J. R. Statist. Soc. B (2000)
62, *Part* 4, *pp.* 795–809

Dealing with label switching in mixture models

Matthew Stephens

University of Oxford, UK

[Received February 1999. Final revision May 2000]

Summary. In a Bayesian analysis of finite mixture models, parameter estimation and clustering are sometimes less straightforward than might be expected. In particular, the common practice of estimating parameters by their posterior mean, and summarizing joint posterior distributions by marginal distributions, often leads to nonsensical answers. This is due to the so-called 'label switching' problem, which is caused by symmetry in the likelihood of the model parameters. A frequent response to this problem is to remove the symmetry by using artificial identifiability constraints. We demonstrate that this fails in general to solve the problem, and we describe an alternative class of approaches, *relabelling algorithms*, which arise from attempting to minimize the posterior expected loss under a class of loss functions. We describe in detail one particularly simple and general relabelling algorithm and illustrate its success in dealing with the label switching problem on two examples.

Keywords: Bayesian approach; Classification; Clustering; Identifiability; Markov chain Monte Carlo methods; Mixture model; Multimodal posterior

1. Introduction

The so-called *label switching* problem arises when taking a Bayesian approach to parameter estimation and clustering using mixture models (see for example Diebolt and Robert (1994) and Richardson and Green (1997)). The term label switching was used by Redner and Walker (1984) to describe the invariance of the likelihood under relabelling of the mixture components. In a Bayesian context this invariance can lead to the posterior distribution of the parameters being highly symmetric and multimodal, making it difficult to summarize. In particular the usual practice of summarizing joint posterior distributions by marginal distributions, and estimating quantities of interest by their posterior mean, is often inappropriate. In this paper we summarize and unite some recently suggested solutions to this problem.

The structure of the paper is as follows. In the next section we introduce the notation, describe the label switching problem in more detail and illustrate the problem on an example data set. In Section 3 we demonstrate that the common strategy of removing label switching by imposing artificial *identifiability constraints* on the model parameters does not always provide a satisfactory solution, and we give a brief overview of other recent approaches to solving the problem: the relabelling algorithms suggested by Stephens (1997a, b) and Celeux (1998), and the decision theoretic approach considered by Celeux *et al.* (2000). Section 4 describes how the relabelling algorithms fit into a decision theoretic framework, and we use this to rederive a relabelling algorithm from Stephens (1997b) which is easily and widely

Address for correspondence: Matthew Stephens, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.
E-mail: stephens@stats.ox.ac.uk

 1369–7412/00/62795

applicable. The success of this algorithm in removing the label switching problem is demonstrated on some examples in Section 5, and results and extensions are discussed in Section 6.

2. The label switching problem

2.1. Notation

Let $\mathbf{x} = x_1, \ldots, x_n$ be independent observations from a mixture density with k (k assumed known and finite) components:

$$p(x|\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_1 f(x; \phi_1, \eta) + \ldots + \pi_k f(x; \phi_k, \eta), \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ are the *mixture proportions* which are constrained to be non-negative and to sum to 1; $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ are (possibly vector) *component-specific* parameters, with ϕ_j being specific to component j, η is a (possibly vector) parameter which is common to all components and f is a density. We write θ for the parameter vector $(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$.

It is sometimes convenient to assume that each observation x_i arose from an unknown component z_i of the mixture, where $z_1, \ldots, z_n$ are realizations of independent and identically distributed discrete random variables $Z_1, \ldots, Z_n$ with probability mass function

$$\Pr(Z_i = j|\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_j \qquad (i = 1, \ldots, n; j = 1, \ldots, k).$$

Conditional on the Zs, $x_1, \ldots, x_n$ are then independent observations from the densities

$$p(x_i|Z_i = j, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = f(x_i; \phi_j, \eta) \qquad (i = 1, \ldots, n).$$

A Bayesian approach to inference requires the specification of a prior distribution $p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ for the parameters of the mixture model. Inference is then based on the posterior distribution $p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta|\mathbf{x})$, and quantities of interest are calculated by integrating out the model parameters over the posterior distribution. For example, the marginal classification probabilities for an observation x_{n+1} are given by

$$\Pr(Z_{n+1} = j|x_{n+1}, \mathbf{x}) = \int \frac{\pi_j f(x_{n+1}; \phi_j, \eta)}{\sum_l \pi_l f(x_{n+1}; \phi_l, \eta)} p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta|x_{n+1}, \mathbf{x}) \, \mathrm{d}\boldsymbol{\pi} \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\eta \tag{2}$$

$$\propto \int \pi_j f(x_{n+1}; \phi_j, \eta) \, p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta|\mathbf{x}) \, \mathrm{d}\boldsymbol{\pi} \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\eta. \tag{3}$$

The accurate approximation of such integrals is now routine through the use of Markov chain Monte Carlo (MCMC) methods (see Gilks *et al.* (1996), for example).

2.2. The label switching problem

For any permutation ν of $1, \ldots, k$, define the corresponding permutation of the parameter vector θ by

$$\nu(\theta) = \nu(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = ((\pi_{\nu(1)}, \ldots, \pi_{\nu(k)}), (\phi_{\nu(1)}, \ldots, \phi_{\nu(k)}), \eta). \tag{4}$$

The root of the label switching problem is that the likelihood

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \{\pi_1 f(x_i; \phi_1, \eta) + \ldots + \pi_k f(x_i; \phi_k, \eta)\} \tag{5}$$

is the same for all permutations of θ. Some implications of this for likelihood analyses are

discussed by Redner and Walker (1984). In a Bayesian analysis, if we have no prior information that distinguishes between the components of the mixture, so our prior distribution $p(\boldsymbol{\pi}, \phi, \eta)$ is the same for all permutations of θ, then our posterior distribution will be similarly symmetric. This symmetry can cause problems when we try to estimate quantities which relate to individual components of the mixture. For example, by symmetry the predictive scaled component densities, given by the right-hand side of expression (3), are the same for every component, and so the marginal classification probabilities, given by equation (2), are $1/k$ for every observation. These classification probabilities are thus useless for clustering the observations into groups. Similarly, the posterior means of the component-specific parameters are the same for all components and are thus, in general, poor estimates of these parameters.

2.3. Example

We illustrate the label switching problem on a data set given by Roeder (1990). The data set consists of the velocities (in thousands of kilometres per second) of distant galaxies diverging from our own, sampled from six well-separated conic sections of the *corona borealis*. It has been analysed under a variety of mixture models by many researchers, including Crawford (1994), Chib (1995), Carlin and Chib (1995), Escobar and West (1995), Phillips and Smith (1996), Richardson and Green (1997) and Stephens (2000). A histogram of the 82 data points is shown in Fig. 1. For illustration we model the data as independent observations from a mixture of $k = 6$ univariate normal distributions:

$$p(x|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \pi_1 \mathcal{N}(x; \mu_1, \sigma_1^2) + \ldots + \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2) \tag{6}$$

where $\mathcal{N}(\cdot; \mu, \sigma^2)$ denotes the density function of the normal distribution with mean μ and variance σ^2.

We fitted model (6) using Gibbs sampling and the semiconjugate priors used by Richardson and Green (1997), where full details of the necessary computational steps can be found. The effects of label switching can be seen in the sampled values of the component means (Fig. 2). Distinct jumps occur in the traces of the means (Fig. 2(a)) as the MCMC scheme moves between relatively well-separated regions of parameter space. Intuitively these regions correspond to some of the 6! ways of labelling the mixture components. Estimates of the marginal posterior distributions of the means (Fig. 2(b)) are all very similar to one

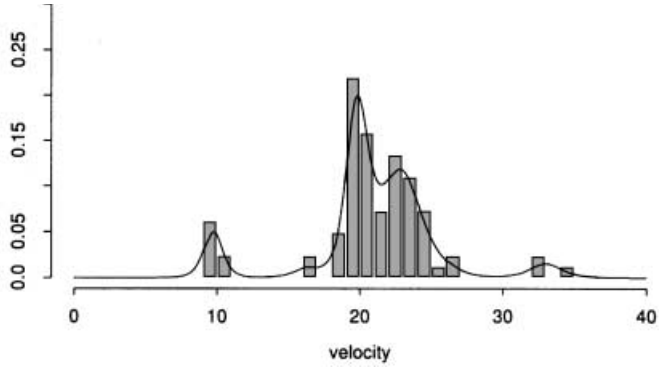

Fig. 1. Histogram of the galaxy data, with bin widths chosen by eye, overlaid with the predictive density estimate based on fitting a mixture of six normal components to the data

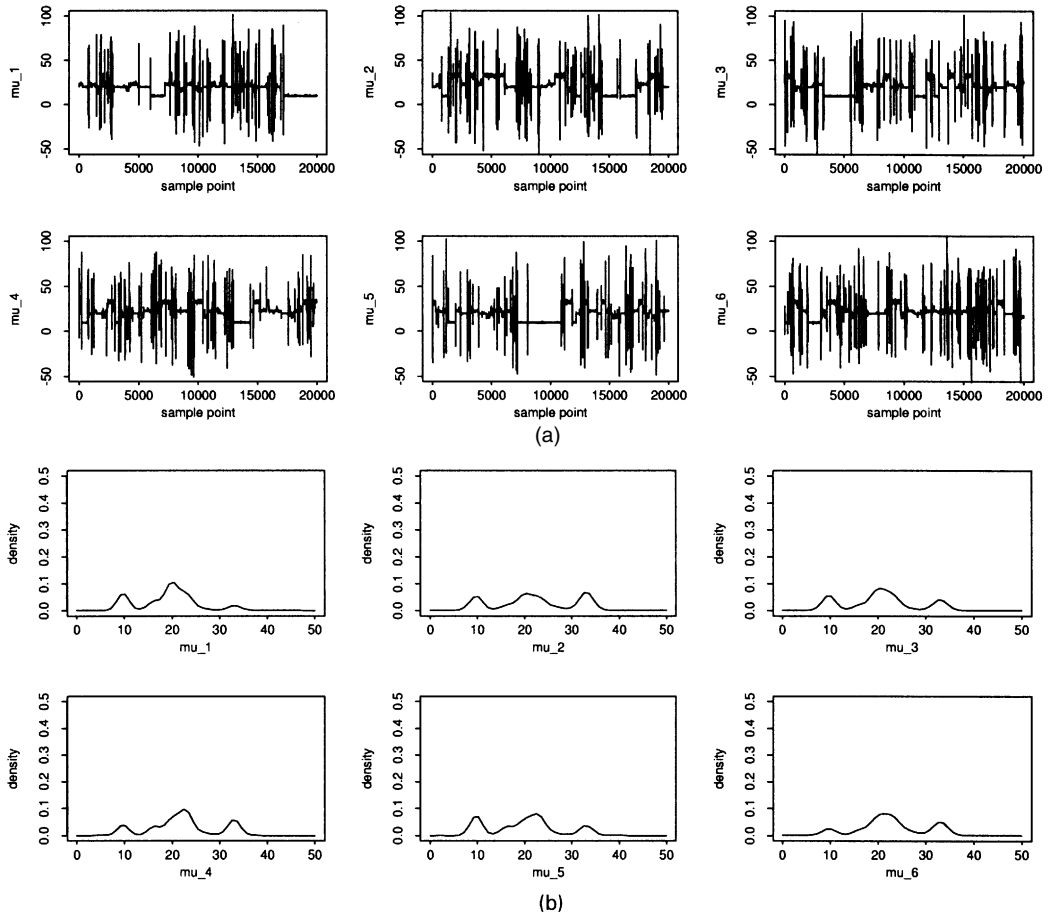

Fig. 2. Illustration of effects of label switching in the raw output of the Gibbs sampler when fitting a mixture of six normal distributions to the galaxy data: (a) traces of component means; (b) estimated marginal posterior densities of component means (the density estimates would all be exactly the same as each other if we ran the MCMC scheme for sufficiently long)

another, and so estimating the means on the basis of the MCMC output is not straightforward. Traces and density estimates for the mixture proportions and variances also behave in this way (data not shown). In contrast, an estimate of the predictive density based on the MCMC output (Fig. 1) is unaffected by the label switching problem, since it does not depend on how the components are labelled.

3. Previous work

3.1. Identifiability constraints

A common response to the label switching problem is to impose an *identifiability constraint* on the parameter space (such as $\pi_1 < \pi_2 < \ldots < \pi_k$ or $\mu_1 < \mu_2 < \ldots < \mu_k$) that can be satisfied by only one permutation of θ for each θ. This breaks the symmetry of the prior (and thus of the posterior) distribution of the parameters and so might seem to solve the label switching problem. However, for any given data set, many choices of identifiability constraint

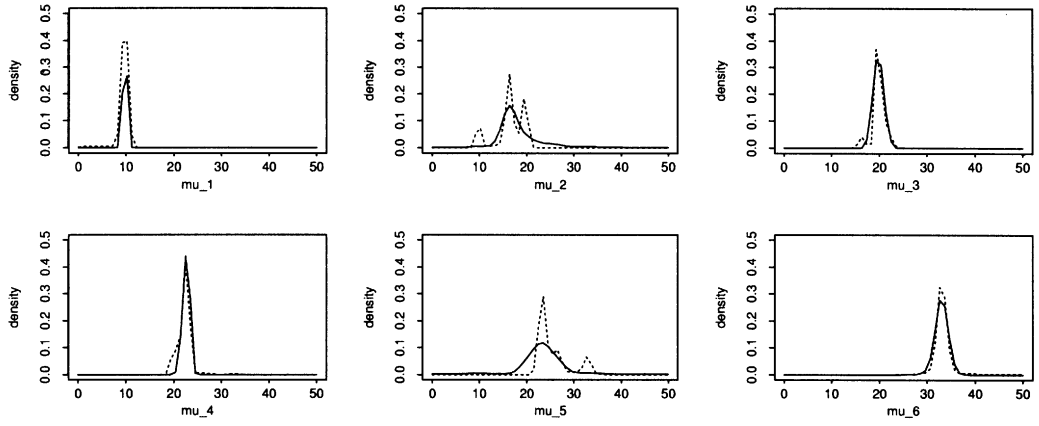

Fig. 3. Estimated posterior densities for the component means when fitting a mixture of six normal distributions to the galaxy data: ·····, obtained by imposing the identifiability constraint $\mu_1 < \mu_2 < \ldots < \mu_6$; ———, obtained by applying algorithm 2 to the raw output of the Gibbs sampler

will be ineffective in removing the symmetry in the posterior distribution. As a result, problems with label switching may remain after imposing an identifiability constraint if the constraint is not carefully chosen.

For example, suppose that we impose the constraint $\mu_1 < \mu_2 < \ldots < \mu_6$ in our example. This can be done by *relabelling* the MCMC sample $(\theta^{(1)}, \ldots, \theta^{(N)})$, applying permutations $\nu_1, \ldots, \nu_N$ such that the constraint is satisfied by the permuted sample $(\nu_1(\theta^{(1)}), \ldots, \nu_N(\theta^{(N)}))$ (see Stephens (1997b), proposition 3.1, for a formal justification). Fig. 3 shows estimates of the marginal posterior densities of the means based on the permuted sample. These densities continue to exhibit multimodality because much of the symmetry in the posterior distribution remains, the effect being most noticeable in the densities of μ_2 and μ_5. Imposing either of the other 'obvious' constraints, $\pi_1 < \pi_2 < \ldots < \pi_6$ or $\sigma_1^2 < \sigma_2^2 < \ldots < \sigma_6^2$, does not solve this problem (data not shown).

3.2. Other approaches

Despite these problems, few researchers have tried anything more sophisticated or devoted much attention to the label switching problem. Exceptions include Celeux *et al.* (1996) and Richardson and Green (1997). Celeux *et al.* (1996) suggest three methods for detecting label switching in simulation studies, but they all rely on the 'true' values of the parameters being known, making them difficult to apply to data analyses. Richardson and Green (1997) demonstrate that imposing different identifiability constraints (i.e. relabelling the MCMC output in different ways) can substantially alter views of the marginal posteriors for the parameters, and they advise that the MCMC output should be post processed according to different choices of labels to obtain the clearest picture of the component parameters. They suggest that labellings be chosen to order on the means, variances, mixture proportions or 'some combination of all three', without being more specific on what that might mean.

Stephens (1997a) suggests relabelling the MCMC output so that the marginal posterior distributions of parameters of interest are, as far as possible, unimodal and describes and demonstrates how this might be achieved in the context of normal mixtures. Other relabelling strategies (one of which is more generally applicable, and we consider in more

detail later) are investigated and compared in Stephens (1997b). Another generally applicable relabelling algorithm, which relabels the MCMC output 'on line' to reduce storage requirements, is mentioned in Celeux (1997) and detailed and demonstrated in Celeux (1998). Celeux *et al*. (2000) compare parameter estimates obtained by using this algorithm with those obtained from a more decision theoretic approach, using algorithms that aim to minimize the posterior expectation of some suggested loss functions. They found that these (apparently different) approaches produced similar results.

In fact each of the relabelling strategies described by Stephens (1997a, b) and Celeux (1998) also has a natural interpretation in terms of attempting to minimize the posterior expectation of some loss function, as we now describe.

4. Relabelling algorithms and the decision theoretic approach

Estimating parameters, clustering observations into groups and summarizing posterior distributions can all be viewed as a problem of choosing a single action a from a set of possible actions $\mathcal{A}$. The decision theoretic approach is to define a loss function $\mathcal{L}$: $\mathcal{A} \times \Theta \to R$, where $\mathcal{L}(a; \theta)$ is the loss incurred for choosing action a when the true value of the parameters is θ, and to choose the action $\hat{a}$ that minimizes the posterior expected loss (or *risk*)

$$\mathcal{R}(a) = E\{\mathcal{L}(a; \theta)|\mathbf{x}\}. \tag{7}$$

Since in the mixture context all permutations of θ give the same likelihood, it makes sense to consider loss functions that are invariant under permutation of θ. We choose to impose this invariance by restricting attention to loss functions of the form

$$\mathcal{L}(a; \theta) = \min_{\nu}[\mathcal{L}_0\{a; \nu(\theta)\}] \tag{8}$$

for some $\mathcal{L}_0$: $\mathcal{A} \times \Theta \to R$.

If $\theta^{(1)}, \ldots, \theta^{(N)}$ are sampled states (after burn-in) from a Markov chain with stationary distribution $p(\theta|\mathbf{x})$, then it is natural to approximate the risk $\mathcal{R}(a)$ by the *Monte Carlo risk*

$$\tilde{\mathcal{R}}(a) = \frac{1}{N}\sum_{t=1}^{N} \min_{\nu_t}[\mathcal{L}_0\{a; \nu_t(\theta^{(t)})\}] = \min_{\nu_1,\ldots,\nu_N}\left[\frac{1}{N}\sum_{t=1}^{N}\mathcal{L}_0\{a; \nu_t(\theta^{(t)})\}\right], \tag{9}$$

and to choose $\hat{a}$ to minimize $\tilde{\mathcal{R}}(a)$. Each iteration of the following algorithm reduces $\tilde{\mathcal{R}}(\hat{a})$:

Algorithm 1. Starting with some initial values for $\nu_1, \ldots, \nu_N$ (setting them all to the identity permutation for example), iterate the following steps until a fixed point is reached.

Step 1: choose $\hat{a}$ to minimize $\Sigma_{t=1}^{N}\mathcal{L}_0\{\hat{a}; \nu_t(\theta^{(t)})\}$.
Step 2: for $t = 1, \ldots, N$ choose ν_t to minimize $\mathcal{L}_0\{\hat{a}; \nu_t(\theta^{(t)})\}$.

The computational complexity of this algorithm will depend on the choice of $\mathcal{L}_0$, and some examples which lead to feasible algorithms are given later. For certain forms of $\mathcal{L}_0$, step 2 can be solved quickly even for large values of k (see Appendix A). The algorithm is guaranteed to reach a fixed point, as each iteration decreases $\tilde{\mathcal{R}}$ and there are only finitely many possible values for the permutations $\nu_1, \ldots, \nu_N$. As with all algorithms which are monotonic in the optimization criterion, the solution that is reached may depend on the starting-point, and there is no guarantee that the algorithm will converge to the global optimal solution. It is therefore prudent to run the algorithm from several starting-points and to choose the $\hat{a}$ that corresponds to the best local optimum found.

Each of the relabelling algorithms suggested in Stephens (1997a, b) can be viewed as a version of algorithm 1 for some action space and loss function. For example, the relabelling scheme described in Stephens (1997a) attempts to permute the sampled values of $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ so that they fit a distribution of the form

$$\mathcal{D}(\boldsymbol{\pi};\, \alpha) \prod_{i=1}^{k} \mathcal{IG}(\sigma_i^2;\, m_i,\, l_i)\, \mathcal{N}(\mu_i;\, u_i,\, \sigma_i^2/n_i), \tag{10}$$

where the hyperparameters (α, m, l, u, n) are to be estimated. ($\mathcal{D}(\cdot;\, \alpha)$ denotes the Dirichlet distribution, and $\mathcal{IG}(\cdot;\, m_i,\, l_i)$ denotes the inverse gamma distribution.) The algorithm suggested in Stephens (1997a) corresponds to algorithm 1 with the action being the estimation of (α, m, l, u, n), and

$$\mathcal{L}_0(\alpha, m, l, u, n;\, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = -\log\left\{\mathcal{D}(\boldsymbol{\pi};\, \alpha) \prod_{i=1}^{k} \mathcal{IG}(\sigma_i^2;\, m_i,\, l_i)\, \mathcal{N}(\mu_i;\, u_i,\, \sigma_i^2/n_i)\right\}. \tag{11}$$

Similarly, the obvious batch version of the on-line relabelling algorithm described by Celeux (1998) is a version of algorithm 1. It corresponds to the action of estimating the means $m = (m_1, \ldots, m_d)$ and variances $s = (s_1, \ldots, s_d)$ of the elements of the parameter vector $\theta = (\theta_1, \ldots, \theta_d)$, using

$$\mathcal{L}_0(m, s;\, \theta) = -\log\left\{\prod_{i=1}^{d} \mathcal{N}(\theta_i;\, m_i,\, s_i^2)\right\}. \tag{12}$$

That is, it is an algorithm that arises naturally if we aim to relabel the sample so that the marginal posterior distributions of the parameters look, as far as possible, normal and independent.

This decision theoretic interpretation of relabelling algorithms places them on a sound theoretical foundation. It also suggests some insight into which of the methods are appropriate for different situations. In particular the two algorithms corresponding to equations (11) and (12) seem most appropriate when primary interest focuses on the posterior distribution of the model parameters. Although this may be important in some contexts, in other contexts we are only indirectly interested in the parameters, for example to cluster the observations into groups.

4.1. A relabelling algorithm for clustering inference

Suppose that we wish to use our mixture model to cluster the observations into k groups, and to give some indication of the uncertainty involved in this clustering. A natural way to do this is to report an $n \times k$ matrix $Q = (q_{ij})$, where q_{ij} represents the probability that observation i is assigned to group j (so each row of Q sums to 1). If we interpret the rows of Q as being independent probability vectors, then Q corresponds to a distribution on k-group clusterings of the data.

Let $P(\theta)$ denote the matrix of classification probabilities $(p_{ij}(\theta))$, where

$$p_{ij}(\theta) = \Pr(Z_i = j | \mathbf{x}, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \frac{\pi_j f(x_i;\, \phi_j, \eta)}{\sum_l \pi_l f(x_i;\, \phi_l, \eta)}. \tag{13}$$

A natural way of measuring the loss for reporting Q when the true parameter values are θ is the Kullback–Leibler divergence from the true distribution on clusterings corresponding to $P(\theta)$, to the distribution on clusterings corresponding to Q:

$$\mathcal{L}_0(Q;\theta) = \sum_{z_1=1}^{k} \cdots \sum_{z_n=1}^{k} p_{1z_1}(\theta) \ldots p_{nz_n}(\theta) \log\left\{ \frac{p_{1z_1}(\theta) \ldots p_{nz_n}(\theta)}{q_{1z_1} \ldots q_{nz_n}} \right\}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}(\theta) \log\left\{ \frac{p_{ij}(\theta)}{q_{ij}} \right\}. \tag{14}$$

For this choice of $\mathcal{L}_0$, algorithm 1 becomes the following algorithm.

Algorithm 2. Starting with some initial values for $\nu_1, \ldots, \nu_N$ (setting them all to the identity permutation for example), iterate the following steps until a fixed point is reached.

Step 1: choose $\hat{Q} = (\hat{q}_{ij})$ to minimize

$$\sum_{t=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\{\nu_t(\theta^{(t)})\} \log\left[\frac{p_{ij}\{\nu_t(\theta^{(t)})\}}{\hat{q}_{ij}} \right]. \tag{15}$$

Step 2: for $t = 1, \ldots, N$ choose ν_t to minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\{\nu_t(\theta^{(t)})\} \log\left[\frac{p_{ij}\{\nu_t(\theta^{(t)})\}}{\hat{q}_{ij}} \right]. \tag{16}$$

It is straightforward to show that step 1 is achieved by

$$\hat{q}_{ij} = \frac{1}{N} \sum_{t=1}^{N} p_{ij}\{\nu_t(\theta^{(t)})\}. \tag{17}$$

Step 2 is most easily achieved by examining all $k!$ possibilities for each ν_t. (For moderate k, step 2 can be solved more efficiently, as described in Appendix A.)

Algorithm 2 was first derived in Stephens (1997b) using the less formal motivation of attempting to cluster together values of $\theta^{(t)}$ that correspond to the same way of labelling the components. The permutations $\nu_1, \ldots, \nu_N$ may be viewed as being chosen to ensure that the permuted sample points $\nu_1(\theta^{(1)}), \ldots, \nu_N(\theta^{(N)})$ all lie in the same symmetric mode. The right-hand side of equation (17) then approximates the posterior mean of the classification probabilities taken over just one of the $k!$ symmetric modes.

5. Examples

5.1. Fitting a mixture of six normal distributions to galaxy data

Returning to our example, we applied algorithm 2 to post-process all 20000 sample points generated by the Gibbs sampler. (In general we recommend discarding an initial burn-in sample before post-processing, but in this case convergence is so rapid that it makes little difference.) We applied algorithm 2 from 10 different starting-points: the first chosen by initializing all permutations to the identity (corresponding to using the raw output of the Gibbs sampler) and nine others chosen by selecting the initial permutations at random. We found that, although the optimum found depended on the starting-point used, all optima gave qualitatively very similar results.

For this example the computational requirements of algorithm 2 are moderate by today's standards. For example, the run that took the raw output of the Gibbs sampler as its starting-point took nine iterations and 18 min to converge (central processor unit times on a Sun UltraSparc 200 workstation, using the transportation algorithm to maximize over the

permutations, as described in Appendix A) and required of the order of $Nkn \approx 10^7$ bytes of storage in our implementation. Both the computational cost per iteration and the memory usage increase with N, k and n. Therefore, for larger data sets, with more components, or longer runs of the MCMC sampler, an on-line algorithm (along the lines of Celeux (1998)) would be preferable. Such an algorithm is given in Section 6.

We can use the $\hat{Q}$ found by using algorithm 2 to cluster the observations into groups by choosing the allocation variables z_i to maximize $\hat{Q}_{iz_i}$ $(i = 1, \ldots, n)$. The clustering obtained is shown in Fig. 4. Of course, the matrix $\hat{Q}$ contains information on the uncertainty associated with the assignment of each individual to its chosen cluster, which is not given in Fig. 4. If a single 'best' clustering is all that is required, a more direct approach would be to define a loss function on clusterings $\mathbf{z} = (z_1, \ldots, z_k)$, such as

$$\mathcal{L}_0(\mathbf{z}; \theta) = -\sum_{i=1}^{n} \log\{p_{iz_i}(\theta)\}, \tag{18}$$

for which algorithm 1 remains straightforward. It seems likely that similar results would be obtained.

It is also interesting to examine the relabelled sample, corresponding to the final values of $\nu_1, \ldots, \nu_N$ found by the algorithm. The label switching in evidence in the raw output of the

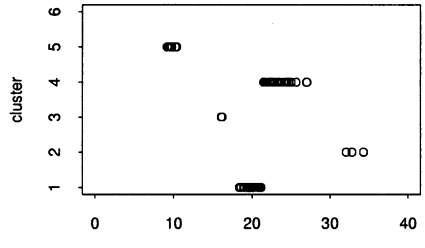

Fig. 4. Clustering of the galaxy data obtained by fitting a mixture of six normal components to the data and using the $\hat{Q}$ found by algorithm 2: the clustering consists of only five groups as no points are allocated to the sixth component

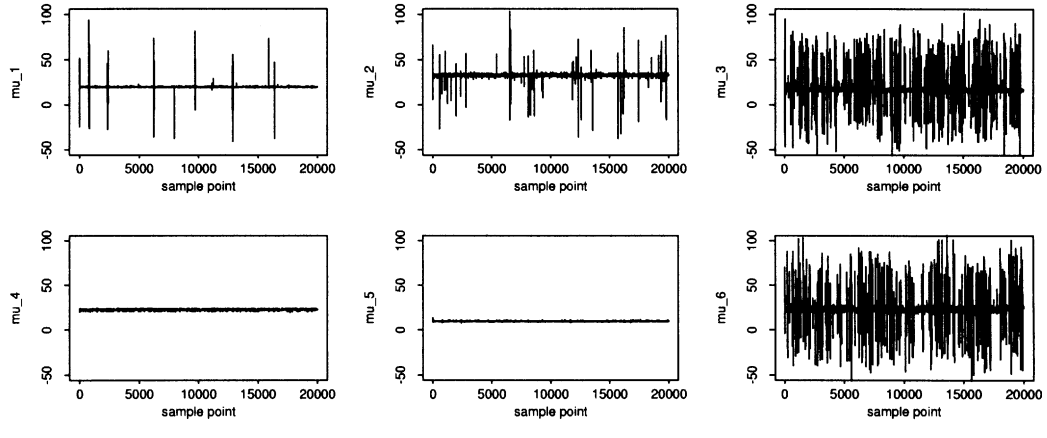

Fig. 5. Trace plots of the component means from the permuted MCMC sample obtained by using algorithm 2: label switching in evidence in the raw output of the Gibbs sampler (Fig. 2(a)) has been successfully removed

Gibbs sampler (Fig. 2(a)) appears to have been eliminated from the permuted sample (Fig. 5). As a result, most of the multimodality in the estimates of the marginal posterior distributions of the means has been removed (Fig. 3), making it straightforward to obtain sensible estimates of the component means. This illustrates an advantage of relabelling algorithms (i.e. those of the form of algorithm 1) over more general algorithms for minimizing more general loss functions, such as those considered by Celeux *et al.* (2000): the permuted sample obtained from a relabelling algorithm may be used to perform inference for *any* quantity of interest, and not just the quantity that was used to derive the algorithm. Under the decision theoretic view this is sensible only if the permuted sample that is obtained is reasonably similar for different (natural) choices of action space and loss function. In our experience (Stephens, 1997b) we have found that this is often so in problems where there is little 'genuine' multimodality (i.e. multimodality which cannot be explained by permuting the parameters) in the posterior distribution of the parameters. Our next example illustrates a situation where the posterior distribution exhibits obvious genuine multimodality.

5.2. *Fitting a mixture of three t_4-components to galaxy data*

We consider fitting a mixture of three t-distributions, each with 4 degrees of freedom, to the galaxy data. Details of the priors and corresponding Gibbs sampling steps used are given in Stephens (1997b). The Gibbs sampler was run for 20000 iterations, and the first $m = 10000$ sample points were discarded as burn-in. Traces of the remaining sampled values of the component means are shown in Fig. 6(a). It is reasonably straightforward to identify by eye that

(a) label switching occurs between the first and second component around iterations 16000 and 17000.
(b) there is genuine multimodality in the posterior distribution of the parameters, exemplified by the change in the means of the first two components, from means near 20 and 23 to means near 34 and 21, for about 300 iterations around iteration 14000.

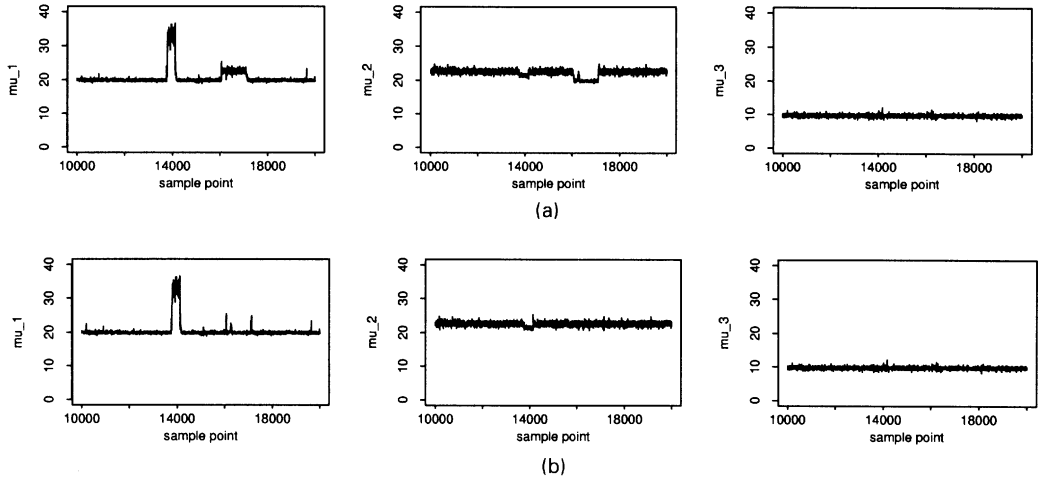

Fig. 6. Sampled values of means of the three components when fitting a mixture of three t_4-distributions to the galaxy data: (a) raw output of the Gibbs sampler; (b) permuted sample from algorithm 2

In this case it would be straightforward to undo the label switching by eye, or by imposing the identifiability constraint $\mu_1 < \mu_2 < \mu_3$. Nevertheless, it is reassuring that using algorithm 2 to post-process the 10000 sample points also successfully undoes the label switching (Fig. 6(b)) and retains a clear picture of the two genuine modes in the distribution. Furthermore, applying algorithm 2 using random permutations of the Gibbs output as a starting-point gave very similar results (data not shown), demonstrating that the algorithm was effectively able to undo the label switching separately for each genuine mode in this case. (A different relabelling algorithm which we tried on this example (see Stephens (1997b)) did not cope so well with the genuine multimodality.)

Since it makes little sense to perform clustering by averaging over the quite different genuine modes, which clearly represent quite different views of the data, we consider the two modes separately. We divided the sample by eye, based on the means of the permuted sample (Fig. 6(b)), into a *minor mode* that consisted of 326 sample points (13786–14111) and a *major mode* that consisted of the remaining 9674 sample points. Estimates of the scaled predictive component densities and corresponding clusterings of the points into three groups, based on the permuted sample, are shown in Fig. 7. The probability mass of each mode can be estimated by the proportion of sample points lying in each mode, giving 0.0326 for the minor mode and 0.9674 for the major mode. However, the sampler mixes poorly between the modes, so more simulation would be required for accurate estimates of these quantities.

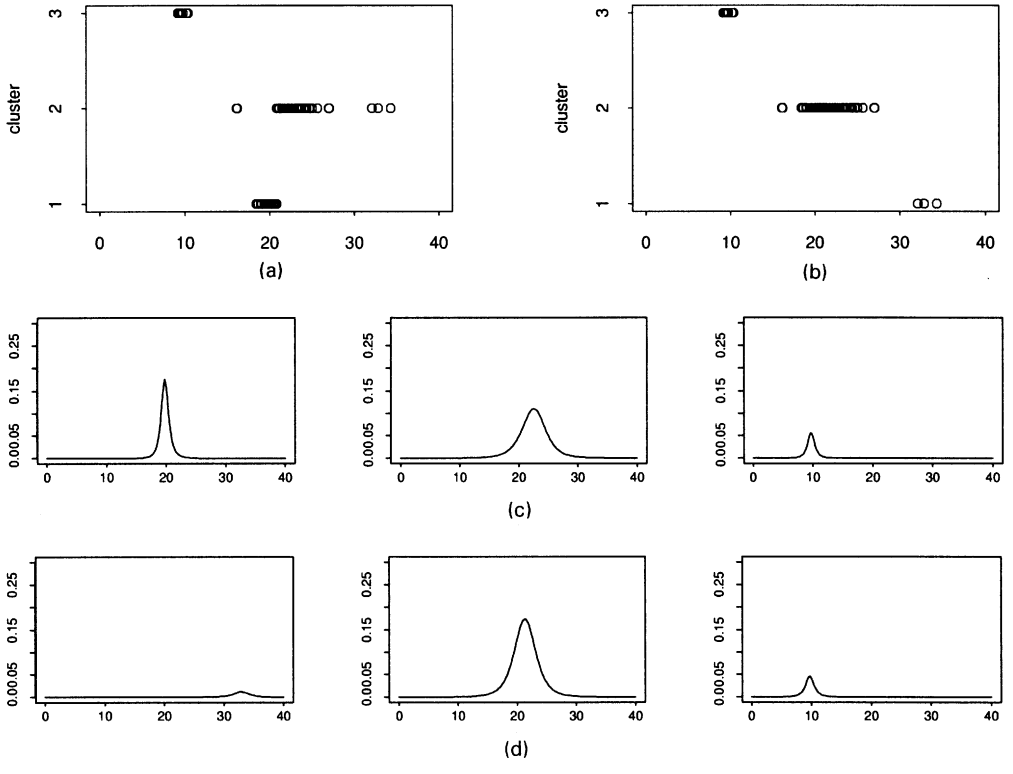

Fig. 7. Inference relating to individual components when fitting a mixture of three t_4-distributions to the galaxy data, separating the major mode and minor mode by eye as described in the text: (a) clustering of data, based on the major mode; (b) clustering of data, based on the minor mode; (c) scaled predictive component densities (3), based on the major mode; (d) scaled predictive component densities (3), based on the minor mode

Although in this case genuine multimodality was detectable by eye, in more complex situations it would be helpful to detect genuine multimodality automatically. One approach which we are currently investigating is to assume the presence of M genuine modes, and to summarize information in the posterior by the 'action' of reporting values $\xi = (\xi_1, \xi_2, \ldots, \xi_M)$ for the relative weights of the modes (with $\Sigma_m \xi_m = 1$), and corresponding classification matrices $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_M)$. Under the loss function

$$\mathcal{L}\{(\xi, \mathbf{Q}); \theta\} = \min_m \{-\log(\xi_m) + \mathcal{L}(Q_m; \theta)\}, \tag{19}$$

where $\mathcal{L}(Q_m; \theta)$ corresponds to equation (14), a natural extension of algorithm 2 can be used to find a locally optimal $(\xi, \mathbf{Q})$: simply add a step that allocates each $\theta^{(t)}$ to the mode $m(t)$ that minimizes $-\log(\xi_{m(t)}) + \mathcal{L}_0\{Q_{m(t)}, \nu_t(\theta^{(t)})\}$, add a step that estimates $\xi_1, \xi_2, \ldots, \xi_M$ by the proportion of points allocated to each mode and perform steps 1 and 2 of the existing algorithm conditionally on the current values of the $m(t)$. Informally this will cluster the sampled values of θ into M groups, each representing a genuine mode of the posterior distribution of the parameters. The choice of M could be done either manually by examining the results obtained for different M or more formally by adding to the loss function a penalty term that depends only on M, although it is not clear what form this penalty term should take. Preliminary experience suggests that the answers obtained by using this approach can be very sensitive to the choice of starting-point of the algorithm, so performing several runs with different starting-points can be important. Similar methods for identifying multimodality may also prove helpful for summarizing information in the posterior distribution in contexts where the label switching problem does not arise.

6. Discussion and extensions

Our examples demonstrate that algorithm 2 provides a more generally satisfactory solution to the label switching problem than imposing obvious identifiability constraints. In our experience, applying different relabelling algorithms, corresponding to different choices of loss function and action space, often leads to similar results (see Stephens (1997b), for example). This is consistent with the results reported in Celeux *et al.* (2000). Algorithm 2 has several appealing features that make it our favourite relabelling algorithm. From a theoretical viewpoint we prefer clustering-motivated approaches to those motivated by parameter estimation, as (while admitting that there will be exceptions) clustering the observations into distinct groups seems a more natural aim in many applications. More importantly, from a practical viewpoint it is simple to implement and is very general: for example it can be applied to any finite mixture model where the component densities can be calculated. In particular, it is straightforward to apply the algorithm to mixtures whose component densities have very high dimension, where imposing identifiability constraints is likely to be more problematic than for the simple univariate mixtures that we considered in our examples. (See Stephens (1997b) for an example of application of the algorithm to bivariate normal distributions.)

6.1. Label switching in other contexts

Although algorithm 2 is very general, there are contexts other than finite mixture models in which the label switching problem occurs, and for which algorithm 2 is inappropriate. For example, in a Bayesian analysis of a hidden Markov model (e.g. Robert *et al.* (2000)) the likelihood is invariant under permutation of the labels of the hidden states. If suitably symmetric prior distributions are used then the label switching problem arises when attempt-

ing to infer, for example, which hidden state z_i gave rise to observation x_i ($i = 1, \ldots, n$). The derivation of algorithm 2 relied on the z_i being independent given parameters θ, which is usually not true in a hidden Markov model. We can deal with this by defining a loss $\mathcal{L}(Q; \mathbf{z})$ for reporting Q when the true clustering of the observations is $\mathbf{z}$ (this replaces defining the loss $\mathcal{L}(Q; \theta)$ for reporting Q when the true value of the parameters is θ.) A natural choice corresponds to

$$\mathcal{L}_0(Q; \mathbf{z}) = -\sum_{i=1}^{n} \log(q_{iz_i}). \tag{20}$$

This leads to a straightforward algorithm which can be applied to any classification or clustering problem where we can construct a Markov chain with stationary distribution $\Pr(\mathbf{z}|\mathbf{x})$ — even when the number of clusters is allowed to vary (e.g. Richardson and Green (1997) and Stephens (2000)), although in this case problems with genuine multimodality are likely to arise.

6.2. An on-line algorithm

Algorithm 2 can be computationally demanding on storage. In our implementation we stored N matrices of classification probabilities, each consisting of nk numbers, and in many modern applications of MCMC sampling N is required to be rather large (though thinning the chain would reduce this burden). Storage requirements could be reduced, at the expense of computational speed, by storing $\theta^{(1)}, \ldots, \theta^{(N)}$, and recalculating the classification matrices $P(\theta^{(1)}), \ldots, P(\theta^{(N)})$ each time that they are needed. None-the-less, for some problems this approach could still require very large amounts of storage, and so it is helpful to note that, following Celeux (1998), we can define an on-line version of algorithm 2 as follows.

Algorithm 3. If at stage t the MCMC sample produces $\theta^{(t)}$, and the current estimate of $\hat{Q}$ is $\hat{Q}^{(t-1)} = (\hat{q}_{ij}^{(t-1)})$,

(a) choose ν_t to minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\{\nu_t(\theta^{(t)})\} \log\left[\frac{p_{ij}\{\nu_t(\theta^{(t)})\}}{\hat{q}_{ij}^{(t-1)}}\right], \tag{21}$$

(b) set

$$\hat{Q}^{(t)} = \frac{t\hat{Q}^{(t-1)} + P\{\nu_t(\theta^{(t)})\}}{t+1}. \tag{22}$$

By analogy with Celeux (1998) we might choose the starting-point $\hat{Q}^{(0)}$ by using a small preliminary MCMC sample in which it is assumed that no label switching occurs. A more reliable (though less convenient) approach would be to choose the starting-point by applying the batch algorithm (algorithm 2) to a small preliminary MCMC sample. We have used algorithm 3 to undo label switching successfully in cases where there is little genuine multimodality (data not shown).

Acknowledgements

This work was undertaken during the tenure of an Engineering and Physical Sciences Research Council studentship, and I gratefully acknowledge the help and advice of my supervisor Professor Brian Ripley. The final form of this paper was greatly improved by

useful discussions with Gilles Celeux and the helpful comments of two referees and the Associate Editor.

Appendix A

If the loss function $\mathcal{L}_0$ is of the form

$$\mathcal{L}_0(a; \theta) = \sum_{j=1}^{k} \mathcal{L}_0^{(j)}(a; \pi_j, \phi_j, \eta),$$

then step 2 of algorithm 1 can be solved quickly, even for large values of k, as follows. It consists of N minimization problems of the form

$$\text{choose } \nu \text{ to minimize } \sum_{j=1}^{k} c\{j, \nu(j)\} \tag{23}$$

where

$$c(j, l) = \mathcal{L}_0^{(j)}(a; \pi_l, \phi_l, \eta).$$

In particular, for the loss function (14) that produced algorithm 2 we have

$$c(j, l) = \sum_{i=1}^{n} p_{il}\{\nu_t(\theta^{(t)})\} \log\left[\frac{p_{il}\{\nu_t(\theta^{(t)})\}}{q_{ij}}\right]. \tag{24}$$

Problem (23) is equivalent to the integer programming problem

$$\text{choose } \{y_{jl}\}\ (j = 1, \ldots, k; l = 1, \ldots, k) \text{ to minimize } \sum_{j=1}^{k}\sum_{l=1}^{k} y_{jl}\, c(j, l)$$

$$\text{subject to } y_{jl} = 0 \text{ or } y_{jl} = 1 \text{ and } \sum_{j=1}^{k} y_{jl} = \sum_{l=1}^{k} y_{jl} = 1, \tag{25}$$

with the correspondence between the problems being, if $\{\hat{y}_{jl}\}$ is an optimal solution to problem (25), then the corresponding optimal solution to problem (23) is $\hat{\nu}(j) = l$ if and only if $\hat{y}_{jl} = 1$.

Problem (25) is a special version of the transportation problem, known as the *assignment problem*, for which efficient algorithms exist (see for example Taha (1989), page 195). We used a Numerical Algorithms Group Fortran routine to solve this problem.

References

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

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

———(1998) Bayesian inference for mixtures: the label-switching problem. In *COMPSTAT 98* (eds R. Payne and P. Green), pp. 227–232. Heidelberg: Physica.

Celeux, G., Chauveau, D. and Diebolt, J. (1996) Stochastic versions of the EM algorithm: an experimental study in the mixture case. *J. Statist. Computn Simuln*, **55**, 287–314.

Celeux, G., Hurn, M. and Robert, C. P. (2000) Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Ass.*, to be published.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *J. Am. Statist. Ass.*, **90**, 1313–1321.

Crawford, S. L. (1994) An application of the Laplace method to finite mixture distributions. *J. Am. Statist. Ass.*, **89**, 259–267.

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

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

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Phillips, D. B. and Smith, A. F. M. (1996) Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), ch. 13, pp. 215–239. London: Chapman and Hall.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195–239.

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

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. Statist. Soc.* B, **62**, 57–75.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Am. Statist. Ass.*, **85**, 617–624.

Stephens, M. (1997a) Discussion on 'On Bayesian analysis of mixtures with an unknown number of components' (by S. Richardson and P. J. Green). *J. R. Statist. Soc.* B, **59**, 768–769.

———(1997b) Bayesian methods for mixtures of normal distributions. *DPhil Thesis*. University of Oxford, Oxford. (Available from `http://www.stats.ox.ac.uk/~stephens`.)

———(2000) Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.*, **28**, in the press.

Taha, H. A. (1989) *Operations Research: an Introduction*, 4th edn. New York: Macmillan.