Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models

M. Sperrin T. Jaki E. Wit

May 6, 2009

Abstract

The label switching problem is caused by the likelihood of a Bayesian mixture model being invariant to permutations of the labels. The permutation can change multiple times between Markov Chain Monte Carlo (MCMC) iterations making it difficult to infer component-specific parameters of the model. Various so-called 'relabelling' strategies exist with the goal to 'undo' the label switches that have occurred to enable estimation of functions that depend on component-specific parameters. Most existing approaches rely upon specifying a loss function, and relabelling by minimising its posterior expected loss. In this paper we develop probabilistic approaches to relabelling that allow estimation and incorporation of the uncertainty in the relabelling process. Variants of the probabilistic relabelling algorithm are introduced and compared to existing loss function based methods. We demonstrate that the idea of probabilistic relabelling can be expressed in a rigorous framework based on the EM algorithm.

Keywords: Bayesian, Identifiability, Label Switching, MCMC, Mixture Model.

1 Introduction

Mixture models have been used as tools to model heterogeneity for over 100 years. Developments in Markov Chain Monte Carlo (MCMC) methods [e.g. Diebolt and Robert, 1994] opened the door for mixture models in a Bayesian framework as they allow efficient exploration of posterior and predictive surfaces of these models. The use of these Bayesian mixture models has given rise to new problems, particularly when estimating component-specific parameters of the model and interpreting marginal posterior densities.

The label switching problem arises as the components of the Bayesian mixture model can be ordered arbitrarily. During one run of an MCMC sampler, the order of components can change multiple times between iterations. To obtain a meaningful interpretation of the components it is necessary to account for these changes, which has been called *relabelling* [e.g. Stephens, 2000]. Various functions of interest, such as recovery of the full mixture posterior and its associated moments, may be invariant to the labelling permutations. For this type of inference, the label switching problem need not concern us. On many occasions, however, it is of interest to infer parameters that are specific to individual components of the mixture model. This may be because the components of the model have some interpretation, in the sense of a one-to-one correspondence to true components in the population, or alternatively we may be using mixture models to carry out semi-parametric density estimation, and the purpose of the relabelling is to provide coherent estimates of the components that make up the density estimate. In either case, we must find methods to 'relabel' the results of an MCMC run so that the components are in the same order at each iteration.

A wide array of strategies exist in the literature for 'relabelling' MCMC output in an attempt to remove the label switching problem — we divide them here into three categories. *Identifiability constraints* involve relabelling the output of the MCMC sampler so that the posterior obtained satisfies a constraint on the component parameters. The constraint is chosen such that exactly one relabelling satisfies the constraint at each iteration of the sampler. *Deterministic relabelling algorithms* select a relabelling at each iteration of the MCMC output that minimises the posterior expectation of some loss function. Naturally, a variety of loss functions have been considered by different authors. *Probabilistic* approaches are a relatively new idea in which one acknowledges that there is uncertainty in the relabelling that should be selected on each iteration of the MCMC output. In contrast, both identifiability constraint and deterministic relabelling algorithms assume that the relabelling that has been carried out is 'correct'.

The contribution of this paper is to develop and extend the idea of probabilistic relabelling, which was introduced originally in Jasra [2005]. We frame probabilistic relabelling as an application of the EM algorithm, where the missing data is the order that the components are in at each iteration of the MCMC. Two novel probabilistic algorithms based on the stochastic EM (SEM) are developed.

We will proceed, in Section 2, by briefly describing some of the relabelling algorithms currently available, before we introduce new strategies for probabilistic relabelling. Section 3 evaluates the performance of the strategies on observed as well as simulated data. We conclude with a discussion of the advantages and disadvantages of the various methods and some future directions in Section 4.

2 Relabelling Strategies

Suppose n observations $y_1, \ldots, y_n$ are taken from a K-component mixture distribution where all the components have the same distributional form, with mixture-specific parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, global parameters $\boldsymbol{\eta}$ and mixing weights $\boldsymbol{\pi}$, summarised by $\boldsymbol{\gamma} = (\pi_1, \ldots, \pi_K; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K; \boldsymbol{\eta})$. The mixture distribution for a single observation Y_i is then given by

$$p\left(y_i|\boldsymbol{\gamma}\right) = \sum_{k=1}^{K} \pi_k f_k\left(y_i|\boldsymbol{\theta}_k, \boldsymbol{\eta}\right)$$

with $K \geq 1$, $\pi_k > 0$ $(k = 1, 2, \ldots, K)$, $\sum_{k=1}^{K} \pi_k = 1$ and $f_k(\cdot|\boldsymbol{\theta}_k, \boldsymbol{\eta})$ is a density function parametrised by $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}$. For convenience we introduce latent variables z_i, $i = 1, \ldots, n$, where $\{z_i = k\}$ indicates membership of the observation y_i to class k, with for $i = 1, \ldots, n$,

$$z_i \overset{\text{i.i.d.}}{\sim} \text{Multinomial}\left\{1, \left(\pi_1, \ldots, \pi_K\right)\right\}$$

Conditional on belonging to class k, observation Y_i will be distributed according to $f_k(\cdot|\boldsymbol{\theta}_k, \boldsymbol{\eta})$,

$$Y_i|(z_i = k) \sim f_k(\cdot|\boldsymbol{\theta}_k, \boldsymbol{\eta})$$

Each z_i is then an unknown categorical variable that denotes the subpopulation from which observation y_i originates. Bayesian inference can be conducted on such a model using MCMC (Diebolt and Robert [1994]). This proceeds, on each iteration r, by drawing a vector of component memberships $\boldsymbol{z}^{(r)}$, and parameter estimates $\boldsymbol{\gamma}^{(r)}$, from the posterior. Throughout this paper, for ease of illustration we will assume that each $f_k(\cdot)$ is a normal distribution with mean μ_k and variance σ_k^2. For the priors we will use the hierarchical 'random beta' model in Richardson and Green [1997], following their suggestions on the hyperparameter choices. For the number of components K we use a Poisson(1) prior as argued for in Nobile and Fearnside [2007].

Let S_K denote the set of all permutations on $\{1, 2, \ldots, K\}$. The label switching problem arises because the likelihood

$$p\left(\boldsymbol{y}|\boldsymbol{\gamma}\right) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \pi_{\nu(k)} f_{\nu(k)}\left(y_i|\boldsymbol{\theta}_{\nu(k)}, \boldsymbol{\eta}\right) \right\}$$

is identical for all $\nu \in S_K$. If exchangeable priors are used (containing no component-specific information) then the posterior has the same property, resulting in the posterior surface having $K!$ symmetric modes, each associated with a different labelling permutation $\nu \in S_K$. This is problematic because each iteration of the MCMC sampler r, $r = 1, \ldots, R$, has an associated permutation $\nu^{(r)} \in S_K$. Then for $r_1 \neq r_2$, it may be that $\nu^{(r_1)} \neq \nu^{(r_2)}$, i.e. the sampler can move from one mode to another between iterations. This makes an ergodic average estimate of a component-specific parameter, e.g.

$$\mathbb{E}[\theta_1] \approx \frac{1}{R} \sum_{r=1}^{R} \theta_1^{(r)} \tag{1}$$

somewhat meaningless. Indeed, if the chain is in equilibrium, then the estimate of $\mathbb{E}[\theta_k]$ should be the same for all k, since such a chain explores equally all the symmetric modes.

The idea of *relabelling* the MCMC output is to account for the permutations $\nu^{(r)}$, $r = 1, \ldots, R$, in such a way that an ergodic average estimate such as (1) is made meaningful. Of course, we generally have limited data, and can never say with certainty whether we truly have agreement $\nu^{(r_1)} = \nu^{(r_2)}$, for $r_1 \neq r_2$. Indeed, in our view the $\nu^{(r)}$s are themselves parameters with associated uncertainty. Define a *relabelled posterior*, $q(\cdot)$, as the posterior density that we obtain when we attempt to account for the permutations $\nu^{(r)}, r = 1, \ldots, R$ across the iterations of an MCMC sampler. This is not unique — firstly there are $K!$ versions of it that correspond to applying a permutation ν to the entire output of an MCMC to yield an equivalent answer. Secondly, we accept that it is not possible to find the 'correct' relabelled posterior due to the uncertainty in estimating the $\nu^{(r)}$s — we approximate this by the various relabelling methods considered in this paper. A version of the relabelled posterior is then useful when one conducts component-specific inference.

2.1 Identifiability Constraints

The first efforts to deal with the label-switching problem involve placing an *Identifiability Constraint* (IC) on the parameter space [e.g. McLachlan and Peel, 2000]. The idea is to define a restricted parameter space $\mathcal{A}$ such that there exists a unique permutation $\nu^* \in S_K$ that satisfies
$(\theta_{\nu^*(1)}, \ldots, \theta_{\nu^*(K)}) \in \mathcal{A}$, for component-specific parameters $\theta_k, k = 1, \ldots, K$. The simplest example in the normal distribution case is the constraint $\mu_1 < \mu_2 < \ldots < \mu_K$, or equally the same constraint on the mixture proportions, or the component variances. More sophisticated alternatives can be found, for example, in Marin et al. [2005].

This approach is simple and works well in many situations. Proposition 3.1 of Stephens [1997a] demonstrates that the relabelling for such a strategy can be carried out after the MCMC has run, provided the priors are exchangeable. Geweke [2007] notes that use of the IC leads, asymptotically in n, to the correct marginals for the true parameter vector $\boldsymbol{\theta}$ being recovered, provided $\boldsymbol{\theta} \in \mathcal{A}$. Nevertheless, for finite n it is found that the parameter estimates are 'pushed apart', that is the difference between the parameters of adjacent components is typically over-estimated [McLachlan and Peel, 2000]. This is a consequence of the fact that we are effectively imposing *a-priori* that the joint prior of $\boldsymbol{\theta}$ must satisfy the constraint, despite originally imposing exchangeable priors, suggesting we know nothing to distinguish the components of the mixture model. Moreover, it can be difficult to find a sensible subspace ,$\mathcal{A}$, when the mixture-specific parameters are multidimensional.

2.2 Deterministic Relabelling Algorithms

The idea of the relabelling algorithm is that we believe that the permutations $\nu^{(r_1)}$ and $\nu^{(r_2)}$ match (for $r_1 \neq r_2$; $r_1, r_2 \in \{1, 2, \ldots, R\}$) when a characteristic about the ${r_1}^{\text{th}}$ iteration under permutation ν^{r_1} is 'close' to that characteristic of the ${r_2}^{\text{th}}$ iteration under permutation ν^{r_2}. There is a vast literature on the application of such algorithms to the label switching problem, all considering different characteristics about each iteration on which to measure closeness, and how one does measure closeness. Stephens [1997a] and Celeux et al. [2000] give methods where the characteristic is the estimates of the parameters on each iteration r, $\boldsymbol{\theta}^{(r)}$. Stephens [2000] produces a method in which the characteristic is the matrix of allocation probabilities of the observations to each component of the mixture, $P^{(r)}$ whilst Nobile and Fearnside [2007] measure closeness in the allocation vector $Z^{(r)}$.

Call the characteristic on which we measure closeness C, and the measure of closeness between two characteristics at iterations r_1 and r_2 as $L(C^{(r_1)}, C^{(r_2)})$, which is a loss function that is large when the discrepancy between $C^{(r_1)}$ and $C^{(r_2)}$ is large. When we apply a permutation $\nu^{(r)}$ to iteration r we will write $\nu^{(r)}(C^{(r)})$.

We are not interested *per se* in pairwise closeness, but closeness of the characteristics across the entire MCMC sample, $\{C^{(1)}, \ldots, C^{(R)}\}$, as we wish the entire sample to be relabelled 'correctly'. To take this into account in an efficient manner, many of the relabelling algorithms adopt a K-means style approach, which can be described in a general manner as follows:

1. Choose C to minimise $\sum_{r=1}^{R} L\left\{C, \nu^{(r)}(C^{(r)})\right\}$. In the K-means analogy, view C as the centroids of the clusters. In common with this analogy, C is usually calculated as the ergodic average of the characteristics $C^{(r)}$, $r = 1, \ldots, R$.

2. For $r = 1, \ldots, R$ choose $\nu^{(r)}$ to minimize $L\left\{C, \nu^{(r)}(C^{(r)})\right\}$, which is equivalent to allocating the observations to the clusters. Stephens [2000] demonstrates that it is usually possible to achieve this quickly, using a variant of the transportation algorithm.

3. Repeat 1 and 2 until an optimal solution is reached.

The algorithm should be run from multiple starting positions (initial permutations of the MCMC iterations) as it is only guaranteed to converge to a local maximum rather than the global maximum (e.g. Stephens [2000]). The approach corresponds to minimising the approximate posterior expectation of the loss function L, with the approximation arising from averaging over the MCMC output. The iterative nature of the algorithm means that it must be run after the MCMC has completed.

ICs and relabelling algorithms have very similar goals, in that they assign meaning to each of the components. For example, under the IC considered above when we talk about the first component we mean 'the component with the smallest mean'. Relabelling algorithms attempt to give components meaning by enforcing some form of stable behaviour between iterations of the MCMC. If the goal of the inference is parameter estimation, it seems sensible to use an algorithm that stabilises the relabelled posterior of the parameters, using for example the algorithm of Stephens [1997a]. Farrar [2006], however, takes the opposing view that one should relabel using a different feature than the one of statistical interest, e.g. relabel based on component allocations when interested in parameter estimates.

A separate class of algorithms are the label invariant loss function approaches introduced by Celeux et al. [2000]. Here, the idea is to measure closeness between iterations of the MCMC without relying on labelling information. For example, one could consider pairwise comparison of the allocation of observations to components [Hurn et al., 2003].

2.3 Probabilistic Relabelling Algorithms

Probabilistic relabelling is first introduced by Jasra [2005]. The idea is that the permutation $\nu^{(r)}$ that is associated with the r^{th} iteration of the MCMC sampler is unknown. Therefore, the permutation can be viewed as having the discrete density $g_r(\nu^{(r)};\boldsymbol{\gamma},y)$ over $\nu^{(r)}\in S_K$, conditional on the data, y, and the full vector of parameters, $\boldsymbol{\gamma}$. Jasra [2005] then shows, using the strong law of large numbers, that one can estimate a quantity of interest $h(\cdot)$ via

$$h(\gamma)=\frac{1}{R}\sum_{r=1}^{R}\sum_{\nu^{(r)}\in S_K} h\left\{\nu^{(r)}(\boldsymbol{\gamma}^{(r)})\right\}\hat{g}_r(\nu^{(r)};\hat{\boldsymbol{\gamma}},y) \tag{2}$$

where $\nu^{(r)}(\boldsymbol{\gamma}^{(r)})$ represents the parameter vector with the component specific parameters permuted by $\nu^{(r)}$. The function of interest $h(\cdot)$ may depend additionally or alternatively on the allocation vector z.

To use this approach we need a way to estimate $g_r(\cdot)$, and we also need to know in advance the vector of true parameters $\boldsymbol{\gamma}$. Jasra [2005] gives various suggestions on how each of these issues may be dealt with. For example, the parameters $\boldsymbol{\gamma}$ can be derived by averaging over a small number of iterations from the MCMC, determined by eye not to have switched labels. The permutation densities $g_r(\cdot)$ are derived by estimating the posterior surface of the relabelled posterior using again a small number of iterations where the labels are deemed not to have switched. This uses a normal approximation, and the idea of estimating the relabelled posterior to deal with label switching was first suggested by Stephens [1997b].

Next we introduce a novel approach to probabilistic relabelling, in which $g_r(\cdot)$ and $\boldsymbol{\gamma}$ are estimated in an iterative fashion. An EM-type approach is adopted, where the missing data are the permutations $\{\nu^{(r)}, r=1,\ldots,R\}$ applied at each stage. The permutation densities, $g_r(\cdot)$, are estimated by conditioning only on the data, y, the current estimate of the parameters, $\boldsymbol{\gamma}$, and the current allocation vector, $z^{(r)}$. Letting $S_k^r=\{i: z_i^{(r)}=k\}$ be the set of indices of the observations belonging, before permutation, to the k^{th} parameter at iteration r, we calculate

$$\hat{g}_r(\nu^{(r)};\hat{\boldsymbol{\gamma}},y,z^{(r)})\propto\prod_{k=1}^{K}\prod_{i\in S_k^r}\hat{\pi}_{\nu(k)}f_{\nu(k)}\left(y_i|\hat{\boldsymbol{\theta}}_{\nu(k)},\hat{\boldsymbol{\eta}}\right) \tag{3}$$

where the right hand side corresponds to the allocated likelihood. So rather than using a normal approximation to the surface of the relabelled posterior, $g_r(\cdot)$ is estimated based on the allocated likelihood of the data under each permutation, the current estimate of the parameters (permuted according to the permutation under consideration) and the current allocation vector $z^{(r)}$. Finally $\hat{g}_r(\cdot)$ is normalised to sum to one over all possible permutations. A detailed derivation of (3) is given in the Appendix.

The usual application of the EM algorithm [Dempster et al., 1977] to the mixture problem views the available data as the observations, and the missing data the membership of the observations to the various components. The framework introduced here, on the other hand, can be interpreted as an EM algorithm where the available data are the output from the MCMC sampler, and the missing data are the permutations $\{\nu^{(r)}, r=1,\ldots,R\}$ applied at each stage. One could loosely consider the approaches suggested by Jasra [2005] as corresponding to a single iteration of such an EM algorithm, with sensible starting values chosen. We propose now a variety of extensions and alternatives that stem from placing probabilistic relabelling in this framework. We suggest first an iterative EM algorithm, which proceeds, after initialising estimates of the parameters $\boldsymbol{\gamma}$ using, for example, an IC, by:

E Step Estimate the densities $\{g_r(\cdot), r = 1, \ldots, R\}$ using the current estimate of $\boldsymbol{\gamma}$, via (3).

M step Update estimates of $\boldsymbol{\gamma}$ using (2), with appropriate choices of $h(\cdot)$. For example, the component weight π_1 may be updated by

$$\hat{\pi}_1 = \frac{1}{R} \sum_{r=1}^{R} \sum_{\nu^{(r)} \in S_K} \pi_1^{(r)} \hat{g}_r(\nu^{(r)}; \hat{\boldsymbol{\gamma}}, y, \boldsymbol{z}^{(r)})$$

As with all EM-type algorithms, convergence to the global maximum is not guaranteed — local modes or saddle points may instead be found. Therefore it is advised to use multiple starting points (different estimates of $\boldsymbol{\gamma}$). We call this EM approach 'EMP' (EM based probabilistic relabelling).

A popular alternative to the EM algorithm is the stochastic EM algorithm (SEM) [Celeux and Diebolt, 1985]. This introduces an extra step 'the S step', where the missing data is simulated from its estimated density. This constitutes drawing $\nu^{(r)}$ multinomially from the discrete density $g_r(\cdot)$. The randomness that this modification introduces helps to avoid the algorithm getting caught in local modes, and provides faster convergence. Additionally, the convergence of the SEM does not depend on the starting position [Celeux and Diebolt, 1985]. A SEM-type probabilistic relabelling strategy is as follows:

E step Estimate the densities $\{g_r(\cdot), r = 1, \ldots, R\}$ using the current estimate of $\boldsymbol{\gamma}$, via (3).

S step Simulate values for the permutations $\{\nu^{(r)}, r = 1, \ldots, R\}$ by drawing multinomially from the corresponding densities $g_r(\cdot)$, calling these simulated permutations $\tilde{\nu}^{(r)}$, $r = 1, \ldots, R$.

M step Update estimates of $\boldsymbol{\gamma}$ by taking ergodic averages over the sample after accounting for the permutations $\tilde{\nu}^{(r)}$, $r = 1, \ldots, R$. For example, the component weight π_1 may be updated by

$$\hat{\pi}_1 = \frac{1}{R} \sum_{r=1}^{R} \pi_1^{(r)}$$

after the inverse of $\tilde{\nu}^{(r)}$ has been applied at each r.

We call this approach 'SEMP' (SEM based probabilistic relabelling).

A final alternative that we suggest acknowledges that $\boldsymbol{\gamma}$ is itself unknown. We consider estimating the permutation densities $g_r(\cdot), r = 1, \ldots, R$ without conditioning on $\boldsymbol{\gamma}$ by integrating $\boldsymbol{\gamma}$ out with respect to its *relabelled* posterior, $q(\boldsymbol{\gamma})$:

$$\hat{g}_r(\nu^{(r)}; y, \boldsymbol{z}^{(r)}) \propto \int \prod_{k=1}^{K} \prod_{i \in S_k^r} \hat{\pi}_{\nu(k)} f_{\nu(k)} \left(y_i | \hat{\boldsymbol{\theta}}_{\nu(k)}, \hat{\boldsymbol{\eta}} \right) q(\boldsymbol{\gamma}) d\hat{\boldsymbol{\gamma}}$$

and approximate the integral by the Monte Carlo estimate over the MCMC sample

$$\hat{g}_r(\nu^{(r)}; y, \boldsymbol{z}^{(r)}) \propto \frac{1}{R} \sum_{r=1}^{R} \left\{ \prod_{k=1}^{K} \prod_{i \in S_k^r} \pi_{\nu(k)}^{(r)} f_{\nu(k)} \left(y_i | \boldsymbol{\theta}_{\nu(k)}^{(r)}, \boldsymbol{\eta}^{(r)} \right) \right\} \tag{4}$$

This leads to the algorithm

E step Estimate the densities $\{g_r(\cdot), r = 1, \ldots, R\}$ using the current estimate of the relabelled posterior density of $\boldsymbol{\gamma}$, via (4).

S step Simulate values for the permutations $\{\nu^{(r)}, r = 1, \ldots, R\}$ by drawing multinomially from the corresponding densities $g_r(\cdot)$, calling these simulated permutations $\tilde{\nu}^{(r)}$, $r = 1, \ldots, R$.

M step Estimate the relabelled posterior density, $q(\boldsymbol{\gamma})$, using the output from the MCMC and the current estimates $\tilde{\nu}^{(r)}$, $r = 1, \ldots, R$.

The M step is, therefore, fundamentally different from a usual EM or SEM algorithm — we estimate an entire posterior rather than point estimates of the parameters. We call this approach 'SEMUP' (SEM based unconditional probabilistic relabelling).

2.4 Comments

For the remainder of the paper we will consider seven different relabelling strategies, the IC, three deterministic relabelling algorithms and the three variants of probabilistic relabelling we introduced in the previous section. The notation used for the methods is defined in Table 1.

Notation	Method	Source
IC	Identifiability constraint	McLachlan and Peel [2000]
PL	Parameter relabelling algorithm	Stephens [1997a]
CPL	Class Probability relabelling algorithm	Stephens [2000]
AL	Allocation vector relabelling algorithm	Nobile and Fearnside [2007]
EMP	EM probabilistic	Section 2.3
SEMP	SEM probabilistic	Section 2.3
SEMUP	SEM unconditional probabilistic	Section 2.3

Table 1: Relabelling algorithms evaluated

One of the disadvantages of relabelling algorithms and ICs is that they apply a specific permutation $\nu^{(r)}$ at each iteration r, with no indication on how uncertain we are that this particular permutation is 'correct'. Using probabilistic methods, uncertainty in relabelling can be quantified by how close to one the probability of the most likely permutation being correct, $\max_{\nu^{(r)}} \left\{ \hat{g}_r(\nu^{(r)}; \hat{\gamma}, y, z^{(r)}) \right\}$, is, for each iteration of the MCMC.

A further advantage of probabilistic relabelling is the improved recovery of the posterior tails, which are often truncated using other methods. Consider 50 simulated observations from $0.5N(0,1) + 0.5N(2,1)$. Figure 1 compares the marginal posteriors for μ_1 (defined as the component mean with smallest ergodic average) under the PL and SEMP methods, assuming that all parameters in the model are unknown. The distributions are quite different in shape with the right hand tail being truncated for the PL algorithm in comparison to the SEMP method, which is compensated by a higher peak. Similar results are observed in all the probabilistic methods. This clearly shows the superior ability of probabilistic relabelling to recover posterior tails.

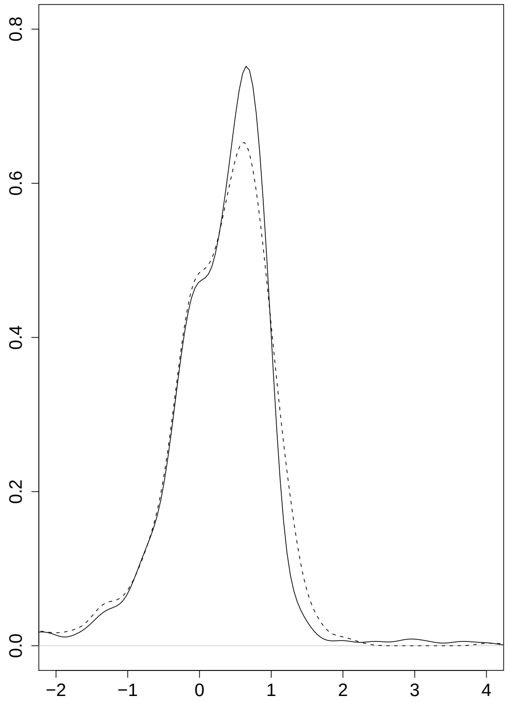

Figure 1: Graphs showing the posteriors for μ_1 (defined as the μ with smallest ergodic average) of the two component mixture model $0.5N(0,1) + 0.5N(2,1)$, for the PL (solid line) and SEMP (dashed line) algorithms.

3 Comparison of Methods

To evaluate the proposed algorithms we will now compare them to existing methods on observed and simulated data. The seven relabelling strategies that will be compared are given in Table 1.

3.1 The Galaxy Data

For the initial comparison we investigate the galaxy data which consist of the velocities of 82 different galaxies [Postman et al., 1986]. A histogram of the data is given in Fig. 2. This dataset has become the benchmark for testing different methods for analysis of mixture data. See Jasra et al. [2005] for a recent investigation into the galaxy data in the mixture modelling context.

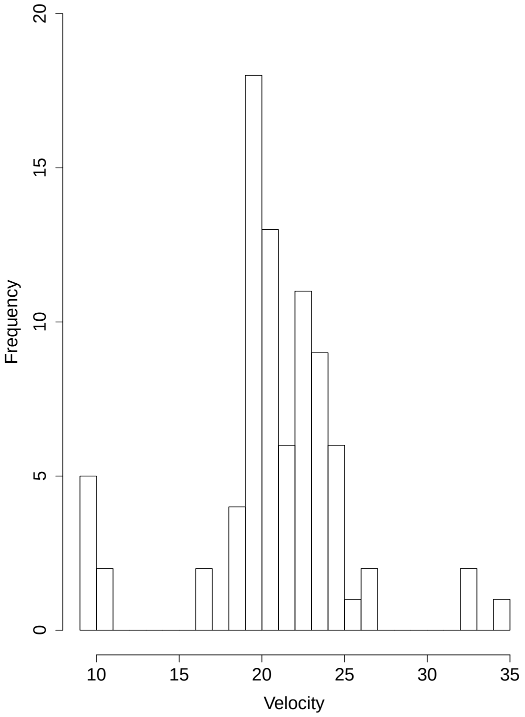

Figure 2: Histogram of the velocities of 82 galaxies.

An MCMC run on this data spends at least 10% of its iterations in each of K =3, 4 and 5 clusters suggesting that any of these choices could be sensible. We refer to Aitkin [2001] for an interesting summary of the differing posteriors for K achieved using different, but apparently similar, methods on this dataset. Here we will look in detail at the relabelling algorithms applied to the $K = 5$ case. As this is a single data set, it is feasible to use all of the output points from the MCMC for the SEMUP algorithm.

A remarkable stability between the different methods can be found as they recover almost identical values to each other for all the parameters. Table 2 shows an example of these results, the component mean of the fourth component, μ_4. The mean changes between methods, which is due to the difference in dealing with tails of the relabelled posterior by the various methods. Looking at the α-quantiles this is further illustrated by the fact that $q_{0.05}$ and $q_{0.95}$ are rather different between the methods. This suggests that there are adjacent components that are poorly separated. For a parameter in a well-separated component, such as the first component that accounts for the observations in the left-hand peak, almost identical results for the α-quantiles are observed for each relabelling method.

Method	Mean	Posterior Quantiles				
		$q_{0.05}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.95}$
IC	23.92	21.81	22.56	23.01	23.58	32.51
PL	22.60	21.33	22.04	22.65	23.09	23.65
CPL	22.39	21.09	21.83	22.43	23.00	23.45
AL	22.49	21.20	21.94	22.55	23.04	23.62
EMP	23.92	21.40	22.09	22.68	23.13	24.13
SEMP	22.60	21.25	22.00	22.63	23.09	24.09
SEMUP	23.37	16.39	22.21	22.88	23.44	34.60

Table 2: Summary of estimated μ_4 for different relabelling methods across all iterations of the MCMC with $K = 5$. Here, μ_4 is defined as the mean with the fourth smallest ergodic average.

Consequently the only major difference between the algorithms can be found in the variance for the estimates of each parameter as the allocation of component estimates from the tails has a large bearing on the estimated variances of the parameters.

Figure 3 gives the probabilities of the two most likely permutations (calculated from (3)) for the Galaxy data with the number of components $K = 3, 4, 5$, for 100 thinned iterations in each case. We have used the SEMP relabelling procedure. For $K = 3$ and 4, there is little or no uncertainty over which permutation of the labels is selected. For $K = 5$, however, it is often the case that there are two permutations with reasonable probabilities of being selected. This suggests that there are two components that are virtually indistinguishable, which implies that it may be beneficial to merge them. In this way, there is potential to use this method to help choose the number of components K.

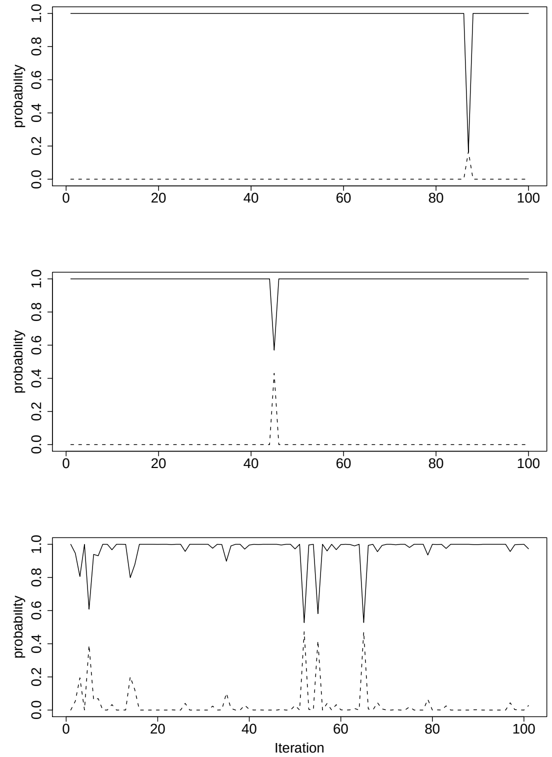

Figure 3: Graphs showing the probabilities of the most likely (solid line) and second most likely (dashed line) permutations at 100 iterations of the MCMC sampler for the galaxy data. The three graphs represent models with differing numbers of components — $K = 3$ (top), K=4 (middle) and K=5 (bottom).

3.2 Simulated Data

For a more thorough evaluation of the different relabelling algorithms we now turn to simulated data. We investigated different simulations in which we draw n observations from $\pi N(0,1) + (1-\pi)N(\mu_2, \sigma_2^2)$, for various combinations of $(n, \pi, \mu_2, \sigma_2^2)$. Each combination is repeated 100 times and the results are averaged over these repeats in order to remove the impact of individual data sets. Since it is computationally not feasible to use the SEMUP algorithm with all available iterations of the MCMC we set the number of posterior points to 100 for use in (4). As well as giving estimates of parameters, we give a measure of closeness of the estimated mixture distribution to the true density that we have simulated from, by simulating 10^6 values from the true density and estimating the Kullback-Leibler distance via

$$\varphi = \frac{10^4}{10^6} \sum_{i=1}^{10^6} \log \left\{ f(x_i; \theta_{\text{true}}) / f(x_i; \hat{\theta}) \right\}$$

where $\hat{\theta}$ is estimated via the various relabelling methods, and we have rescaled by 10^4 from a usual average to give more readable results.

For situations where the difference between two components is large, that is when either μ_2 was very different from zero or σ_2^2 was very different from 1 (e.g. $\sigma_2^2 = 0.1$ or $\sigma_2^2 = 10$), all relabelling algorithms, unsurprisingly, performed well as label switching occurs rarely. We therefore omit the details of these simulations and focus on situations where the two components are very similar. Tables 3-5 provide the details of some of the most interesting situations considered.

For the case where $(\pi, \mu_2, \sigma_2^2) = (0.5, 2, 1)$ and the sample size is varied as $n = 50$ and $n = 100$ (Table 3) it is immediately striking that for all relabelling algorithms except the IC, the estimates of μ_1 and μ_2 are pushed toward each other with the effect being strongest for the CPL and AL methods, and a moderate effect for the probabilistic strategies. Further, for all relabelling methods, the variances are severely over-estimated and neither feature is improved by an increase sample size, even when raised to $n = 500$ (not shown).

Both of these problems can be attributed to posterior weight on the possibility of both components being in the middle of the dataset with similar means and different variances. This solution, however, yields a rather different interpretation of the components than the one used to generate the data. The high standard deviation of the simulations indicates a high uncertainty in the 'correct' interpretation of the mixture distribution. In terms of the predictive error φ, PL and the probabilistic approaches are best performers in both sample sizes.

When looking at the results for very similar components ($\mu_2 = 0.1$, Table 4) we see the converse feature of the average estimates of μ_1 and μ_2 being pushed apart from each other. This is caused by the components being virtually indistinguishable so the MCMC responds by moving one component excessively to the left and the other excessively to the right. These opposing results are an illustration of the limitations of using ergodic average estimates for the parameters. For this situation interestingly the CPL and AL method perform better than the other methods, while probabilistic relabelling methods are in the middle. It is also interesting to see that, contrary to the previous set of situations, the estimates of the variance are more or less on target for all algorithms considered. In this case, the predictive error φ is minimised by CPL and AL, although SEMUP performs fairly well.

In Table 5 the components are more distinguishable ($\mu_2 = 2$), but the mixing weights are rather different, with $\pi = 0.1$. In this case, μ_1 and σ_1 are both severely over-estimated while μ_2 and σ_2 are estimated accurately for all relabelling strategies with none of the methods appearing to be superior to the others. Additionally π is also over-estimated strongly which can be attributed to the asymmetry in the posterior distribution. The predictive error φ is smallest for the PL method while it is largest for the IC.

n	θ	IC	PL	CPL	AL	EMP	SEMP	SEMUP
50	μ_1	−0.02 (0.21)	0.19 (0.12)	0.64 (0.19)	0.65 (0.18)	0.20 (0.09)	0.20 (0.16)	0.34 (0.16)
	μ_2	1.90 (0.12)	1.69 (0.03)	1.24 (0.09)	1.23 (0.08)	1.69 (0.01)	1.68 (0.06)	1.54 (0.06)
	σ_1^2	1.63 (0.39)	1.62 (0.31)	1.58 (0.51)	1.58 (0.51)	1.62 (0.37)	1.63 (0.38)	1.66 (0.39)
	σ_2^2	1.61 (0.47)	1.62 (0.55)	1.66 (0.34)	1.66 (0.34)	1.62 (0.48)	1.61 (0.48)	1.58 (0.46)
	π	0.50 (0.06)	0.51 (0.09)	0.49 (0.38)	0.49 (0.39)	0.51 (0.11)	0.46 (0.08)	0.47 (0.11)
	φ	205	97	166	169	96	86	83
100	μ_1	0.07 (0.18)	0.23 (0.23)	0.58 (0.36)	0.58 (0.36)	0.27 (0.28)	0.28 (0.25)	0.30 (0.26)
	μ_2	1.91 (0.19)	1.75 (0.23)	1.39 (0.39)	1.40 (0.38)	1.71 (0.30)	1.70 (0.29)	1.68 (0.27)
	σ_1^2	1.52 (0.39)	1.52 (0.39)	1.50 (0.39)	1.50 (0.39)	1.51 (0.35)	1.52 (0.39)	1.52 (0.36)
	σ_2^2	1.47 (0.36)	1.47 (0.37)	1.49 (0.36)	1.49 (0.36)	1.49 (0.40)	1.47 (0.35)	1.47 (0.39)
	π	0.50 (0.05)	0.50 (0.07)	0.49 (0.24)	0.49 (0.24)	0.50 (0.09)	0.50 (0.09)	0.49 (0.09)
	φ	110	65	188	184	66	67	70

Table 3: Average parameter estimates over 100 iterations for different relabelling strategies when $(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = (0.5, 0, 2, 1, 1)$ for $n = 50$ and $n = 100$. Values in parentheses give the standard deviations of the estimates.

θ	IC	PL	CPL	AL	EMP	SEMP	SEMUP
μ_1	−0.60 (0.24)	−0.42 (0.28)	−0.22 (0.30)	−0.21 (0.30)	−0.47 (0.30)	−0.44 (0.28)	−0.36 (0.29)
μ_2	0.67 (0.22)	0.47 (0.24)	0.27 (0.26)	0.27 (0.26)	0.52 (0.25)	0.50 (0.25)	0.42 (0.25)
σ_1^2	0.95 (0.25)	0.95 (0.31)	0.94 (0.26)	0.94 (0.26)	0.95 (0.24)	0.95 (0.27)	0.96 (0.28)
σ_2^2	0.92 (0.23)	0.91 (0.29)	0.92 (0.23)	0.92 (0.23)	0.92 (0.23)	0.91 (0.24)	0.91 (0.24)
π	0.49 (0.09)	0.49 (0.15)	0.47 (0.29)	0.47 (0.29)	0.49 (0.14)	0.48 (0.14)	0.50 (0.15)
φ	211	37	0.4	0.4	66	50	18

Table 4: Average parameter estimates over 100 iterations for different relabelling strategies when $(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = (0.5, 0, 0.1, 1, 1)$ for $n = 100$. Values in parentheses give the standard deviations of the estimates.

θ	IC	PL	CPL	AL	EMP	SEMP	SEMUP
μ_1	0.75 (0.40)	0.91 (0.47)	1.07 (0.55)	1.08 (0.55)	0.85 (0.45)	0.91 (0.49)	0.95 (0.49)
μ_2	2.32 (0.20)	2.17 (0.20)	2.00 (0.21)	1.99 (0.21)	2.22 (0.23)	2.16 (0.23)	2.12 (0.20)
σ_1^2	1.39 (0.36)	1.46 (0.46)	1.36 (0.43)	1.35 (0.42)	1.35 (0.36)	1.39 (0.38)	1.38 (0.41)
σ_2^2	1.02 (0.29)	0.95 (0.25)	1.05 (0.26)	1.05 (0.27)	1.05 (0.31)	1.02 (0.29)	1.02 (0.32)
π	0.39 (0.10)	0.36 (0.12)	0.28 (0.18)	0.28 (0.18)	0.38 (0.12)	0.39 (0.13)	0.40 (0.14)
φ	184	51	57	59	125	106	110

Table 5: Average parameter estimates over 100 iterations for different relabelling strategies when $(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = (0.1, 0, 2, 1, 1)$ for $n = 100$. Values in parentheses give the standard deviations of the estimates.

Overall the results indicate that none of the methods compared are performing uniformly better than any of the others leaving the ultimate decision on which method to use to the user. The CPL and AL methods are unstable in terms of the predictive error φ as they perform well when the components are very hard to distinguish, but show poor performance when the components are more separated. Consistent results for φ are obtained for the PL and the probabilistic methods. Based on these results it is, however, evident that the use of ergodic averages can often be detrimental. Due to the large variation in the parameter estimates we believe the SEMUP method is more appropriate as it is probabilistic and moreover avoids conditioning on the parameter estimates. It does, however, depend on the accuracy of the approximation in (4) through the value of R.

4 Discussion

In this paper we have developed a new class of probabilistic methods for the label switching problem in Bayesian mixture models. The main advantages of these approaches are on the one hand that the tails of the posterior distributions are recovered and on the other hand uncertainty associated with relabelling can be incorporated, features that are not present for deterministic relabelling algorithms. The computation time of the probabilistic methods are either substantially lower than or on par with the existing deterministic methods with the exception of the IC. It is shown through analysis of an observed dataset as well as simulation that the parameter estimates obtained by probabilistic relabelling are virtually the same as for the deterministic approaches suggesting that the above advantages come without any loss.

We also introduce an algorithm for probabilistic relabelling, called SEMUP, that does not rely on ergodic average estimates of parameters as we integrate over a relabelled posterior. Although there is some additional computation required to approximate the relevant integral that also introduced a trade-off between speed and accuracy, the additional time was found to be reasonable for single datasets.

During the evaluation of the methods it was pointed out that some information about the choice of K, the number of components, can be derived from probabilistic relabelling algorithms. Although the full extent of the relevance of probabilistic relabelling for choosing K is still to be evaluated carefully, it does show promise. The uncertainty in the relabelling can be used as an indication that too many components are in the model, since high uncertainty in relabelling suggests that there is ambiguity between adjacent components, implying that it may be better to merge them. Further work will need to be done to get a better understanding of this.

A Derivation of (3)

First, $g_r(\nu_r; \hat{\gamma}, y, z^{(r)})$ is defined as the probability that permutation ν_r is 'correct', given the data y, the current estimate of the parameters $\hat{\gamma}$, and the allocation vector $z^{(r)}$, for the r^{th} iteration of the sampler. In an abuse of notation when we write ν_r henceforth we mean 'permutation ν_r is correct'.

Then

$$\mathbb{P}[\nu_r|y, z^{(r)}, \hat{\gamma}] = \frac{\mathbb{P}[y|\nu_r, z^{(r)}, \hat{\gamma}]\mathbb{P}[z^{(r)}|\nu_r, \hat{\gamma}]\mathbb{P}[\nu_r|\hat{\gamma}]}{\mathbb{P}[y|\hat{\gamma}, z^{(r)}]\mathbb{P}[z^{(r)}|\hat{\gamma}]}$$

Now, the terms in the denominator do not depend on ν_r, and we assume that each permutation is equally likely, so we are left with

$$\begin{aligned}\mathbb{P}[\nu_r|y, z^{(r)}, \hat{\gamma}] \propto& \mathbb{P}[y|\nu_r, z^{(r)}, \hat{\gamma}]\mathbb{P}[z^{(r)}|\nu_r, \hat{\gamma}] \\ =& \prod_{i=1}^{n} \mathbb{P}[y_i|\nu_r, z_i^{(r)}, \hat{\gamma}]\mathbb{P}[z_i^{(r)}|\nu_r, \hat{\gamma}] \\ =& \prod_{k=1}^{K} \prod_{i \in S_k^r} \hat{\pi}_{\nu(k)} f_{\nu(k)} \left(y_i | \hat{\boldsymbol{\theta}}_{\nu(k)}, \hat{\boldsymbol{\eta}} \right)\end{aligned}$$

which is the form given in (3).

References

M. Aitkin. Likelihood and Bayesian analysis of mixtures. *Stat. Model.*, 1:287–304, 2001.

G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Comput. Stat. Q.*, 2:73–82, 1985.

G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *J. Am. Stat. Assoc.*, 95:957–970, 2000.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Ser. B*, 39:1–38, 1977.

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

D. Farrar. Approaches to the label-switching problem of classification, based on partition-space relabeling and label-invariant visualization. Technical report, Statistical consulting center and department of statistics, Virginia Polytechnic, 2006.

J. Geweke. Interpretation and inference in mixture models: simple MCMC works. *Comp. Stat. and Data Analysis*, 51:3529–3550, 2007.

M. A. Hurn, A. Justel, and C. P. Robert. Estimating mixtures of regressions. *J. Comp. Graph. Stat.*, 12:55–79, 2003.

A. Jasra. *Bayesian Inference for Mixture Models via Monte Carlo.* PhD thesis, Imperial College London, 2005.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain monte carlo methods and the label switching problem in Bayesian mixture modelling. *Stat. Sci.*, 20:50–67, 2005.

J. M. Marin, K. L. Mengersen, and C. P. Robert. *Bayesian modelling and inference on mixtures of distributions*. Elsevier, 2005.

G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

A. Nobile and A. T. Fearnside. Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Stat. and Comput.*, 17(2):147–162, 2007.

M. Postman, J. P. Huchra, and M. J. Geller. Probes of large-scale structure in the Corona Borealis region. *Astro. J.*, 92:1238–1246, 1986.

S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. Ser. B*, 59:758–764, 1997. With discussion.

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

M. Stephens. *Bayesian Methods for Mixtures of Normal Distributions*. PhD thesis, University of Oxford, 1997a.

M. Stephens. Discussion of On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. Ser. B*, 59:768–769, 1997b.