknown number of components is now possible using the methods of Escobar and West (1995, Dirichlet process mixtures), Mengersen and Robert (1996, distributional distances), Richardson and Green (1997, reversible jump MCMC) and Stephens (2000a, birth-and-death MCMC). Due to the above developments, implementation of Bayesian mixtures has become increasingly popular in many academic disciplines, such as biological sequence analysis (Boys and Henderson, 2003), econometrics (Frühwirth-Schnatter, 2001; Hurn, Justel and Robert, 2003), machine learning (Beal, Ghahramani and Rasmussen, 2002) and epidemiology (Green and Richardson, 2002).

One of the main challenges of a Bayesian analysis using mixtures is the nonidentifiability of the components. That is, if exchangeable priors are placed upon the *parameters* of a mixture model, then the resulting posterior distribution will be invariant to permutations in the labelling of the parameters. As a result, the marginal posterior distributions for the parameters will be identical for each mixture component. Therefore, during MCMC simulation, the sampler encounters the symmetries of the posterior distribution and the interpretation of the labels switches. It is then meaningless to draw inference directly from MCMC output using ergodic averaging. Label switching significantly increases the effort required to produce a satisfactory Bayesian analysis of the data, but is a prerequisite of convergence of an MCMC sampler and therefore must be addressed. While convergence in MCMC simulation is a complex issue, we regard a *minimum requirement* of convergence for a mixture posterior to be such that we have explored all possible labellings of the parameters. We justify this choice in our examples in Section 3. For a discussion of convergence issues, see Robert and Casella (2004).

A difficulty in the Bayesian analysis of mixtures, when the number of components is unknown, is the sensitivity of the posterior distribution for the number of components to changes in the prior distribution for the parameters. Aitkin (2001) noted apparent difficulties in Bayesian analyses of mixture models and we discuss these concerns in this paper.

1.1 Interpretation of Mixture Models

In general, there are two ways in which mixture models can be interpreted. First is the missing data formulation. We assume that data $\mathbf{x} = (x_1, \ldots, x_n)$ are i.i.d. with distribution

$$x_i | z_i = j, \quad \phi_j \sim f(x_i; \phi_j) \tag{1}$$

for $j = 1, \ldots, k$, and the latent variables $\{z_n\}$ deconvolve the distribution of the data, with $p(z_i = j|\boldsymbol{\theta}) = \pi_j$ (with ϕ_j and $\boldsymbol{\theta}$ to be defined in the next section). However, if the i.i.d. assumption is relaxed, for example to Markovian dependence, we return the so-called hidden Markov model (HMM); see Baum and Petrie (1966) and Robert, Rydén and Titterington (2000). Therefore label switching is not restricted to "standard" mixture models (e.g., Richardson and Green, 1997), but to any model with conditional structure such as (1).

The second interpretation is through a semiparametric construction. As noted above, due to the ability of mixture models to approximate nonstandard distributions, they can be seen as alternatives to nonparametric models. The missing data approach is appropriate in terms of clustering and semiparametricity in areas such as density estimation.

1.2 Illustrative Example: The Crab Data

To illustrate some of the issues discussed in Section 1, we consider the famous crab data set analyzed by Pearson (1894). The data are shown in Figure 1 and comprise measurements of the ratio of forehead to body length of 1000 crabs, and were the focus of one of the first major analyses of data by a mixture model. The measurements were provided to Pearson by W. F. R. Weldon, who speculated that there were two new subspecies present. Following Pearson (1894), we use a two component normal mixture model to analyze

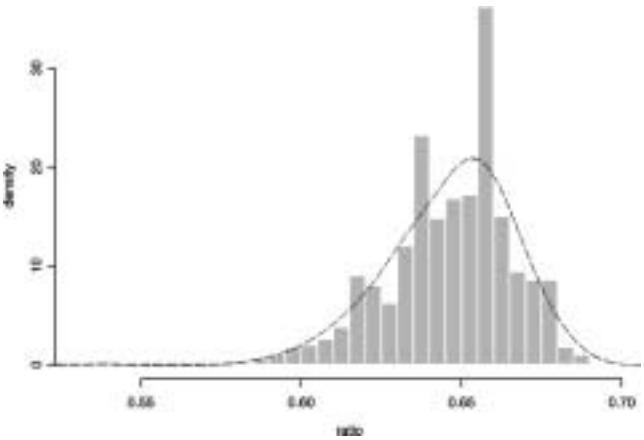

FIG. 1. *Histogram of the crab data with a kernel density estimate* (*dashed*) *overlaid.*

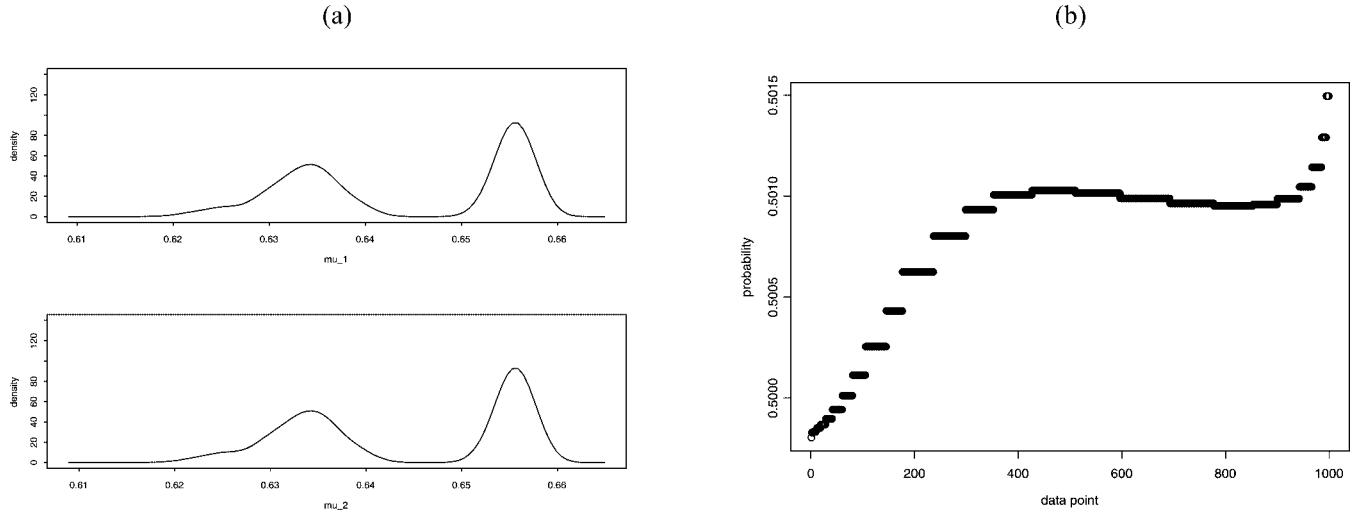

FIG. 2. (a) *Marginal posterior density estimates and* (b) *classification probabilities for the crab data. We fitted a two component mixture model to the data, and the output is the last* 88,000 *samples from a reversible jump sampler which were permuted for effect.*

these data. Our priors for the parameters are described in Section 2 and are exchangeable with respect to the labelling of the components.

In Figure 2 we observe the marginal posterior density estimates for the means [Figure 2(a)] and the classification probabilities [Figure 2(b)]. The classification probability, for this example, is the probability that a data point is in component/class 1, based on our MCMC output. The symmetries in the posterior distribution are immediately seen, with the posterior means being the same for each component, as well as the classification probabilities all being close to $1/2$.

There appears, however, to be significant information in the output. This is because there are two modes in the posterior for the means, which represent the two possible populations in the data. Label switching masks this information and we need a way to deal with it.

1.3 Solutions to the Label Switching Problem

For Bayesian mixtures the invariance of the likelihood to permutations in the labelling is not a problem that is as easily solved as in the frequentist approach. In the case of the latter, simple inequality constraints [artificial identifiability constraints (ICs)] on the parameter space can be used to break the symmetry in the likelihood (see McLachlan and Peel, 2000). For example, if the component parameters are θ_1 and θ_2, a possible constraint is $\theta_1 < \theta_2$. In the Bayesian context these constraints do not always perform adequately.

To demonstrate the above, consider the well-known Galaxy data (see, e.g., Stephens, 1997a). The data set was first presented by Postman, Huchra and Geller (1986) and consists of the velocities (in 10^3 km/s) of distant galaxies diverging from our own, taken from six well separated conic sections of the Corona Borealis: they can be observed in Figure 3. The data were originally of size 83, but we leave one observation out, in accordance with the analyses of Roeder (1990), Richardson and Green (1997) and Stephens (1997a). Since Richardson and Green (1997) found high posterior support for between five and seven components, we fit the random beta model (see Section 2.2 for further details) of Richardson and Green (1997) with a fixed number of six components to the data. We ran a Gibbs sampler (the fixed dimensional updates in

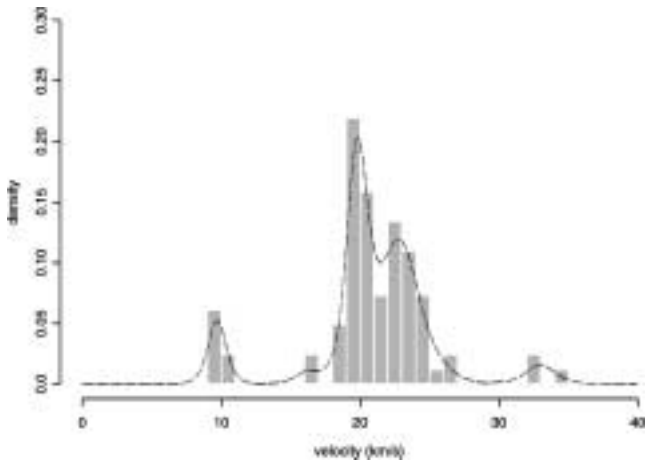

FIG. 3. *Histogram of the Galaxy data. We have overlaid the histogram with a kernel density estimate* (*dashed*).

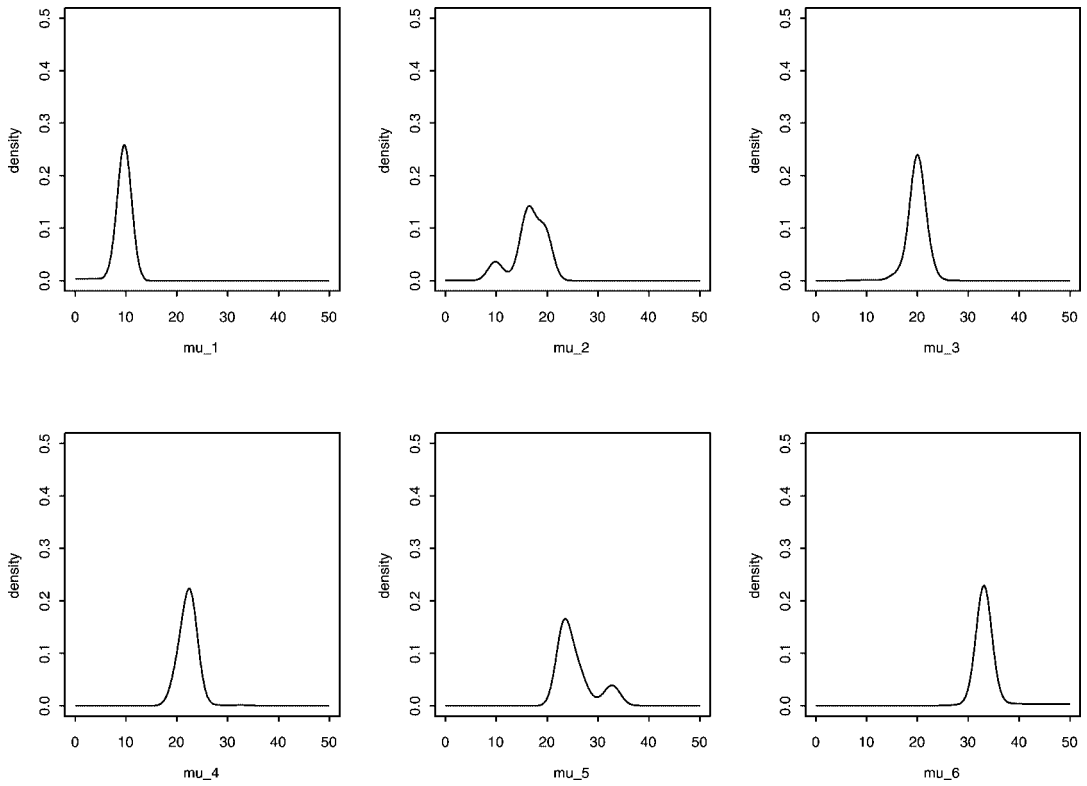

FIG. 4. *Marginal posterior density estimates of the sampled means of the galaxy data set. The means were permuted to obey the constraint* $\mu_1 < \cdots < \mu_6$. *We fitted a six component normal mixture to the data. The output is the last* 20,000 *iterations from the Gibbs sampler.*

Richardson and Green, 1997) for 20,000 iterations post burn in.

In Figure 4 we can observe the marginal posterior density estimates for the means under the identifiability constraint $\mu_1 < \cdots < \mu_6$, where μ_j denotes the mean parameter of the jth normal component. We can see that there is evidence of multimodality in components two and five, and it appears that the symmetry in the posterior has not been removed.

This problem is typical of MCMC mixture analysis and consequently there have been many ideas proposed to deal with label switching. Along with artificial identifiability constraints, Stephens (1997a, 2000b) and Celeux (1998) developed *relabelling algorithms* to perform a k-means type clustering of the MCMC samples. Additionally, Celeux, Hurn and Robert (2000) and Hurn, Justel and Robert (2003) used *label invariant loss functions*—a decision theoretic procedure. Related to ICs is the *random permutation sampler* of Frühwirth-Schnatter (2001), which was designed both to improve the mixing of an MCMC sampler and to be a convenient way to apply identifiability constraints. In this article we provide a review of these methods.

One simple solution to the label switching problem is to adopt the *maximum a posteriori* (MAP) estimator, which is equivalent to penalized maximum likelihood (see Ciuperca, Ridolfi and Idier, 2003, e.g.). As a result, the label switching problem is only of concern during simulation. However, one of the main attractions of using a Bayesian approach is the ability to reflect the uncertainties related to our inference. Clearly MAP estimation does not allow this. This aspect is of particular importance in mixture analysis, due to the likely genuine multimodality (modes which cannot be explained by permuting the labels of the parameters) of the posterior distribution (in our experience this occurs quite often). As a result, we do not believe that MAP estimates provide a general solution to the label switching problem, because of the inability of the estimate to accommodate competing explanations of the data.

1.4 Outline

The article is organized as follows. In Section 2 we introduce some notation and a particular mixture model that we will be studying. In Section 3 we review various MCMC sampling strategies for mixtures. When the number of components is fixed, it was established by Celeux, Hurn and Robert (2000) that the Gibbs sampler is not always appropriate

for sampling from a mixture posterior. This is because of the inability of the Gibbs sampler to traverse the support of highly multimodal distributions. We emphasize that we can simulate from a mixture posterior using Metropolis–Hastings updates without completion (simulation of the missing class labels) and that tempering MCMC (Neal, 1996) may be used. We also consider reparameterizations, as discussed by Celeux, Hurn and Robert (2000), and variable dimension samplers. Next, we examine the existing solutions to the label switching problem. We begin in Section 4 with identifiability constraints, then relabelling algorithms (Section 5) and finally label invariant loss functions (Section 6). In Section 7 we discuss some of the potential problems with prior specification in Bayesian mixture models with an unknown number of components. In Section 8 we conclude with our views on applying the methods reviewed as well as a future research area in Bayesian mixture modeling.

2. NOTATION AND MIXTURE MODELS

Throughout this article we use the following notation. We let $p(\cdot)$ represent a generic probability density. Denote data $\mathbf{x} = (x_1, \ldots, x_n)$ which is assumed to be independently and identically distributed (i.i.d.) with mixture distribution

$$p(x_i|\boldsymbol{\theta}, k) = \sum_{j=1}^{k} \pi_j f(x_i; \phi_j),$$

where f is some parametric component density/mass function, k is possibly unknown and finite, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ are component specific parameters, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ are the mixture proportions or weights and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) = ((\pi_1, \phi_1), \ldots, (\pi_k, \phi_k))$. Denote the parameter space $\phi \in \Phi \subseteq \mathbb{R}^p$ with $\mathcal{S}^{k-1} \times \Phi^k = \Theta^k$, where $\mathcal{S}^{k-1} = \{(\pi_1, \ldots, \pi_{k-1}) : \pi_1, \ldots, \pi_{k-1} \geq 0 \cap \pi_1 + \cdots + \pi_{k-1} \leq 1\}$.

Define a *permutation* σ of the labels $1, \ldots, k$ of a parameter vector $\boldsymbol{\theta}$ as

$$\sigma(\boldsymbol{\theta}) = \big(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(k)}\big),$$

where $\sigma \in S_k$, the set of all $k!$ permutations of $1, \ldots, k$.

The nonidentifiability in the posterior arises as

$$p\big(\mathbf{x}|\sigma(\boldsymbol{\theta}), k\big) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} \pi_{\sigma(j)} f\big(x_i; \phi_{\sigma(j)}\big) \right\}$$

is identical for all $\sigma \in S_k$. Hence if $p(\boldsymbol{\theta}) \equiv p(\sigma(\boldsymbol{\theta}))$ $\forall \sigma \in S_k$, then so is the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{x})$. As a result, if there are k components in a mixture model and there is one mode under a given labelling, there are $k!$ symmetric modes in the posterior distribution.

2.1 Random Beta Model

A well-known mixture model that we use for our examples is the random beta model of Richardson and Green (1997). The model is as follows: data $x_1, \ldots, x_n$ are i.i.d. with distribution

$$x_i|\boldsymbol{\theta}, k \sim \sum_{j=1}^{k} \pi_j \mathcal{N}(\mu_j, \lambda_j^{-1}),$$

where $\mathcal{N}(\mu, \lambda^{-1})$ denotes the normal distribution with mean μ and precision λ. The priors, which are the same for each component $j = 1, \ldots, k$, are taken to be

$$\begin{aligned} \mu_j &\sim \mathcal{N}(\xi, \kappa^{-1}), \\ \lambda_j|\beta &\sim \mathcal{G}a(\alpha, \beta), \\ \beta &\sim \mathcal{G}a(g, h), \\ \boldsymbol{\pi} &\sim \mathcal{D}(\delta), \end{aligned}$$

where $\mathcal{D}(\delta)$ is the symmetric Dirichlet distribution with parameter δ and $\mathcal{G}a(\alpha, \beta)$ is the gamma distribution, shape α, scale β. If k is unknown, we assume $k \sim \mathcal{U}_{\{1,\ldots,k_{\max}\}}$, where $\mathcal{U}_{\{1,\ldots,k_{\max}\}}$ is the uniform distribution on the integers $1, \ldots, k_{\max}$ with $k_{\max}$ known.

The purpose of the hierarchical structure on the variances is to reduce the effect of the prior on the posterior; improper priors are generally unavailable for mixtures (see Gruet, Philippe and Robert, 1999, for an example of improper priors in the mixture context). A problem with the above prior, when k is unknown, arises due to the Lindley–Bartlett paradox (Lindley, 1957; Bartlett, 1957). Jennison (1997) noted that, in the limit as $\kappa \to 0$ and $\beta \to \infty$, the posterior distribution for k favors models with fewer components. We illustrate this phenomenon in Section 7.

Quantities in which we often are interested are the classification probabilities, defined as

$$p(z_i = j|\mathbf{x}, k) = \int_{\Theta} \frac{\pi_j f(x_i; \phi_j)}{\sum_{l=1}^{k} \pi_l f(x_i; \phi_l)} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}.$$

Then if we were interested in a single "best" clustering, we might take the groups which are formed by the maximal classification probabilities.

3. MCMC SAMPLERS FOR MIXTURE MODELS

As we saw in Section 1.2, label switching creates a problem at the inferential stage of analysis. However, it does provide a useful convergence diagnostic at the simulation stage. That is, we know a priori that the

mixture posterior has $k!$ symmetric modes. Thus failure to visit them reveals that an MCMC sampler has not converged. Many different sampling schemes have been proposed for mixture models. We first review the most popular samplers that are available for simulating from standard mixtures with a known number of components.

3.1 Gibbs Sampler

Following Diebolt and Robert (1994), perhaps the most popular methods to simulate from a mixture posterior distribution uses data augmentation and the Gibbs sampler, that is, by simulating the unobserved $\mathbf{z}$. However, the highly multimodal nature of a mixture distribution often makes the Gibbs sampler inappropriate for this task. To illustrate such a case, we simulated 100 data points from $x_i \sim 1/4\{\mathcal{N}(-3, 0.55^2) + \mathcal{N}(0, 0.55^2) + \mathcal{N}(3, 0.55^2) + \mathcal{N}(6, 0.55^2)\}$ and then used the random beta model, with $k = 4$. We ran the sampler for 150,000 iterations post burn-in, the results of which are presented in Figure 5(a).

The most striking feature of Figure 5(a) is that the sampler appears to be performing well, in the sense that it has picked out the means from the data. The apparent "good" performance of the sampler is offset by the fact that it has only been able to visit one of the 4! symmetric modes in the posterior distribution. It may be the case that if we ran the sampler for more iterations, we would visit another symmetric mode. However, it is clear that the Gibbs sampler is unable to freely move around the space of this distribution. Such behavior is highly undesirable since it is possible that there are many regions of the posterior support that are not being explored by the sampler.

We have shown that the Gibbs sampler cannot always visit the $k!$ symmetric modes of a posterior mixture distribution easily. We note that "[f]rom a statistical viewpoint, exploration of the $k!$ modal regions is redundant" (Celeux, Hurn and Robert, 2000). Indeed, if we wish to explore all of the $k!$ symmetric modes, we could randomly permute the output from the sampler; that is, simply add a Metropolis–Hastings move that proposes a random permutation of the labels, which is accepted with probability 1 (as used by Frühwirth-Schnatter, 2001). Clearly, this course of action is only appropriate if the posterior distribution is not genuinely multimodal (which would not be known a priori to simulation). This is because, if a Gibbs sampler is unable to move around the support of a multimodal distribution and there exists genuine multimodality, then the sampler will not mix well (or at all) between the modes.

3.2 Metropolis–Hastings with Tempering Updates

Since the Gibbs sampler cannot visit all of the modes of a mixture target, we need to consider alternative methods. Cappé, Robert and Rydén (2001) made the following statement:

> We will not use completion to run our (MCMC) algorithm. That is to say, the latent variables $\{z_n\}$ is not to be simulated by the algorithm.... We believe that this choice is bound to accelerate convergence

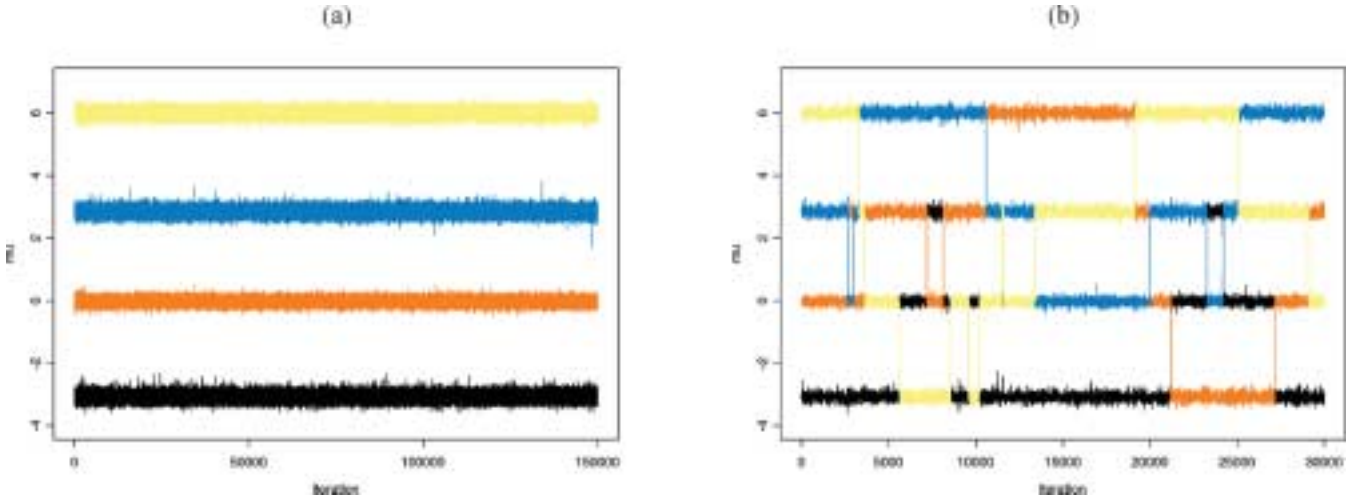

FIG. 5. *Trace plot of the sampled means for the simulated data of Section* 3.1. *We fitted a four component normal mixture to the data. The output is* (a) *the last* 150,000 *iterations from the Gibbs sampler and* (b) *the last* 150,000 *iterations from the tempering sampler* (*every fifth*). *The initial labelling is* μ_1 (*black*), μ_2 (*orange*), μ_3 (*blue*) *and* μ_4 (*yellow*).

of the algorithm by the drastic reduction in the dimensionality of the space.

We consider this approach by discarding the latent variables and updating the parameters using Metropolis–Hastings moves. Celeux, Hurn and Robert (2000) reported that random walk proposal mechanisms are too local, in the sense that the sampler cannot move freely around the $k!$ symmetric modes. They used tempering MCMC, which was developed by Neal (1996). We now introduce this method and apply it to the simulated data set of the previous section. We note that more advanced methods exist, for example, population or evolutionary Monte Carlo (EMC; Liang and Wong, 2001). We do not review these methods here, other than to note that population based MCMC works by embedding the target distribution of interest into a sequence of related distributions and sampling from $p^*(\cdot) \propto \prod_{j=0}^{m} p_j(\cdot)$, where $p_0(\cdot)$ is the original target distribution: for a review, see Liu (2001) and for an extension to the transdimensional case, see Jasra, Stephens and Holmes (2005).

Tempering MCMC uses what is essentially a Metropolis–Hastings kernel to sample from the posterior distribution, in which case it is often beneficial to reparameterize the mixture proportions in the random beta model. This is because Metropolis–Hastings moves may not perform well on a constrained space such as the simplex of mixing proportions. We choose the reparameterization $\pi_j = v_j / \sum_{l=1}^{k} v_l$ with $v_j > 0$ $\forall\, j$. We modify the prior for $\boldsymbol{\pi}$ as $v_j \sim \mathcal{G}(\delta, 1)$, with $v_j \perp\!\!\!\perp v_l$ $\forall\, j \neq l$, where $A \perp\!\!\!\perp B$ means A is independent of B. As a result, our reparameterized model is equivalent to the original model.

3.2.1 *Tempering MCMC.* Suppose we have a target distribution $p_0(\boldsymbol{\theta})$ which has many isolated modes. Now suppose we have a sequence of m related distributions $p_1(\boldsymbol{\theta}), \ldots, p_m(\boldsymbol{\theta})$. The final distribution p_m is (potentially) quite different from p_0, but is thought to be easier to sample from. The objective is to use these distributions to assist in the movement around the support of the target.

To propose a new state in the chain $\boldsymbol{\theta}'$, we use an up–down scheme described with pseudocode in Figure 6 (note that $a \wedge b$ means $\min\{a, b\}$). The figure tells us that we may need to draw from the intermediate distributions p_j via a Markov chain kernel. We note that this kernel itself may be a cycle of Metropolis–Hastings kernels; this is particularly useful if $\boldsymbol{\theta}$ is of high dimension.

Currently in state $\boldsymbol{\theta}$.

To propose a new state in the chain $\boldsymbol{\theta}'$

Draw $\hat{\boldsymbol{\theta}}_1$ from $\boldsymbol{\theta}$ using $\hat{T}_1$.

Draw $\hat{\boldsymbol{\theta}}_2$ from $\hat{\boldsymbol{\theta}}_1$ using $\hat{T}_2$.

$\vdots$

Draw $\hat{\boldsymbol{\theta}}_m$ from $\hat{\boldsymbol{\theta}}_{m-1}$ using $\hat{T}_m$.

Draw $\check{\boldsymbol{\theta}}_{m-1}$ from $\hat{\boldsymbol{\theta}}_m$ using $\check{T}_{m-1}$.

$\vdots$

Draw $\check{\boldsymbol{\theta}}_2$ from $\check{\boldsymbol{\theta}}_3$ using $\check{T}_2$.

Draw $\boldsymbol{\theta}'$ from $\check{\boldsymbol{\theta}}_2$ using $\check{T}_1$.

where $\hat{T}_j$, $\check{T}_j$ is a transition kernel that satisfies detailed balance with respect to p_j, $j = 1, \ldots, m$. Then the new state is accepted with probability

$$1 \wedge \frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} \cdots \frac{p_m(\hat{\boldsymbol{\theta}}_{m-1})}{p_{m-1}(\hat{\boldsymbol{\theta}}_{m-1})} \frac{p_{m-1}(\hat{\boldsymbol{\theta}}_m)}{p_m(\hat{\boldsymbol{\theta}}_m)} \cdots \frac{p_0(\boldsymbol{\theta}')}{p_1(\boldsymbol{\theta}')}.$$

FIG. 6. *Transition dynamics for tempering MCMC.*

To apply the method for mixtures we suppose $p_0(\cdot)$ is the posterior distribution. We then let $p_j(\cdot) \propto p_0(\cdot)^{1/\zeta_j}$, $j = 1, \ldots, m$, where $1 > \zeta_1 > \cdots > \zeta_m > 0$ (the ζ's act as a temperature parameter). The objective is during the first m simulations to flatten out the target, allowing us to walk freely on the space. Then, for the next $m - 1$ steps we return to a state that receives high posterior support under the target. To have sufficiently high acceptance probability the intermediate steps (i.e., the ζ's) should not have a large difference. We can add further simulations from $p_m(\cdot)$ to encourage movement between the modal regions.

3.2.2 *Tempering for the random beta model.* To apply tempering MCMC for the reparameterized random beta model, we make some modifications to the algorithm. First we add a Metropolis–Hastings step, so that with probability ω we perform a deterministic cycle of Metropolis–Hastings steps, implemented in the following manner. Draw a new $\boldsymbol{\mu}' = (\mu_1', \ldots, \mu_k')$ via an additive normal random walk. This move is accepted with probability $1 \wedge p(\boldsymbol{\mu}'|\cdots)/p(\boldsymbol{\mu}|\cdots)$, where

$$p(\boldsymbol{\mu}|\cdots) \propto p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{v})\, p(\boldsymbol{\mu}) \tag{2}$$

and $|\cdots$ denotes conditioning on all other variables. For $\boldsymbol{\lambda}' = (\lambda_1', \ldots, \lambda_k')$ and $\boldsymbol{v}' = (v_1', \ldots, v_k')$ we use reflective proposals, that is, normal random walks bounced off a barrier at zero.

The second modification is to use tempering to simulate from the full conditionals, that is, to sample from $p(\boldsymbol{\mu}|\cdots)$, $p(\boldsymbol{\lambda}|\cdots)$ and $p(\boldsymbol{v}|\cdots)$. We note

that this is a valid MCMC sampler since any kernel (which is a cycle) that is invariant with respect to the (full) conditional distributions for all $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$ will have invariant distribution $p(\cdot)$ (Tierney, 1994).

Our choice may seem odd, since the full conditionals may not be as multimodal as the full posterior. However, if we consider (2) we can see that this is of mixture form and is likely to have many modes.

Another reason we do this is because in other simulations (on a hidden Markov model), we obtained huge rejection rates when sampling from the full posterior. It may be the case that a reduction in the dimensionality of the parameters may facilitate higher acceptance rates. Our approach may lead to longer computing time than a single tempering move on $p(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{v}|\beta, \mathbf{x})$, but it may also require fewer iterations to converge due to higher acceptance rates.

The intermediate steps are performed using the same random walk proposal as above. Also, the levels are taken to be the same for each full conditional. The Metropolis–Hastings steps allow us to explore a modal region and the tempering allows us to move between modes. For all of our examples we set $\omega = \frac{1}{2}$.

Other proposals and reparameterizations that we have found to work well are as follows. For the precisions, a multiplicative random walk with lognormal proposals has been used. For the weights, a reparameterization onto the logit space, updating via an additive normal random walk, can prove to be an effective strategy. Additionally, if the sampler exhibits slow movement around the parameter space, we often use heavy-tailed (e.g., Cauchy) proposals to improve mixing.

3.3 Simulated Data Example

We now return to the simulated data at the beginning of the section, using the tempering sampler to draw from the posterior. We ran the MCMC sampler for sufficiently long post burn-in (until the sampled parameters seemed to stabilize), with appropriate thinning to take into account rejections of our moves. The number of steps for the tempering was 55, starting at $\zeta_1 = 2$ and increasing by 2 at each level. The choice of steps was tuned in prior simulations to achieve reasonable acceptance rates, which were (in the order $\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{v}$) for the Metropolis–Hastings moves $(0.15, 0.44, 0.23)$ and for the tempering moves $(0.007, 0.03, 0.016)$. The tempering acceptance rates appear to be quite low, but since they are used for global moves we are satisfied with the performance of our algorithm.

In Figure 5(b) we can observe the output. From this figure we can see the correct label switching behavior, the sampler visiting the majority [in fact, $(4! - 2)$] of the symmetric modes in the posterior distribution (mixing over colors). We note that for full convergence we would need to ensure that the sampler visits all 4! modes, but the behavior of the sampler is more than satisfactory.

3.4 Variable Dimension Samplers

Following Richardson and Green (1997) the standard way to simulate from a mixture with an unknown number of components is reversible jump MCMC (Green, 1995; for an up-to-date review, see Green, 2003). Reversible jump is simply an adaptation of the Metropolis–Hastings method, where the measure theoretic construction is necessary because of the lack of common dominating measure when jumping between distributions of differing dimension. Stephens (2000a) and Cappé, Robert and Rydén (2003) have considered continuous time samplers. For continuous time samplers the standard accept–reject step of a reversible jump sampler is replaced by new states that are always accepted, but that occur via either a marked point process, or more generally a Markov jump process with appropriate stationary distribution.

In a comprehensive comparison between reversible jump and continuous time samplers, Cappé, Robert and Rydén (2003) demonstrated that there was little difference between the two. Theoretically, they showed that a sequence of reversible jump samplers converges to a continuous time sampler (in Skorohod topology). In terms of performance, they showed that the continuous time sampler was less efficient in that the CPU time, on average, was longer. The main differences appear to be that, first, the continuous time sampler can visit unlikely regions in the support of the posterior, thus yielding a sort of springboard between different modal regions (Cappé, Robert and Rydén, 2003), and second, that for the continuous time sampler there is a "free" Rao–Blackwellization to reduce the variance of the Monte Carlo (MC) estimates of integrals. We have found that, in practice, these latter differences do not have a significant impact on the performance of the sampler (when compared to reversible jump) or the inference resulting from it.

3.4.1 *Completion.* Another aspect of interest is that of completion for variable dimension samplers. In our experience, for relatively small data sets (e.g., the Galaxy data), the sampler which simulates the missing

data comprehensively outperforms a sampler without completion. This is in terms of CPU time, mixing and convergence speed. For large data sets (e.g., $n = 5000$), we found that the sampler without completion converged faster (we would generally expect this due to the reduction in size of the state space), but at a cost of longer CPU time.

As noted by Richardson and Green (1997) and Stephens (2000a), mixing within k is often improved by using variable dimension samplers. This is because the sampler now has the ability to move around the modes of a distribution (conditional on k) via a model of lower or higher dimension, that is, by jumping out of the k mixture model space, moving around in a model of different dimension and then returning to a different region in the k space, thus escaping valleys in the posterior probability distribution that a fixed dimensional sampler is unable to scale. Indeed for the crab data example in Section 1 we used a reversible jump sampler, since the fixed dimension sampler was unable to visit a genuine mode in the posterior. Therefore, variable dimension samplers provide an alternative method to sample from a mixture posterior with a known number of components. This is at the cost of extra programming effort and inefficiency due to the fact that the sampler will not necessarily stay in the model of interest for the entire run.

3.4.2 *Multivariate mixtures.* Simulation from multivariate mixtures with an unknown number of components is also an issue of importance. Stephens (2000a) used a continuous time sampler to draw from a bivariate mixture posterior. More generally, Dellaportas and Papageorgiou (2004) constructed a reversible jump sampler for multivariate mixtures of normals, with split/merge moves operating on the space of eigenvalues and eigenvectors of the covariance matrix. Dellaportas and Papageorgiou (2004) demonstrated their method on two-, three- and five-dimensional data, and reported reasonable mixing properties for their algorithm.

3.5 Dirichlet Process Mixtures

An additional way to construct a mixture with an unknown number of components is the Dirichlet process mixture (DPM); see, for example, Escobar and West (1995). In this approach the prior distribution for the parameters of the mixture is treated as unknown, say G, and a Dirichlet process (DP) prior is placed upon it. The discreteness of G under the DP is exploited as follows. In any sample of parameters $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ (corresponding to the n possible components), there is positive probability of two or more points coinciding, thus reducing the components to k $(\leq n)$. Typically the full conditionals can be sampled from using Gibbs or possibly Metropolis–Hastings updates. West (1997) noted that DPMs are perhaps less general than the mixtures focused on in this article, stating that the approach is "more geared towards density estimation and related objectives than mixture deconvolution or parameter estimation."

3.6 Comments

We have considered samplers for mixtures. We have demonstrated that the Gibbs sampler is not always appropriate for sampling from a mixture posterior, though for some cases such as the Galaxy data example (Section 1), the samples were more than adequate. We saw that tempering updates allowed full exploration of the mixture posterior.

To use tempering MCMC (as we have used it) we need to be able to calculate the marginal likelihood. Clearly, for standard mixtures this is straightforward. However, if the latent variables have special structure (e.g., Markovian), this calculation may be difficult. We have found that for HMMs, marginal update schemes (which require the "forward" step of the forward–backward algorithm of Baum et al., 1970) can be computationally slow and sometimes lead to large autocorrelations and rejection rates in Metropolis–Hastings moves (as found by Boys and Henderson, 2003).

There are other approaches for simulating from a mixture posterior, including perfect samplers: see Casella, Mengersen, Robert and Titterington (2002). Marin, Mengersen and Robert (2005) provide a review. Also work on exact simulation for change-point problems via the forward–backward method can be found in Fearnhead (2004).

We now review the various ways to deal with label switching, beginning with artificial identifiability constraints. To use the methods that we review, we recommend conditioning on k as argued by Robert (1997). Therefore, we always consider label switching for mixtures with a fixed number of components. We also assume that we have a sampler that can visit the $k!$ symmetric modes of the posterior.

4. ARTIFICIAL IDENTIFIABILITY CONSTRAINTS

4.1 The Method

We define an *identifiability constraint* as a condition on the parameter space Θ, such that only one

permutation can satisfy it. An example of such a constraint is $\mu_1 < \cdots < \mu_k$ in the univariate random beta model. This identifiability constraint is *artificial*, as it does not arise from any genuine knowledge or belief about the model, but is rather an apparent inferential convenience.

We refer to ICs in the way they were initially used (as in Diebolt and Robert, 1994, or Dellaportas, Stephens, Smith and Guttman, 1996). In other words, at every iteration of the sampler, we permute the samples so they satisfy the constraint. When applying ICs, it is best to search for identifiability constraints that lead to density estimates of the parameters that are as unimodal as possible (as stated by Richardson and Green, 1997). Indications of inappropriate identifiability constraints include exaggerated skewness and multimodality in the density estimates.

The motivation behind imposing an IC is the following. Since the likelihood and prior are invariant to the labelling of the parameters, if we impose an identifiability constraint on the parameter space, we break the symmetry in the posterior and the labelling problem should be solved. We would therefore focus on one of the $k!$ symmetric modes and output from the MCMC sampler can then be interpreted.

An identifiability constraint need not be imposed before any simulation takes place. Stephens (1997a) proved that inference conditional on an identifiability constraint can be performed when the constraint is imposed after the MCMC run (see Proposition 3.1 and Corollary 3.2 of Stephens, 1997a). Such procedures are equivalent to *changing the prior distribution*. That is, since the marginal posterior distributions are the same for each label and the likelihood is invariant to permutations, we define a new prior $p_n(\boldsymbol{\theta})$ such that

$$p_n(\boldsymbol{\theta}) = k! p(\boldsymbol{\theta}) \mathbb{I}_{(\boldsymbol{\theta} \in C)},$$

where C is the constraint, $\mathbb{I}_{(\cdot)}$ is the indicator function and $p(\boldsymbol{\theta})$ is the unconstrained prior.

An alternative approach to identifiability constraints was provided by Frühwirth-Schnatter (2001). Frühwirth-Schnatter used a "random permutation" sampler (RPS). That is, at every iteration of an MCMC sampler, a Metropolis–Hastings move is used to propose a new permutation of the labels. This ensures the sampler visits all $k!$ symmetric modes. Frühwirth-Schnatter then applied exploratory data analysis on the MCMC output from the RPS by applying ICs. We search for constraints that give the clearest picture of all of the parameters (i.e., much the same as the recommendations of Richardson and Green, 1997, discussed above).

4.2 Comments on the Method

Identifiability constraints have come under much scrutiny in the literature. Celeux (1997) and Celeux, Hurn and Robert (2000) and Stephens (1997a, 1997b, 2000b) all voiced their concerns about imposing an identifiability constraint.

Much of the initial attention was confined to the effect on the MCMC sampler, such as the implications of truncating the support of the posterior in terms of simulation (as mentioned by Celeux, 1997, and Celeux, Hurn and Robert, 2000). However, as stated above, since identifiability constraints can be imposed after simulation, we can simulate from the unconstrained posterior distribution and then impose an IC. As a result, there is no problem in terms of an adverse effect on simulation.

The use of exchangeable priors is normally an attempt to be weakly informative (e.g., Richardson and Green, 1997). However, if the identifiability constraint used does not correctly isolate one of the $k!$ symmetric modes, we would hesitate to call such a specification weakly informative. This is because the prior will become highly influential on our inference, as demonstrated by Celeux, Hurn and Robert (2000). We note that they called this "disturbing." We contend that it is only to be *expected*, since different constraints correspond to different models.

One problem with identifiability constraints is the choice of constraint. Frühwirth-Schnatter (2001) suggested that "if the components of the state specific parameters have some physical meaning, then an expert in the field will have some idea in which way the groups or states differ and might be able to offer such an identifiability constraint." This seems reasonable. However, we stress that if such expert opinion is available, an effort to produce subjective priors should be made. This may mean that there is no label switching at all, although label switching can still occur under subjective priors.

A more general difficulty of using ICs is in multivariate problems. Finding suitable identifiability constraints in such situations is almost impossible. Moreover, it can be difficult to anticipate the overall effect of such an action. We saw at the beginning of this paper that identifiability constraints do not always work, so we consider this example in more detail.

4.3 Example: Galaxy Data Revisited I

We now use the random beta model of Richardson and Green (1997) to illustrate that identifiability constraints can often induce informative priors that do not

Intialise algorithm with permutations $\sigma_1, \ldots, \sigma_N$.

Repeat until a fixed point is reached.

1. Choose $\hat{a}$ to minimise $\sum_{t=1}^{N} L_0(a, \sigma_t(\boldsymbol{\theta}^{(t)}))$.

2. For $t = 1, \ldots, N$ choose σ_t to minimise $L(\hat{a}, \sigma_t(\boldsymbol{\theta}^{(t)}))$.

FIG. 7. *A general relabelling algorithm.*

necessarily reflect the objective of their exchangeable versions. To do this, we return to the output from the Galaxy data in Section 1.3.

We computed the estimated means, conditional on the identifiability constraint $\mu_1 < \cdots < \mu_6$, through ergodic averaging and compared them with the means estimated by the relabelling algorithm in Figure 7 (which we discuss in the next section; we believe the estimates reflect one of the 6! symmetric modes of the posterior). The results can be seen in Table 1. For most of the means we can see that the new prior induced by the identifiability constraint produces very different results to the "correct" clustering of the MCMC samples; hence, the constraint is more influential than was intended.

TABLE 1
Estimated means

Parameter	Constraint	KL
μ_1	8.07	9.71
μ_2	16.46	19.01
μ_3	19.90	19.88
μ_4	22.21	22.71
μ_5	25.62	22.86
μ_6	34.84	32.92

NOTE. The constraint column is the estimated means under the identifiability constraint $\mu_1 < \cdots < \mu_6$. The KL column is the estimated means under the relabelling algorithm in Figure 8.

5. RELABELLING ALGORITHMS

5.1 The Method

Relabelling algorithms were developed by Stephens (1997a, 1997b, 2000b) and Celeux (1998). The idea is as follows: Suppose we define a loss function $L : \mathcal{A} \times \Theta \to [0, \infty)$ such that

$$L(a, \boldsymbol{\theta}) = \min_{\sigma \in S_k} \{L_0(a, \sigma(\boldsymbol{\theta}))\},$$

where $\mathcal{A}$ is the action space. Then the optimal action a^* is

$$a^* = \arg\min_a \int_\Theta L(a, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}. \tag{3}$$

Since the integral in (3) cannot be computed exactly, we use an MC estimate. We suppose that we draw N samples from the posterior distribution (denote these $\boldsymbol{\theta}^{(t)}$, $t = 1, \ldots, N$) and then apply the algorithm in Figure 7 to remove label switching.

The idea behind the algorithm is the following. Since the likelihood is invariant to permutations in the labelling, we seek to minimize the loss of performing an action a associated with $\boldsymbol{\theta}$ by selecting the permutation that minimizes this loss and then minimizing the posterior expected loss.

Stephens (2000b) derived an algorithm for clustering inference based on reporting an $n \times k$ matrix, Q, of classification probabilities, that is, q_{ij} is the probability that observation i is assigned to class j. Let $P(\boldsymbol{\theta})$ denote the true matrix of classification probabilities, where $p_{ij}(\boldsymbol{\theta}) = \pi_j f(x_i; \phi_j) / \sum_{l=1}^{k} \pi_l f(x_i; \phi_l)$. Stephens used the Kullback–Liebler (KL) divergence to measure the loss of reporting Q when the true probabilities are $P(\theta)$. The algorithm is given in Figure 8. We refer to this algorithm as the KL algorithm.

To apply a relabelling algorithm, we must choose m (well dispersed) starting points, since the algorithm is only guaranteed to converge to a local minimum. We then select the permutations and quantities that give the optimal solution. If storage requirements are too substantial, then an online version can be implemented. Additionally, step 2 of Figure 7 can be performed efficiently (for medium k) using the transportation algorithm; see Stephens (2000b) for details.

5.2 Comments on the Method

The form of a relabelling algorithm is exactly that of a k-means type clustering algorithm: Stephens took advantage of the special nature of the problem at hand. In our view, the method is an *automatic way to apply* (*or induce*) *an identifiability constraint.* That is, under an inferential objective the permutations of the labelling

Intialise algorithm with permutations $\sigma_1, \ldots, \sigma_N$.

Repeat until a fixed point is reached.

1. Choose $\hat{Q}$ to minimise

$$\sum_{t=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij} \left\{\sigma_t(\boldsymbol{\theta}^{(t)})\right\} \log \left\{ \frac{p_{ij}\{\sigma_t(\boldsymbol{\theta}^{(t)})\}}{q_{ij}} \right\}.$$

2. For $t = 1, \ldots, N$ choose σ_t to minimise

$$\sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij} \left\{\sigma_t(\boldsymbol{\theta}^{(t)})\right\} \log \left\{ \frac{p_{ij}\{\sigma_t(\boldsymbol{\theta}^{(t)})\}}{\hat{q}_{ij}} \right\}.$$

FIG. 8. *Stephens's KL algorithm for clustering inference.*

are induced or discovered so that all of the samples are labelled in the same way. We can then draw inference on any quantity by using ergodic averaging on the permuted samples.

We feel that the fact that the method is simply a way to apply an identifiability constraint (i.e., that the statistical model is changed) is underappreciated in the literature; it is not a fully decision theoretic method. That is, under a fully decision theoretic method, for every quantity of interest, we derive a loss function for estimation, which is not the case for relabelling algorithms (indeed, it would not be sensible since we may have different quantities that were estimated conditional on different identifiability constraints).

A method related to relabelling algorithms was given by Marin, Mengersen and Robert (2005), who found the MAP estimate of the parameters based on all of the MCMC samples. Then, to permute the MCMC samples, they found the permutation that minimizes the canonical scalar product between the MAP estimator and the sample. This method is simple to use, but has a one major drawback when the parameter space features many genuine modes. In this case the MAP estimate will ignore minor modes and may lead to inappropriate identifiability constraints being used.

We now demonstrate that relabelling algorithms apply or induce an identifiability constraint in the following example.

5.3 Example: Crab Data Revisited

We used Stephens's KL algorithm to deal with the label switching of the crab data example. The density estimates of the relabelled marginal posteriors and the classification probabilities can be seen in Figure 9.

Application of Stephens' KL algorithm has induced the identifiability constraint $\mu_1 < \mu_2$ (note that this was for this example and is not a general mathematical result). This allows us to perform inference that is supported by Figure 9(b), which shows that the classification probabilities are far more discriminated under the relabelled samples.

We note that if we wanted to estimate the means say, it would be most sensible to take the estimate over the permuted MCMC samples. This is because, if we applied another algorithm with a different loss function, we might not obtain the same constraint (in this example it is unlikely another constraint would transpire). However, Stephens (2000b) reported that it is generally the case, when there is no genuine multimodality, that different relabelling algorithms often produce similar permutations.

Another potential problem with ICs (and hence relabelling algorithms), when the data in the components are poorly separated, is the following. Gruet, Philippe and Robert (1999) found that one of the components overwhelms the others, which become negligible. We explore this in the following example.

5.4 Example: Rao's Plant Data

For our next example we consider the plant data of Rao (1948). The data consist of the heights (in centimeters) of 454 plants of two different types. Rao (1948) used a two component normal mixture model to analyze the data and the data were presented in a fre-

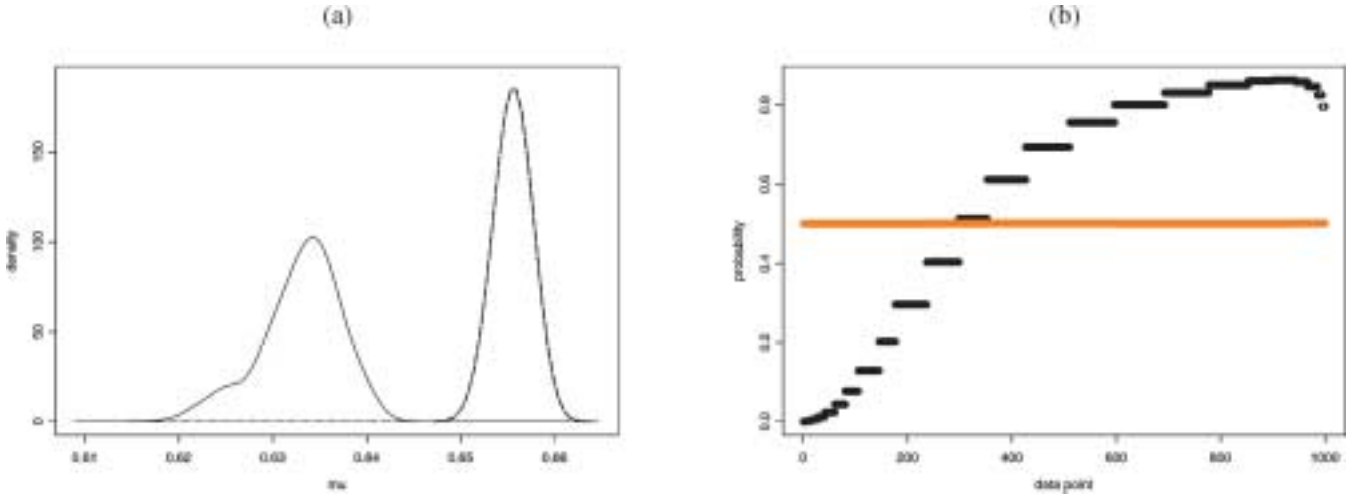

FIG. 9. (a) *Marginal posterior density estimates* [μ_1 (*unbroken*), μ_2 (*dashed*)] *and* (b) *classification probabilities for the crab data. We used Stephens's KL algorithm* (*Figure* 8) *to deal with the label switching. For* (b) *the orange dots are the classification probabilities under the unconstrained prior.*

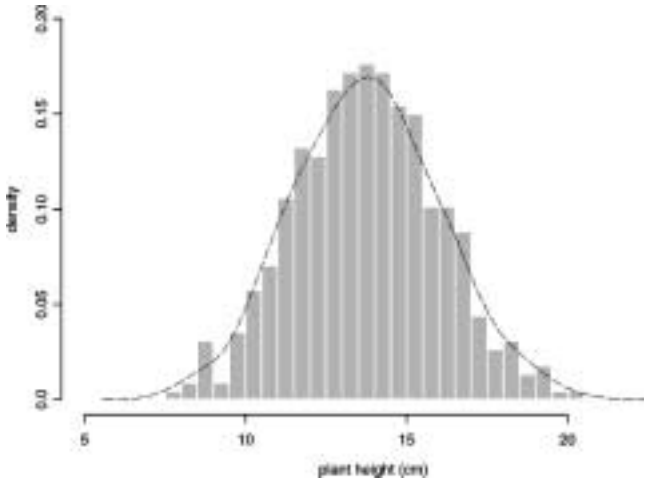

FIG. 10. *Histogram of plant data. We overlaid the histogram with a kernel density estimate* (*dashed*).

quency table, so they are essentially interval censored. To remove this effect, we added a $\mathcal{U}_{[-0.5,0.5]}$ random variable (where $\mathcal{U}_{[a,b]}$ is the continuous uniform distribution on $[a, b]$) to each data point (which we interpreted as the midpoint of each class interval).

The physical motivation for a mixture is clear, but let us observe the data given in Figure 10. We can see that if the data truly comprise a two component mixture, it is difficult to see evidence of this from the histogram (note that the sampling density/histogram need not be bimodal to be a two component mixture). Thus we suggest that the data in the components are poorly separated.

To analyze these data, we use the random beta model with $k = 2$ and use the KL algorithm to deal with label switching. We used tempering MCMC to sample from the posterior distribution. The trace plots of the parameters for 10,000 iterations post burn-in are given in Figure 11. From Figure 11(a) we can observe that there is significant label switching, since the plots for both components are similar.

We then used the KL algorithm to undo label switching. While the object of inference is not clustering, we can observe in Figure 11(b) that the algorithm has worked well. This is because the algorithm appears to have isolated one of the two symmetric modes in the posterior distribution.

We can see from Table 2 that (for the relabelled samples) component 1 dominates component 2 (since $\pi = 0.804$). However, we feel that this is *not* a defect of using a relabelling algorithm (and hence an IC). This is because the relabelled MCMC output appears to be correct (no bias of the constraint), as demonstrated by the fact that the relabelled components are so different, that is, that the plot on the left in Figure 11(b) has lower variance than the plot on the right. Although the inference from the mixture model (based upon the relabelled samples) appears to be incorrect (e.g., Rao estimated the mixture proportion as 0.566, but we note that our data may not give identical results to the original analysis because we have perturbed them), this does not matter from the perspective of dealing with label switching. The relabelling algorithm has performed well and exposed the view of the data that one component provides sufficient explanation of the data. In such cases, it may be more appropriate to determine subjective priors or different component densities.

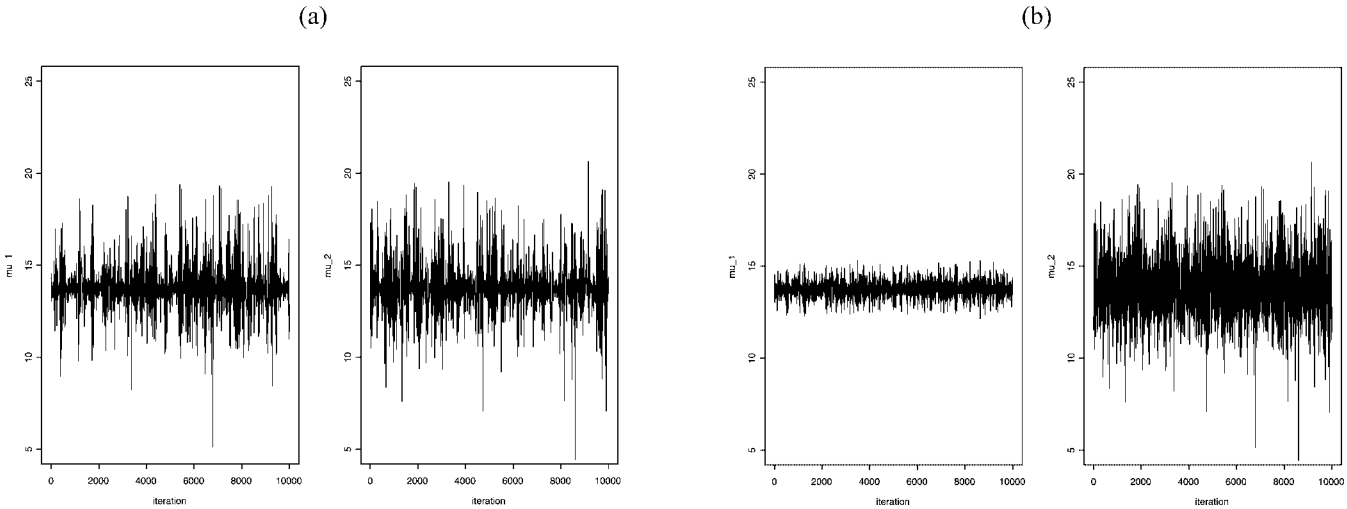

FIG. 11. *Tempering MCMC trace plots for plant data*: (a) *the means as returned by our algorithm for* 10,000 *iterations*; (b) *the relabelled means*.

TABLE 2
Parameter estimates (ergodic averages) for the plant data based on the unconstrained prior and the permuted samples

Parameter	Raw output	KL
μ_1	13.824	13.747
μ_2	13.819	13.897
λ_1^{-1}	3.773	4.690
λ_2^{-1}	3.66	3.075
π	0.519	0.804

5.5 Justification of Relabelling Algorithms

Stephens (1997a) gave possible justifications for using relabelling algorithms. The first justification was "revisionist Bayes." As stated in Section 4.1, applying an identifiability constraint is equivalent to changing the prior distribution. Therefore, when we use a relabelling algorithm we are essentially changing the statistical model. How might we justify such an approach? Stephens suggested that we are returning to the modeling stage and forcing identifiability on the components of the mixture. Many Bayesians may find this explanation distinctly unsatisfactory, but we agree that it is a credible justification, because if we seek to draw meaningful inference from the parameters of a mixture model, we need to be able to interpret them. If we use the extra information from the data (and hence the MCMC sampler), we can do so.

The second justification was "mode hunter." This is simply the view that we seek to find the permutations of the labelling which minimize the loss of computing the quantity of interest, and that we can calculate no other quantity.

Under the revisionist Bayes justification one actually believes in the permutations applied and is convinced that the model is representative of the real world problem. This is not the case under the mode hunter view.

5.6 Discussion

In this section we have shown that relabelling algorithms impose an IC. The exact nature of the constraint depends on both the loss function chosen and the MCMC samples themselves.

We have stated that relabelling algorithms should be used with care when the data in the components are similar. Note that inference from the parameters in this case is not meaningless. There are many instances in the literature where analysis of data when the components are poorly separated has occurred, for example, Rao (1948) and more recently Gruet, Philippe and Robert (1999). Additionally there may be situations where we do not know a priori that the components are poorly separated, for example, if we were analyzing high-dimensional data.

6. LABEL INVARIANT LOSS FUNCTIONS

6.1 The Method

Define a loss function $L : \mathcal{A} \times \Theta \to [0, \infty)$ such that

$$L(a, \boldsymbol{\theta}) = L\big(a, \sigma(\boldsymbol{\theta})\big) \quad \forall \sigma \in S_k.$$

Using such a loss function solves the labelling problem immediately. The way in which the method is applied (as in Celeux, Hurn and Robert, 2000, and Hurn, Justel and Robert, 2003) is the following. Compute the posterior expected loss

$$\begin{aligned}\mathbb{E}[L(a, \boldsymbol{\theta})|\mathbf{x}] &= \int_\Theta L(a, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \\ &\approx \frac{1}{N}\sum_{i=1}^{N} L\big(a, \boldsymbol{\theta}^{(t)}\big).\end{aligned} \tag{4}$$

Normally (4) cannot be minimized analytically, and so stochastic optimization methods (e.g., simulated annealing) are implemented.

An example of a particular loss function used by Hurn, Justel and Robert (2003) for clustering inference is

$$\begin{aligned}L(a, \mathbf{z}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \big\{&\mathbb{I}_{(z_i=z_j)}\big(1 - \mathbb{I}_{(a_i=a_j)}\big) \\ &+ \mathbb{I}_{(a_i=a_j)}\big(1 - \mathbb{I}_{(z_i=z_j)}\big)\big\},\end{aligned}$$

where a_i is the allocation that we give for the ith data point. This loss function is based on a pairwise comparison of the allocation of data points. If the true pair of data points is in the same class and our decision is that it is not, then we lose one. Conversely, if we choose the correct allocation we lose zero. To compute the posterior expected loss, we have

$$\begin{aligned}&\mathbb{E}[L(a, \mathbf{z})|\mathbf{x}] \\ &\quad= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \big\{p(z_i = z_j|\mathbf{x})\big(1 - \mathbb{I}_{(a_i=a_j)}\big) \\ &\qquad + \mathbb{I}_{(a_i=a_j)}\, p(z_i \neq z_j|\mathbf{x})\big\}.\end{aligned} \tag{5}$$

We then estimate (5) from our MCMC output (it is invariant to the labelling) and minimize with respect to a. For examples of other loss functions, see Celeux, Hurn and Robert (2000) and Hurn, Justel and Robert (2003).

6.2 Comments on the Method

From the Bayesian point of view the method of label invariant loss functions is more satisfactory than identifiability constraints. That is, we draw inference conditional only on the data. The method is fully decision theoretic: for every quantity of interest we must construct a loss function, and perform the expectation and then the minimization procedure highlighted above. This approach acknowledges both that the marginal posteriors for the labels are the same and that there is significant information in the MCMC output. It is therefore a fully Bayesian procedure.

The main difficulty with this method is the computational cost. Performing a simulated annealing algorithm for many loss functions will often be computationally expensive and may not be feasible for some functions.

A second drawback is the fact that the method is restricted to a class of loss functions that may or may not make sense for the decision problem at hand. That is, the method requires that the loss functions be invariant to the labelling of the parameters and that it is computationally feasible to minimize the posterior expected loss. Whether loss functions can be constructed within this class depends on the statistical objectives.

7. ROLE OF THE PRIOR IN BAYESIAN MIXTURE MODELING

One of the difficulties in Bayesian modeling is specifying a prior distribution when there is little information to be used. This is often the case in Bayesian analyses via mixture models with an unknown number of components and can lead to some additional (i.e., not label switching) inferential difficulties. We now demonstrate one such problem for the random beta model.

7.1 Example: Galaxy Data Revisited II

We reanalyzed the Galaxy data using the random beta model with k unknown. For seven different settings of κ (the normal prior precision on the component mean parameter), we ran a reversible jump sampler (similar to Richardson and Green, 1997, except that we used a split/merge move that is much the same as described by Cappé, Robert and Rydén, 2003, Appendix C) for 500,000 iterations, taking the burn-in to be 100,000 iterations. The results are given in Table 3.

In Table 3 we can observe Lindley's paradox. As we seemingly place less prior information on the component means (i.e., as $\kappa \to 0$), the prior becomes more informative for the number of components. Jennison (1997) noted that placing a prior on κ can help to reduce this effect (as used by Stephens, 2000a) and presumably this is one of the reasons why Richardson and Green (1997) placed a prior on β. In our opinion Lindley's paradox demonstrates that there is generally no natural way to represent "prior ignorance" in this hierarchical modeling context. That is, it will often be the case that many reasonable prior specifications (for the parameters) lead to differing inferences with respect to the number of components. This latter point was made by Aitkin (2001).

TABLE 3
Sensitivity of the posterior distribution for k for the Galaxy data

	Range of k with		
κ	$p(k\|\mathbf{x}) \geq 0.05$	$p(k\|\mathbf{x}) \geq 0.001$	$\max_k p(k\|\mathbf{x})$
$\frac{1}{R^2}$	3–9	3–15	6
$\frac{4}{R^2}$	5–13	3–24	8
$\frac{9}{R^2}$	8–16	4–30	10
$\frac{16}{R^2}$	11–18	5–30	14
$\frac{25}{R^2}$	8–22	5–30	19
$\frac{64}{R^2}$	21–30	6–30	30
$\frac{100}{R^2}$	23–30	2–30	30

NOTE: Each row is based on a reversible jump sampler run for 500,000 iterations with a 100,000 iteration burn-in. The priors were set to be $\alpha = 2$, $\delta = 1$, $\xi = M$, $g = 0.2$, $h = 100g/\alpha R^2$ and $k_{\max} = 30$, where R and M denote the range and midpoint of the observed data.

7.2 Further Discussion

Aitkin (2001) compiled a set of Bayesian analyses of the Galaxy data, ranging from the approach of Richardson and Green (1997) to Escobar and West (1995). Aitken noted that although each method has a sensible (weakly informative) prior specification, the posterior distributions for the number of components are markedly different. Indeed Aitkin (2001) noted:

> The complexity of the prior structures needed for Bayesian analysis, the obscurity of their interaction with the likelihood, and the widely different conclusions they lead to over different specifications, leave the user completely unclear about "what the data say" from a Bayesian point of view about the number of components in the mixture.

This is a traditional frequentist criticism of Bayesian inferential methods, but does illustrate the need for appropriate consideration of all elements of a Bayesian

model. Of course, these concerns are only really of interest when performing clustering or discrimination analysis. In the case of the latter, Stephens (2000a) noted that priors should be set so that components are as different as possible so as to avoid overfitting (too many components). When the objective of inference is prediction (e.g., computing density estimates), model averaging procedures may be used, because the influence of the prior on the number of components is of less concern.

We believe that meaningful data analysis can be performed using Bayesian mixture models, but an appreciation of the effect of the prior on the number of components is needed. Our general practice is to set priors according to the information available (e.g., data dependent as in Richardson and Green, 1997) and then to perform a sensitivity analysis to measure the influence of the prior specification on the number of components, especially using simulated data. Once we are certain we understand the implications of the prior on the posterior, we proceed with a data analysis, accepting that different priors will lead to different posteriors.

Constructing priors for Bayesian mixture models is an area which still requires further research. We recommend the discussion of Richardson and Green (1997) and Stephens (2000a) as possible starting points.

8. SUMMARY AND RECOMMENDATIONS

In this article we have reviewed MCMC samplers for mixtures, posed solutions to the label switching problem and discussed the sensitivity of posterior inference for the number of mixture components to prior specifications.

To construct an MCMC sampler for fixed dimensional, univariate mixtures, we generally use the following strategy. We begin by coding a Metropolis–Hastings algorithm without completing the data (if we can compute the marginal likelihood). We then seek to find the proposal densities and parameters that provide reasonable mixing (i.e., autocorrelations in the chain not too large). If the sampler is unable to move around the symmetric modes of the target, we consider a couple of ideas.

First, if we are unsure as to the number of components in the mixture, we add reversible jump steps, which normally provides adequate mixing. Second, we try to use either tempering or evolutionary Monte Carlo. We have found that tempering is often effective, but in some difficult situations (highly separated modes in the posterior) we were often unable to tune the tempering sampler to enable the correct movement around the target space. In our experience, EMC works extremely well and we have never found an example where a properly tuned algorithm does not traverse the state space correctly.

We feel that the Gibbs sampler run with completion is often not worth programming (unless it can be quickly implemented, in BUGS, e.g.), since the chance of it failing to converge is too high.

To choose a method to deal with label switching, we have used the following criteria in our applied work. In situations for which we are only interested in a single inference from our mixture model (e.g., clustering the data), we often use a label invariant loss function. This choice is made because it often needs less programming effort than a relabelling algorithm. Conversely, if we are concerned with many label dependent quantities, we prefer the relabelling approach because it avoids performing a large number of simulated annealing algorithms. In our experience, ICs (not through relabelling algorithms) are only of use in situations where it is obvious how to undo the label switching [e.g., Figure 5(b)].

In this article we have detailed the progress of Bayesian mixture modeling: So what challenges need to be addressed in the future? One area of current research in bioinformatics is gene clustering (see, e.g., Yeung, Fraley, Murua, Raftery and Ruzzo, 2001), which can be performed using mixture models. A drawback of using a Bayesian mixture model is the difficulty of simulating from a high-dimensional, variable-dimensional target measure, which is characteristic of such problems (an example of a small data set is 2000 data points in six dimensions). Current reversible jump and continuous time samplers are unable to move efficiently around the sample space and new simulation methods are required to apply Bayesian methodology in such contexts; see Jasra, Stephens and Holmes (2005) for a potential approach.

ACKNOWLEDGMENTS

The first author was supported by a UK Engineering and Physical Sciences Research Council studentship. The second author is partially supported by the Medical Research Council (MRC). We also thank an Executive Editor and three referees for their comments on earlier versions.

REFERENCES

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

BARTLETT, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44** 533–534.

BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37** 1554–1563.

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171.

BEAL, M. J., GHAHRAMANI, Z. and RASMUSSEN, C. E. (2002). The infinite hidden Markov model. In *Neural Information Processing Systems* **14** (T. G. Diettrich, S. Becker and Z. Ghahramani, eds.) 577–584. MIT Press, Cambridge, MA.

BERNARDO, J. M. and GIRÒN, F. J. (1988). A Bayesian analysis of simple mixture problems. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 67–78. Oxford Univ. Press.

BOYS, R. J. and HENDERSON, D. A. (2003). Data augmentation and marginal updating schemes for inference in hidden Markov models. Technical report, Univ. Newcastle.

BOYS, R. J. and HENDERSON, D. A. (2004). A Bayesian approach to DNA sequence segmentation (with discussion). *Biometrics* **60** 573–588.

CAPPÉ, O., ROBERT, C. P. and RYDÉN, T. (2001). Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. Technical report, Univ. Paris Dauphine.

CAPPÉ, O., ROBERT, C. P. and RYDÉN, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 679–700.

CASELLA, G., MENGERSEN, K. L., ROBERT, C. P. and TITTERINGTON, D. M. (2002). Perfect samplers for mixtures of distributions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 777–790.

CELEUX, G. (1997). Discussion of "On Bayesian analysis of mixtures with an unknown number of components," by S. Richardson and P. J. Green. *J. Roy. Statist. Soc. Ser. B* **59** 775–776.

CELEUX, G. (1998). Bayesian inference for mixtures: The label-switching problem. In *COMPSTAT 98—Proc. in Computational Statistics* (R. Payne and P. J. Green, eds.) 227–232. Physica, Heidelberg.

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

CIUPERCA, G., RIDOLFI, A. and IDIER, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scand. J. Statist.* **30** 45–59.

DELLAPORTAS, P. and PAPAGEORGIOU, I. (2004). Multivariate mixtures of normals with an unknown number of components. Technical report, Athens Univ.

DELLAPORTAS, P., STEPHENS, D. A., SMITH, A. F. M. and GUTTMAN, I. (1996). A comparative study of perinatal mortality using a two-component mixture model. In *Bayesian Biostatistics* (D. A. Berry and D. K. Stangl, eds.) 601–616. Dekker, New York.

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

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

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588.

FEARNHEAD, P. (2004). Exact and efficient Bayesian inference for multiple changepoint problems. Technical report, Univ. Lancaster.

FRÜHWIRTH-SCHNATTER, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Amer. Statist. Assoc.* **96** 194–209.

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

GREEN, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 179–196. Oxford Univ. Press.

GREEN, P. J. and RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.* **97** 1055–1070.

GRUET, M.-A., PHILIPPE, A. and ROBERT, C. P. (1999). MCMC control spreadsheets for exponential mixture estimation. *J. Comput. Graph. Statist.* **8** 298–317.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

HURN, M., JUSTEL, A. and ROBERT, C. P. (2003). Estimating mixtures of regressions. *J. Comput. Graph. Statist.* **12** 55–79.

JASRA, A., STEPHENS, D. A. and HOLMES, C. C. (2005). Population-based reversible jump Markov chain Monte Carlo. Technical report, Imperial College London.

JENNISON, C. (1997). Discussion of "On Bayesian analysis of mixtures with an unknown number of components," by S. Richardson and P. J. Green. *J. Roy. Statist. Soc. Ser. B* **59** 778–779.

LIANG, F. and WONG, W. H. (2001). Real parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Statist. Assoc.* **96** 653–666.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* **44** 187–192.

LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

MARIN, J.-M., MENGERSEN, K. L. and ROBERT, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In *Bayesian Modelling and Inference on Mixtures of Distributions. Handbook of Statistics* **25** (D. Dey and C. R. Rao, eds.). North-Holland, Amsterdam.

MCLACHLAN, G. J. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, Chichester.

MENGERSEN, K. L. and ROBERT, C. P. (1996). Testing for mixtures: A Bayesian entropic approach (with discussion). In *Bayesian Statistics 5* (J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith, eds.) 255–276. Oxford Univ. Press.

NEAL, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statist. Comput.* **4** 353–366.

NEWCOMB, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *Amer. J. Math.* **8** 343–366.

PEARSON, K. (1894). Contribution to the mathematical theory of evolution. *Philos. Trans. Roy. Soc. London Ser. A* **185** 71–110.

POSTMAN, M., HUCHRA, J. P. and GELLER, M. J. (1986). Probes of large-scale structure in the Corona Borealis region. *Astronomical J.* **92** 1238–1246.

RAO, C. R. (1948). The utilization of multiple measurements in problems of biological classification (with discussion). *J. Roy. Statist. Soc. Ser. B* **10** 159–203.

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

ROBERT, C. P. (1997). Discussion of "On Bayesian analysis of mixtures with an unknown number of components," by S. Richardson and P. J. Green. *J. Roy. Statist. Soc. Ser. B* **59** 758–764.

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York.

ROBERT, C. P., RYDÉN, T. and TITTERINGTON, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 57–75.

ROEDER, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. Amer. Statist. Assoc.* **85** 617–624.

STEPHENS, M. (1997a). Bayesian methods for mixtures of normal distributions. D.Phil. dissertation, Dept. Statistics, Univ. Oxford.

STEPHENS, M. (1997b). Discussion of "On Bayesian analysis of mixtures with an unknown number of components," by S. Richardson and P. J. Green. *J. Roy. Statist. Soc. Ser. B* **59** 768–769.

STEPHENS, M. (2000a). Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *Ann. Statist.* **28** 40–74.

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

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

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester.

WEST, M. (1997). Discussion of "On Bayesian analysis of mixtures with an unknown number of components," by S. Richardson and P. J. Green. *J. Roy. Statist. Soc. Ser. B* **59** 783–784.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. and RUZZO, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17** 977–987.