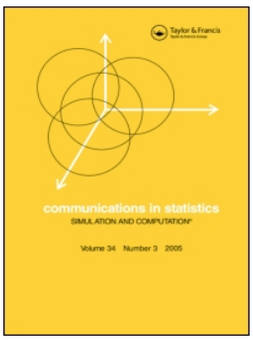

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

# Relabel mixture models via modal clustering

Qiang Wu & Weixin Yao

**To cite this article:** Qiang Wu & Weixin Yao (2017) Relabel mixture models via modal clustering, Communications in Statistics - Simulation and Computation, 46:5, 3406-3418, DOI: 10.1080/03610918.2015.1089287

**To link to this article:** https://doi.org/10.1080/03610918.2015.1089287

Accepted author version posted online: 30 Oct 2015.
Published online: 28 Dec 2016.

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 46, NO. 5, 3406–3418
http://dx.doi.org/10.1080/03610918.2015.1089287

Taylor & Francis
Taylor & Francis Group

# Relabel mixture models via modal clustering

Qiang Wu[a] and Weixin Yao[b]

[a]Department of Biostatistics, East Carolina University, Greenville, NC, USA; [b]Department of Statistics, University of California, Riverside, CA, USA

**ABSTRACT**

Effectively solving the label switching problem is critical for both Bayesian and Frequentist mixture model analyses. In this article, a new relabeling method is proposed by extending a recently developed modal clustering algorithm. First, the posterior distribution is estimated by a kernel density from permuted MCMC or bootstrap samples of parameters. Second, a modal EM algorithm is used to find the $m!$ symmetric modes of the KDE. Finally, samples that ascend to the same mode are assigned the same label. Simulations and real data applications demonstrate that the new method provides more accurate estimates than many existing relabeling methods.

**ARTICLE HISTORY**
Received 23 June 2015
Accepted 26 August 2015

**KEYWORDS**
Bayesian analysis; EM algorithm; Finite mixture models; Kernel density estimation; Label switching; Modal clustering

**MATHEMATICS SUBJECT CLASSIFICATION**
62F15; 62G05; 62H30

## 1. Introduction

Mixture models are very popular tools to model the population when it is heterogeneous and consists of several homogeneous subgroups. Mixture models can be used for cluster analysis, latent class analysis, discriminant analysis, image analysis, survival analysis, disease mapping, meta analysis, and more. They provide extremely flexible descriptive models for distributions in data analysis and inference. For a general introduction to mixture models, see Lindsay (1995), Böhning (1999), McLachlan and Peel (2000), and Frühwirth-Schnatter (2006).

One important feature of the mixture likelihood is its invariance to permutations of component labels which is referred to as *label switching* by Redner and Walker (1984). Therefore, when one wants to use simulations or bootstraps to conduct inference for the maximum likelihood estimate (MLE) of component related parameters, such as component parameters, component densities, or classification probabilities, difficulties arise as the components of the MLE can be ordered arbitrarily. Given a sequence of MLEs, to obtain a meaningful interpretation of the components, it is necessary to relabel all components of the MLEs such that they have a consistent label meaning. For Bayesian mixtures, if the prior is invariant to permutations of component labels, so is the posterior. Hence, in one run of an MCMC sampler the order of components might change multiple times between iterations. If we want to infer parameters that are specific to individual components of the mixture model, we must find methods to relabel the samples so that the components are in the same order at each iteration. In this article, we focus on the label switching problem mainly under the setting of Bayesian mixtures, but the proposed relabeling method can also be applied to Frequentist mixture models in simulation studies or bootstrap procedures.

**CONTACT** Qiang Wu wuq@ecu.edu Department of Biostatistics, East Carolina University, Greenville, NC 27834, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

Many methods have been proposed to solve the label switching problem for Bayesian mixtures. Diebolt and Robert (1994), Dellaportas et al. (1996), and Richardson and Green (1997) use order constraints (such as $\mu_1 < \mu_2$ for a two component normal mixture) to relabel the MCMC samples. Celeux (1998) and Stephens (2000) propose a relabeling algorithm, which minimizes the posterior expectation of certain loss function. Yao and Lindsay (2009) propose relabeling the samples based on matching posterior modes between successive iterates. Sperrin et al. (2010) and Yao (2012a) propose some probabilistic approaches to account for the uncertainty of the relabeling. Other labeling methods include, for example, Frühwirth-Schnatter (2001), Hurn et al. (2003), Chung et al. (2004), Jasra et al. (2005), Marin et al. (2005), Geweke (2007), Grün and Leisch (2009), Cron and West (2011), Yao (2012b), and Papastamoulis and Iliopoulos (2010). Some of these methods are *online* and proposed to solve the label switching problem on the fly (see, e.g., Celeux, 1998; Stephens, 2000).

In this article, we propose a new label switching method by extending the modal clustering approach of Li et al. (2007). Due to permutation symmetry, the posterior distribution of parameters from an $m$-component mixture model has a total of $m!$ symmetric regions. If there is one major mode in each region, then there are a total of $m!$ symmetric major modes. Relabeling the samples is equivalent to determining which symmetric region each sample belongs to. We propose to first estimate the posterior distribution by fitting a nonparametric kernel density to all the samples and their permuted images. Then using each sample as the initial value, we find the converged mode of the estimated kernel density based on an EM type algorithm. If two samples converge to the same mode, then we say that they are in the same symmetric region and therefore receive the same label. Note that the number of modes for the kernel density estimate (KDE) decreases when the bandwidth parameters increase. Therefore, the bandwidths can be naturally chosen by increasing them from some small values and stopping when there are $m!$ symmetric major modes. Unlike Yao and Lindsay (2009), the new method does not depend on the Bayesian model used to generate the MCMC samples. So the proposed algorithm of the new method is applicable to any finite mixture models. Compared to most of the existing labeling methods, the proposed labeling method directly uses the geometric structure of the posterior distribution and thus bears a nice interpretation. The proposed relabeling method can also be applied online to save storage and boost computational efficiency. In addition, our simulation studies demonstrate that the new method provides more accurate estimates than many existing relabeling methods.

The organization of this article is as follows. The label switching problem is formally introduced in Section 2 followed by Section 3 for several illustrative examples. The newly proposed label switching method via modal clustering is presented in Section 4. Section 5 includes some simulations as well as applications of the new method to the illustrative examples. Finally, some discussions are given in Section 6.

## 2. The label switching problem

In the mixture model analysis, an $m$-component mixture density can be expressed as

$$p(x; \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j f(x; \boldsymbol{\lambda}_j), \tag{2.1}$$

where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_m, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$ is the vector of unknown component specific parameters and $f(x; \boldsymbol{\lambda})$ is the component density whose functional form is usually known, such as the normal density $\phi(x; \mu, \sigma^2)$. We assume that the number of components $m \geq 2$ is known in

advance. The prior component proportions $\pi_1, \ldots, \pi_m > 0$ satisfy $\sum_j \pi_j = 1$. For notational ease, we sometimes refer to them in a vector form $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$. The model identifiability requires that $\boldsymbol{\lambda}_j \neq \boldsymbol{\lambda}_k$ for all $1 \leq j \neq k \leq m$. For any permutation $\boldsymbol{\omega} = (\omega(1), \ldots, \omega(m))$ of component labels $(1, \ldots, m)$, let us define the permuted parameter vector of $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta}^{\boldsymbol{\omega}} = (\pi_{\omega(1)}, \ldots, \pi_{\omega(m)}, \boldsymbol{\lambda}_{\omega(1)}, \ldots, \boldsymbol{\lambda}_{\omega(m)}).$$

A special feature of the mixture model is that the density function (2.1) is invariant under any permutation of the component labels, i.e., $p(x; \boldsymbol{\theta}) = p(x; \boldsymbol{\theta}^{\boldsymbol{\omega}})$ for all permutations $\boldsymbol{\omega}$. For this reason, the identifiability of model (2.1) is usually defined up to permutations of the component labels.

For independent observations $\mathbf{x} = (x_1, \ldots, x_n)$ from (2.1), their likelihood function is

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta}), \tag{2.2}$$

which is also invariant to permutations of the component labels, i.e., $L(\boldsymbol{\theta}; \mathbf{x}) = L(\boldsymbol{\theta}^{\boldsymbol{\omega}}; \mathbf{x})$ for all $\boldsymbol{\omega}$. The MLE of $\boldsymbol{\theta}$ can be found by maximizing (2.2) using the EM algorithm of Dempster et al. (1977). How to label the MLE is usually irrelevant unless simulation or bootstrap procedures are used. If one wishes to synthesize results from simulations or bootstraps, then it is necessary to relabel the multiple MLEs in a consistent way.

In Bayesian analysis, given a prior distribution $p(\boldsymbol{\theta})$ of the model parameters, inference on $\boldsymbol{\theta}$ is drawn based on the posterior distribution

$$p(\boldsymbol{\theta}; \mathbf{x}) = p(\boldsymbol{\theta}) L(\boldsymbol{\theta}; \mathbf{x}) / p(\mathbf{x}), \tag{2.3}$$

where $p(\mathbf{x}) = \int p(\boldsymbol{\theta}) L(\boldsymbol{\theta}; \mathbf{x}) d\boldsymbol{\theta}$ is the marginal distribution of $\mathbf{x}$. When (2.3) is not analytically available, an MCMC sampler such as the Gibbs sampler can be used to generate samples from (2.3) for the inferential purpose. Diebolt and Robert (1994) and Richardson and Green (1997) have more details on drawing MCMC samples for mixture models. Particularly, when the prior distribution is invariant to permutations of the component labels, i.e., $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}^{\boldsymbol{\omega}})$ for all $\boldsymbol{\omega}$, so is the posterior distribution (2.3). Indeed, due to permutation symmetry, the posterior distribution has a total of $m!$ symmetric regions over which the posterior appears a mirror image of each other. Therefore, all component-specific parameters calculated directly from the posterior, such as the posterior marginal density or classification probabilities, are exactly the same for all components and thus meaningless for the inference. In practice, the MCMC sampler may have multiple jumps among different symmetric regions. In fact, it is desirable for the sampler to explore all $m!$ regions for the sake of convergence detection (Jasra et al., 2005). As a result, it is often meaningless to draw inference from the MCMC samples directly without solving the labeling issue. Samples located in the same symmetric region should be labeled in the same manner so the label switching problem is equivalent to a region identification or a clustering problem. It usually requires some post-generation treatments of the MCMC samples. But some relabeling methods are online and proposed to solve the labeling issue on the fly in order to reduce storage and boost computational efficiency (see, e.g., Celeux, 1998; Stephens, 2000).

## 3. Illustrative examples

In this section, we describe three datasets as illustrative examples: the acidity data of Crawford et al. (1992) and Crawford (1994), the fish data of Titterington et al. (1985), and the old

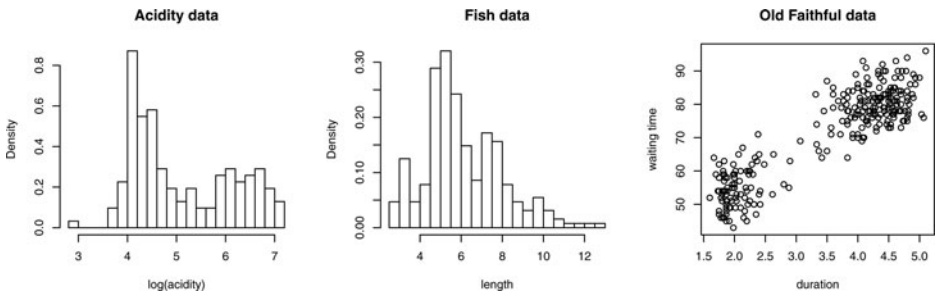


**Figure 1.** Histograms of the acidity data and the fish data, as well as a scatter plot of the old faithful data.

faithful data of Stephens (1997) and Dellaportas and Papageorgious (2006). Histograms of the acidity data and the fish data, as well as a scatter plot of the old faithful are shown in Fig. 1. The acidity data and the fish data are analyzed using the following $m$-component univariate Bayesian normal mixture model, while the old faithful data are analyzed using the multivariate one.

**Bayesian Normal Mixtures.** *Let* $\mathbf{z}_i = (z_{i1}, \cdots, z_{im})$ *for* $i = 1, \cdots, n$ *be the latent class variables where* $z_{ij} = 1$ *if the* $i$*th observation is from the* $j$*th component and* $z_{ij} = 0$ *otherwise. An* $m$*-component univariate normal mixture can be modeled by*

$$\Pr(z_{ij} = 1|\boldsymbol{\pi}) = \pi_j, \text{ and } x_i|z_{ij} = 1, \mu_j, \sigma_j^2 \sim N(\mu_j, \sigma_j^2),$$

*where* $N(\mu, \sigma^2)$ *is a univariate Normal distribution with mean* $\mu$ *and variance* $\sigma^2$*. Similarly, an* $m$*-component multivariate normal mixture model can be written as*

$$\Pr(z_{ij} = 1|\boldsymbol{\pi}) = \pi_j, \text{ and } x_i|z_{ij} = 1, \boldsymbol{\mu}_j, \Sigma_j \sim N(\boldsymbol{\mu}_j, \Sigma_j),$$

*where* $N(\boldsymbol{\mu}, \Sigma)$ *is a multivariate Normal distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\Sigma$*.*

*For Bayesian analysis, we adopt the following conjugate priors*

$$\boldsymbol{\pi} \sim D(\delta, \ldots, \delta),\ \sigma_j^2 \sim i\Gamma(\alpha, \beta),\ \mu_j \sim N(\xi, \kappa\sigma_j^2)$$

*for the univariate case, and*

$$\boldsymbol{\pi} \sim D(\delta, \ldots, \delta),\ \Sigma_j \sim iW(\Psi, \nu),\ \mu_j \sim N(\boldsymbol{\xi}, \kappa\Sigma_j)$$

*for the multivariate case, where* $i\Gamma(\alpha, \beta)$ *is an inverse-Gamma distribution with mean* $\beta/(\alpha - 1)$*,* $D(\delta, \cdots, \delta)$ *is a Dirichlet distribution with concentration parameter* $\delta$*, and* $iW(\Psi, \nu)$ *is an inverse-Wishart distribution with* $\nu$ *degrees of freedom and scale matrix* $\Psi$*. A Gibbs sampler is used to draw 20,000 MCMC samples from the posterior distribution of the parameters after 20,000 burn-in steps.*

The acidity data concern the logarithm of an acidity index measured on a sample of 155 lakes in north-central Wisconsin and have been previously analyzed using a mixture of normal distributions by Crawford et al. (1992), Crawford (1994), and Richardson and Green (1997). Although a 3-component normal mixture model has the greatest posterior probability in their analysis, there is some ambiguity as far as the number of components. This imposes great difficulties on any label switching methods. For this reason, we remove the lower outlier (log(acidity) = 2.93) and fit a 3-component normal mixture to the rest of the data. Prior hyperparameters are chosen to be $\delta = 1, \alpha = 2, \beta = 0.1, \kappa = 10$, and $\xi = \bar{x}$ (the sample mean). A trace plot of the mean parameters is given in Fig. 2 using the 20,000 MCMC samples.

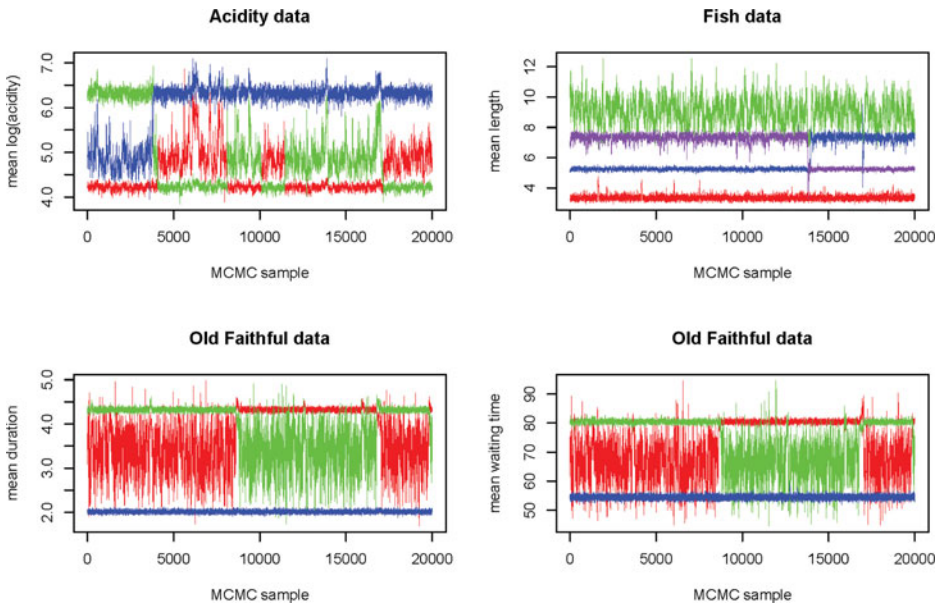


**Figure 2.** Trace plots of the mean parameters using the 20,000 MCMC samples from the acidity data, the fish data, and the old faithful data, respectively.

The fish data of Titterington et al. (1985) contain 256 observations of fish length in inches. The heterogeneity is likely from the age groups of the fish, but the age of fish is much harder to determine. Thus, normal mixtures have been fitted to model the unobserved heterogeneity by Titterington et al. (1985). Both Fig. 1 and the analysis of Titterington et al. (1985) suggest a mixture of $m = 4$ components with modes roughly at (3.25, 5.00, 7.75, 9.75) inches. Prior hyperparameters are chosen to be $\delta = 1$, $\alpha = 2$, $\beta = 1$, $\kappa = 10$, and $\xi = \bar{x}$ in our analysis. Figure 2 shows a trace plot of the mean parameters using the 20,000 MCMC samples.

The old faithful data analyzed by Stephens (1997) and Dellaportas and Papageorgious (2006) consist of 272 bivariate observations of the duration of eruption and the waiting time between eruptions of the old faithful geyser. Dellaportas and Papageorgious (2006) show that there are most likely three clusters. We analyze the data using a 3-component multivariate normal mixture model with prior hyperparameters chosen to be $\delta = 1$, $\nu = 4$, $\Psi = S$ (the sample covariance matrix), $\kappa = 10$, and $\xi = \bar{x}$. Trace plots of the mean duration and mean waiting time are given in Fig. 2 using the 20,000 MCMC samples. Label switching is observed in all three examples.

## 4. Label switching via modal clustering

The label switching problem in Bayesian mixture model analysis is equivalent to identifying the $m!$ symmetric modal regions of the posterior. When the posterior has a unique mode and monotone elsewhere within each region, the regions can be identified by the locations of their modes and there must exist an ascending path from every point in the region to its mode. In order to locate the modal region for each MCMC sample, we propose running an EM type algorithm using each MCMC sample as an initial value and find the corresponding converged mode. If two samples converge to the same mode, then they are determined in the same symmetric region and, therefore, receive the same label.

For the purpose of label switching, we record all $m!$ permutation images of the original MCMC samples and use them as a training set. Then the posterior distribution can be estimated from the training set by a kernel density

$$\hat{p}(\boldsymbol{\theta}) = \frac{1}{m!N} \sum_{i=1}^{m!N} K(\boldsymbol{\theta}; \boldsymbol{\theta}_i, H), \tag{4.1}$$

where $K(\cdot)$ is a multivariate kernel function and $N$ is the number of original MCMC samples. Let $d \geq 2$ be the dimension of the involved parameters. The bandwidth parameters are given by the $d \times d$ matrix $H$. The most frequently used kernel is the multivariate Gaussian kernel $K(\cdot; \boldsymbol{\mu}, H) = \phi(\cdot; \boldsymbol{\mu}, H^2)$, and a simple choice for the bandwidth matrix is a diagonal one $H = \text{diag}(h_1, \ldots, h_d)$, which results in a multiplicative kernel.

As can be seen, the KDE $\hat{p}(\boldsymbol{\theta})$ is a mixture density itself with $m!N$ components. The mode of $\hat{p}(\boldsymbol{\theta})$ that each original MCMC sample ascends to can be found by the nonparametric clustering approach of Li et al. (2007) which we summarize in the following.

From some initial parameter values $\boldsymbol{\theta}^{(0)}$, the modal EM (MEM) algorithm of Li et al. (2007) solves a local maximum of $\hat{p}(\boldsymbol{\theta})$ by iterating through two steps until a convergence criterion is met: starting with $r = 0$,

**Step 1:** compute $p_i^{(r)} = K(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}_i, H)/m!N\hat{p}(\boldsymbol{\theta}^{(r)})$ for $i = 1, \cdots, m!N$;

**Step 2:** update $\boldsymbol{\theta}^{(r+1)} = \text{argmax}_{\boldsymbol{\theta}} \sum_i p_i^{(r)} \log K(\boldsymbol{\theta}; \boldsymbol{\theta}_i, H)$.

The convergence criterion can be defined on the absolute differences between iterations such as $|\hat{p}(\boldsymbol{\theta}^{(r+1)}) - \hat{p}(\boldsymbol{\theta}^{(r)})|$, $\| \boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)} \|$, and $\max |\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}|$ or on the relative absolute differences such as $|\hat{p}(\boldsymbol{\theta}^{(r+1)}) - \hat{p}(\boldsymbol{\theta}^{(r)})|/\hat{p}(\boldsymbol{\theta}^{(r)})$, $\| \boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)} \| / \| \boldsymbol{\theta}^{(r)} \|$, and $\max |\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}|/|\boldsymbol{\theta}^{(r)}|$. See Li et al. (2007) for more details about the ascending property of the MEM algorithm.

When the Gaussian kernel is used, step 2 acquires a unique maximum and simplifies to a weighted average

**Step 2′:** update $\boldsymbol{\theta}^{(r+1)} = \sum_i p_i^{(r)} \boldsymbol{\theta}_i$.

The bandwidth parameters governed by $H$ play an important role on the smoothness and the number of local modes possessed by the KDE (4.1). As the bandwidths increase, the KDE becomes smoother and features less local modes. Following Li et al. (2007), the bandwidth matrix $H$ can be chosen such that the KDE has exactly $m!$ modes corresponding to the $m!$ symmetric regions. Based on our limited empirical experience, the rule-of-thumb bandwidths $\hat{H} = (m!N)^{-1/(d+4)}\text{diag}(s_1, \cdots, s_d)$ (Scott, 1992), where $s_1, \cdots, s_d$ are the standard deviations of the permuted MCMC samples, usually works well or could be a good starting point to determine the appropriate bandwidths.

With appropriately chosen bandwidths, the MEM algorithm is then applied to each original MCMC sample. Samples that lead to the same mode are determined to belong to the same region and, therefore, receive the same label. For example, suppose $m = 2$ and there are two symmetric modal regions of the posterior. Let one of the two symmetric modes, denoted by $\hat{\boldsymbol{\theta}}$, be the reference mode/label. The label switching issue is solved if all permuted samples have the same label as $\hat{\boldsymbol{\theta}}$. Our proposed relabeling method is to run the MEM algorithm using each original MCMC sample as an initial value. If a sample converges to $\hat{\boldsymbol{\theta}}$ based on the MEM algorithm, then it is said to be in the modal region of $\hat{\boldsymbol{\theta}}$ and has the same label as $\hat{\boldsymbol{\theta}}$. However, if a sample, say $\boldsymbol{\theta}_i$, converges to $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$, a permuted image of $\hat{\boldsymbol{\theta}}$, then $\boldsymbol{\theta}_i$ is in the modal region of $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$ and thus has the same label as $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$. Therefore, a permutation that brings $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$ back to $\hat{\boldsymbol{\theta}}$ will successfully relabel $\boldsymbol{\theta}_i$ consistently with $\hat{\boldsymbol{\theta}}$.

The number of MCMC samples required to effectively estimate the kernel density for the purpose of label switching could be much smaller than that necessary for an effective inference. So the label switching method using the MEM algorithm can be made online for a better computational efficiency and a reduced storage. If we let the training set consist of permuted MCMC samples obtained during the first few steps of the sampling process, then the label of each sample in the following steps can be determined on the fly. This online implementation is also very helpful in convergence detection. According to our simulation studies and real data applications in Section 5, the first few hundred original MCMC samples may be enough for the purpose.

## 5. Numerical studies

In this section, we conduct some simulation studies to show the effectiveness of the new label switching method and apply the new method to the three examples introduced in Section 3. In all applications, a multiplicative Gaussian kernel is adopted. The rule-of-thumb bandwidths are first attempted and increased if the label switching problem was not successfully resolved. The MEM algorithm is claimed a convergence if the relative changes in consecutive parameter values does not exceed $10^{-4}$. Two converged modes are determined the same if they differ by no more than 1%. A Linux computer with a 2.60GHz CPU is used for the computation.

### 5.1. Simulations

Datasets of size $n = 200$, 300, and 400 are simulated, respectively, from the following three models

$$C1 : \frac{1}{2}N(0, 1) + \frac{1}{2}N(3, 1.5),$$
$$C2 : \frac{1}{3}N(0, 0.5) + \frac{1}{3}N(0, 2) + \frac{1}{3}N(5, 1),$$

and

$$\begin{aligned} C3 : \frac{1}{4}N\left(\begin{pmatrix} 4.5 \\ -2.5 \end{pmatrix}, \begin{pmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{pmatrix}\right) \\ + \frac{1}{4}N\left(\begin{pmatrix} -3 \\ 4 \end{pmatrix}, \begin{pmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{pmatrix}\right) + \frac{1}{4}N\left(\begin{pmatrix} 6.5 \\ 7 \end{pmatrix}, \begin{pmatrix} 4 & 2.5 \\ 2.5 & 4 \end{pmatrix}\right) \\ + \frac{1}{4}N\left(\begin{pmatrix} 7 \\ -3 \end{pmatrix}, \begin{pmatrix} 4 & 2.5 \\ 2.5 & 9 \end{pmatrix}\right). \end{aligned}$$

One hundred replicates are generated from each model in order to compare the efficiency of the parameter estimates. Model C1 simulates a case where components are highly mixed. The first two components of model C2 have the same mean and differ only by their variances. Model C3 is a bivariate model used by Papastamoulis and Iliopoulos (2010) and Rodríguez and Walker (2012). For each simulated dataset, the (univariate or multivariate) Bayesian normal mixture model given in Section 3 is used to generate 20,000 MCMC samples of parameter estimates after 20,000 burn in steps. Prior hyperparameters are chosen to be $\delta = 1$, $\alpha = 2$, $\beta = 1$, $\kappa = 10$, $\xi = \bar{x}$ for the univariate case and $\delta = 1$, $\nu = 4$, $\Psi = S$, $\kappa = 10$, $\xi = \bar{x}$ for the multivariate case. During the Gibbs sampling, the components are randomly permuted to simulate the label switching phenomenon.

In order to show the effectiveness of the new label switching method (MEM), we compare it to several existing label switching methods including the order restriction method (OR), the Kullback-Leibler based relabeling method of Stephens (1997) and Stephens (2000) (KL), the equivalence class representative relabeling method of Papastamoulis and Iliopoulos (2010) (ECR), and the data based relabeling method of Rodríguez and Walker (2012) (DATA). The Mean Deviation Error

$$\text{MDE} = \frac{1}{N}\sum_{i=1}^{N} \parallel \boldsymbol{\theta}_i^{\boldsymbol{\omega}_i} - \boldsymbol{\theta} \parallel$$

is used to measure the accuracy of the parameter estimates after the label switching problem is resolved, where $\parallel \cdot \parallel$ indicates the Euclidean distance. The mean and standard deviation of the MDE are computed from the 100 replicates.

For model C1, the OR method when applied to the mean parameters is quite successful because the two clusters of parameter estimates are well separated. See the first column of Fig. 3 for an example. Among the five relabeling methods, ECR has the worse performance according to the MDE (see Table 1). All other methods are comparable, but our MEM algorithm outperforms most others. For our MEM algorithm, using the first 500 instead of all 20,000 original MCMC samples for the KDE provides about the same relabeling results. This approves the online application of the MEM algorithm.

For model C2, the OR method applied to the mean parameters fails because two of the components have the same mean. Using the component probabilities or the variance parameters for the label switching purpose will not help either. The next worst performed method is the KL method. See Fig. 3 for an example. According to the MDE given in Table 2, our MEM algorithm beats all other methods by at least 9% in total MDE (for $\boldsymbol{\theta}$). Again, an online application of the MEM algorithm using the first 500 original MCMC samples for KDE is proved to be successful.

For model C3, since the data are bivariate, the OR method is not expected to work well when applied to a single dimension. It is also not clear how to implement the OR method to multi-dimensional data. Evidence can be found in Fig. 3 and Table 3. To show the flexibility of the MEM algorithm, only the two-dimensional mean parameters are used when the MEM algorithm is implemented to solve the label switching problem. As can be seen from Table 3, all methods except the OR method have very similar performance in the MDE.

### 5.2. Real data applications

For the acidity data of Crawford et al. (1992) and Crawford (1994), the fish data of Titterington et al. (1985), and the old faithful data of Stephens (1997) and Dellaportas and Papageorgious (2006), the MEM algorithm is implemented to solve the label switching problem. For the old faithful data, only the two-dimensional mean parameters are used for the label switching purpose. Figure 4 shows a great success of the MEM algorithm for all three examples. Table 4 details the posterior means and standard deviations of the parameter estimates for the three examples. An online application of the MEM algorithm using the first 500 original MCMC samples gives very similar results to Fig. 4 and Table 4.

To reconstruct the mixture models, the plug-in density estimates are obtained using the posterior means of the parameter estimates. Figure 5 superimposes the plug-in density estimates over their corresponding histograms or scatter plot. As can be seen, the mixtures have

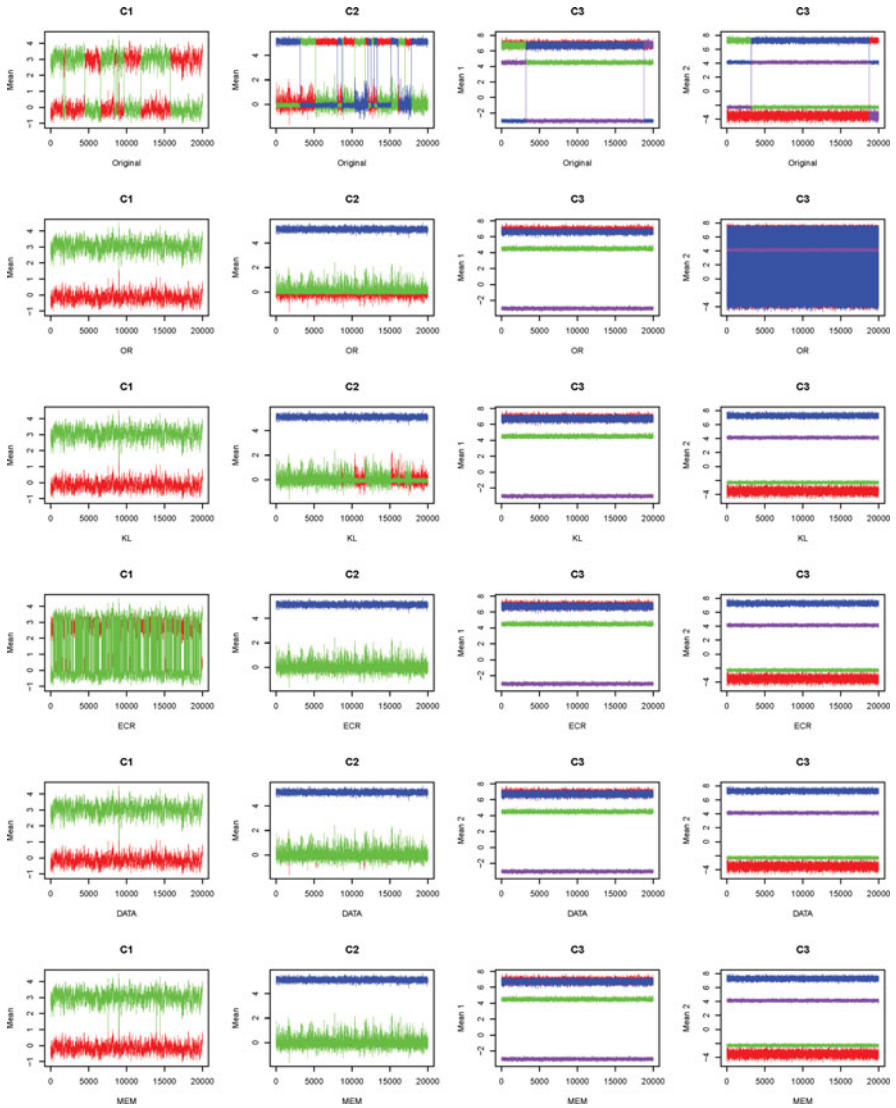


**Figure 3.** An example of trace plots of the mean parameters from the original MCMC samples and the relabeled samples for all three models C1, C2, and C3.

been successfully reconstructed and fit the data very well. This proves a great success of the MEM algorithm.

## 6. Discussions

In this article, a new label switching method using the modal clustering algorithm of Li et al. (2007) is proposed for Bayesian mixture model analysis. This method does not depend on the Bayesian model used to generate the MCMC samples so it can also be applied to Frequentist settings such as simulations and bootstraps. In addition, the new method has a nice geometric interpretation based on the permutation symmetry of the posterior. The numerical studies show that the new method is very effective in detecting the major modes of the kernel density when the tuning parameters are appropriately adjusted.

**Table 1.** Summaries of the MDE of parameter estimates for model C1 by different relabeling methods.

| Method | $\pi$ | | $\mu$ | | $\sigma^2$ | | $\theta$ | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| OR | 0.190 | 0.062 | 0.636 | 0.190 | 0.940 | 0.244 | 1.189 | 0.291 |
| KL | 0.190 | 0.062 | 0.665 | 0.232 | 0.938 | 0.244 | 1.214 | 0.327 |
| ECR | 0.190 | 0.062 | 0.991 | 0.595 | 1.030 | 0.326 | 1.535 | 0.641 |
| DATA | 0.190 | 0.062 | 0.650 | 0.203 | 0.935 | 0.242 | 1.198 | 0.301 |
| MEM (20K) | 0.190 | 0.062 | 0.638 | 0.190 | 0.939 | 0.244 | 1.190 | 0.292 |
| MEM (500) | 0.190 | 0.062 | 0.639 | 0.191 | 0.940 | 0.246 | 1.191 | 0.294 |

MEM (20K) stands for the MEM algorithm using the 20,000 original MCMC samples for the KDE, while MEM (500) uses the first 500 MCMC samples for the KDE.

**Table 2.** Summaries of the MDE of parameter estimates for model C2 by different relabeling methods.

| | $\pi$ | | $\mu$ | | $\sigma^2$ | | $\theta$ | |
|---|---|---|---|---|---|---|---|---|
| Method | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| OR | 0.167 | 0.059 | 0.609 | 0.299 | 2.523 | 0.445 | 2.678 | 0.430 |
| KL | 0.167 | 0.059 | 0.615 | 0.303 | 2.308 | 0.556 | 2.470 | 0.557 |
| ECR | 0.167 | 0.059 | 0.610 | 0.300 | 2.130 | 0.588 | 2.295 | 0.577 |
| DATA | 0.167 | 0.059 | 0.609 | 0.299 | 2.137 | 0.566 | 2.299 | 0.560 |
| MEM (20K) | 0.167 | 0.059 | 0.609 | 0.299 | 1.923 | 0.470 | 2.090 | 0.482 |
| MEM (500) | 0.167 | 0.059 | 0.609 | 0.299 | 1.924 | 0.470 | 2.091 | 0.482 |

MEM (20K) stands for the MEM algorithm using the 20,000 original MCMC samples for the KDE, while MEM (500) uses the first 500 MCMC samples for the KDE.

While the new label switching method outperforms all four existing methods in the MDE, it suffers from a major drawback of a high computational burden. Table 5 summarises the per dataset computational time in seconds. The average computational time is in the order of OR $<$ ECR $<$ DATA $<$ KL $<$ MEM. When the 20,000 original MCMC samples are used for the KDE in the MEM algorithm, the computational time is in the magnitude of a few hours per dataset. Hopefully, this high computational burden can be compensated by an upgrade in the hardware. In our simulation studies, a PC with 24 CPU cores is utilized so the per dataset computational time is reduced to about 5–30 minutes. However, an online application of the MEM algorithm using the first 500 original MCMC samples to fit the KDE does help cut the computational time into a fraction while maintaining the accuracy.

Nevertheless, in some situations where the number of components $m$ is large, the MEM algorithm can still be diffiicult to handle computationally because of a large training set. Here are some thoughts of possible refinements to the method when $m$ is indeed large. In practice, if the MCMC only visited a few, out of $m!$, model regions, it may be unnecessary to permute

**Table 3.** Summaries of the MDE of parameter estimates for model C3 by different relabeling methods.

| | $\pi$ | | $\mu$ | | $\Sigma$ | | $\theta$ | |
|---|---|---|---|---|---|---|---|---|
| Method | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| OR | 0.066 | 0.015 | 1.704 | 1.607 | 3.414 | 0.894 | 4.131 | 1.695 |
| KL | 0.066 | 0.015 | 0.745 | 0.142 | 3.076 | 0.686 | 3.191 | 0.673 |
| ECR | 0.066 | 0.015 | 0.745 | 0.142 | 3.076 | 0.686 | 3.191 | 0.673 |
| DATA | 0.066 | 0.015 | 0.745 | 0.142 | 3.076 | 0.686 | 3.191 | 0.673 |
| MEM (20 K) | 0.066 | 0.015 | 0.745 | 0.142 | 3.076 | 0.686 | 3.191 | 0.673 |
| MEM (500) | 0.066 | 0.015 | 0.745 | 0.142 | 3.076 | 0.686 | 3.191 | 0.673 |

MEM (20 K) stands for the MEM algorithm using the 20,000 original MCMC samples for the KDE, while MEM (500) uses the first 500 MCMC samples for the KDE

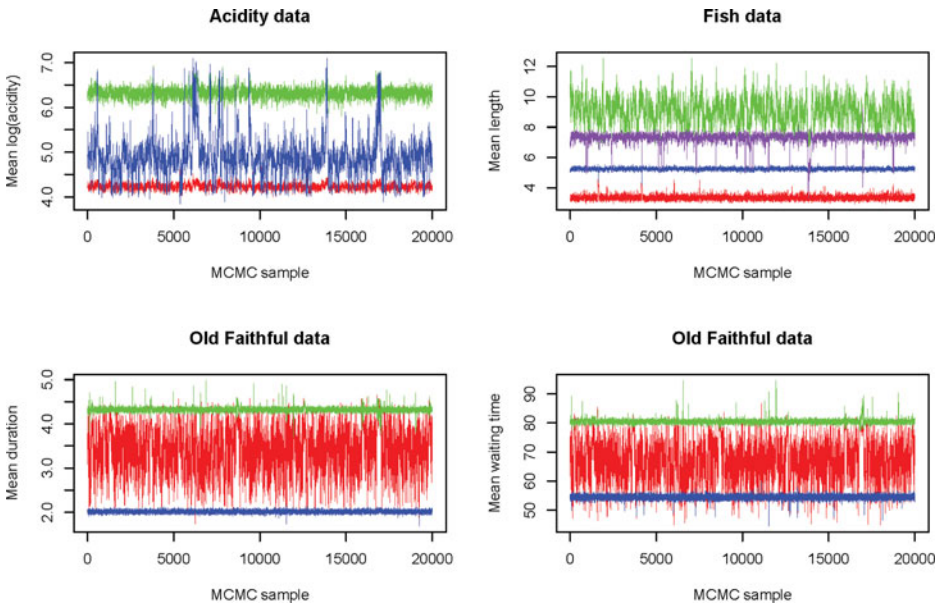


**Figure 4.** Trace plots of the mean parameters using the 20,000 MCMC samples from the acidity data, the fish data, and the old faithful data, respectively. The MEM algorithm is implemented to solve the label switching problem.

each original MCMC sample all $m!$ ways. For example, when the MCMC only visited two symmetric modal regions, we could permute each MCMC sample once to match the other label so that the permuted samples are only twice (instead of $m!$ times) the size of the original. The MEM algorithm can then be applied to such permuted MCMC samples to find the two modal regions. But, how do we know how many regions were visited by the MCMC before any label switching algorithm is applied? Well, as long as there are enough original MCMC samples from each visited modal region, we can run the MEM algorithm on the original samples. If one wishes to refine the results, we can permute the MCMC samples according to how many regions that were found in the first run and reapply the MEM algorithm to the permuted samples.

**Table 4.** Posterior means and standard deviations of parameter estimates for the acidity data, the fish data, and the old faithful data, respectively

| | $\pi$ | | $\mu$ | | $\sigma^2/\Sigma$ | |
|---|---|---|---|---|---|---|
| Data | Mean | SD | Mean | SD | Mean | SD |
| Acidity | 0.442 | 0.089 | 4.237 | 0.084 | 0.066 | 0.027 |
| | 0.191 | 0.083 | 4.906 | 0.413 | 0.132 | 0.098 |
| | 0.367 | 0.063 | 6.324 | 0.119 | 0.215 | 0.071 |
| Fish | 0.115 | 0.026 | 3.364 | 0.154 | 0.246 | 0.115 |
| | 0.478 | 0.060 | 5.251 | 0.115 | 0.322 | 0.109 |
| | 0.232 | 0.078 | 7.289 | 0.326 | 0.531 | 0.338 |
| | 0.174 | 0.085 | 9.028 | 0.840 | 2.509 | 1.068 |
| Old Faithful | 0.339 | 0.034 | 2.013 | 0.031 | 0.070 0.568 | 0.060 0.666 |
| | | | 54.45 | 0.700 | 0.568 35.79 | 0.666 8.611 |
| | 0.075 | 0.074 | 3.405 | 0.479 | 0.414 4.793 | 0.258 3.289 |
| | | | 67.01 | 6.732 | 4.793 75.39 | 3.289 49.48 |
| | 0.586 | 0.081 | 4.323 | 0.052 | 0.157 0.716 | 0.091 1.080 |
| | | | 80.49 | 0.772 | 0.716 32.42 | 1.080 12.93 |

The MEM algorithm is implemented to solve the label switching problem.

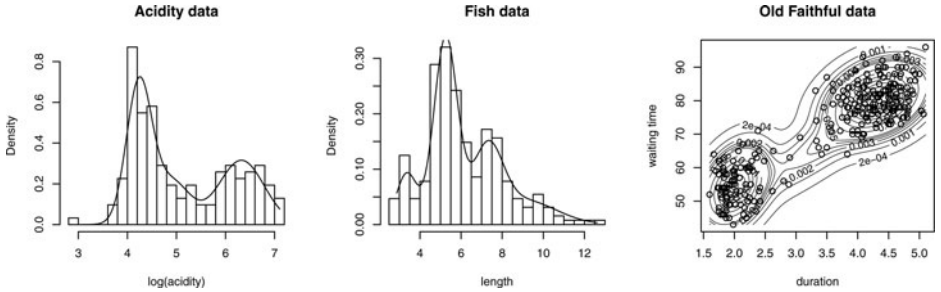


**Figure 5.** Plug-in density estimates (using posterior means) for the acidity data, the fish data, and the old faithful data, respectively.

**Table 5.** Summaries of per dataset computational time in seconds for the simulation studies

| Model | C1 | | C2 | | C3 | |
|---|---|---|---|---|---|---|
| Method | Mean | SD | Mean | SD | Mean | SD |
| OR | 1.79 | 0.16 | 1.92 | 0.23 | 1.87 | 0.17 |
| KL | 37.4 | 13.4 | 116 | 38.4 | 864 | 89.2 |
| ECR | 13.3 | 1.12 | 18.8 | 1.73 | 107 | 9.89 |
| DATA | 26.8 | 2.42 | 35.0 | 3.18 | 112 | 12.2 |
| MEM (20K) | 6940 | 4389 | 18797 | 7754 | 32137 | 4953 |
| MEM (500) | 237 | 155 | 497 | 218 | 588 | 56.3 |

A major difficulty in mixture model analysis is to determine the number of components. Although in our simulations and real data applications, the number of components is assumed known in advance, it is most often not the case in real life. A mis-specification of the number of components can lead to great difficulties in the MCMC sampling and thus the relabeling procedure. Even if the number of components is correctly specified, most Bayesian sampling procedures, such as the one in Section 3, do not always guarantee that all MCMC samples fall within the $m!$ symmetric regions. It happens that some of them may seem outside the $m!$ symmetric regions due to random noises. Most existing relabeling methods, for example, the OR method, ignore this fact and conduct the relabeling process anyway. However, if this does happen, the MEM algorithm will complain that more than $m!$ modes are detected. This may seem an instability issue, but we see it as an advantage because that extra information can be utilized to double check the modeling and the Bayesian sampling process.

## Acknowledgments

The authors thank two anonymous referees for their insightful comments that led to an improved manuscript. Dr. Yao's research is supported by NSF grant DMS-1461677 and the Department of Energy with Award No. 10006272.

## References

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications*. Boca Raton, FL: Chapman and Hall/CRC.

Celeux, G. (1998). Bayesian inference for mixtures: The label switching problem. In: Payne, R., Green, P., eds. *In Compstat 98-Proc. in Computational Statistics*, Heidelberg, Germany: Physica, pp. 227–232.

Chung, H., Loken, E., Schafer, J. L. (2004). Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician* 58:152–158.

Crawford, S. L. (1994). An application of the laplace method to finite mixture distributions. *Journal of American Statistical Association* 89:259–267.

Crawford, S. L., Degroot, M. H., Kadane, J. B., Small, M. J. (1992). Modeling lake-chemistry distributions-approximate bayesian methods for estimating a finite-mixture model. *Technometrics* 34:441–453.

Cron, A., West, M. (2011). Efficient classification-based relabeling in mixture models. *The American Statistician* 65:16–20.

Dellaportas, P., Papageorgious, I. (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing* 16(1):57–68.

Dellaportas, P., Stephens, D. A., Smith, A. F. M., Guttman, I. (1996). A comparative study of perinatal mortality using a two-component mixture model. In: Berry, D., Stangl, D., eds., *Bayesian Biostatistics*. Boca Raton, FL: CRC Press. pp. 601–616.

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

Diebolt, J., Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of Royal Statistical Association: Series B* 56:363–375.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of American Statistical Association* 96:194–209.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple mcmc works. *Computational Statistics and Data Analysis* 51:3529–3550.

Grün, B., Leisch, F. (2009). Dealing with label switching in mixture models under genuine multimodality. *Journal of Multivariate Analysis* 100:851–861.

Hurn, M., Justel, A., Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* 12:55–79.

Jasra, A., Holmes, C. C., A., S. D., (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science* 20:50–67.

Li, J., Ray, S., Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research* 8:1687–1723.

Lindsay, B. G. (1995). Mixture models: Theory, geometry, and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics v 5*, Hayward, CA. Institure of Mathematical Statistics.

Marin, J.-M., Mengersen, K. L., Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In: Dey, D., Rao, C., eds., *Handbook of Statistics: Volume 25*. Amsterdam, North Holland: Elsevier, pp. 459–507.

McLachlan, G. J., Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Papastamoulis, P., Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics* 19:313–331.

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

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

Rodríguez, C. E., Walker, S. G. (2012). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics* 23(1):25–45.

Scott, D. W. (1992). *Multivariate Density Estimation*. Chichester, New York: John Wiley & Sons.

Sperrin, M., Jaki, T., Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing* 20(3):357–366.

Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions*. PhD dissertation, Department of Statistics, Oxford, UK: University of Oxford.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of Royal Statistical Association: Series B* 62:795–809.

Titterington, D. M., Smith, A. F. M., Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Yao, W. (2012a). Bayesian mixture labeling and clustering. *Communications in Statistics - Theory and Methods* 41:403–421.

Yao, W. (2012b). Model based labeling for mixture models. *Statistics and Computing* 22:337–347.

Yao, W., Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of American Statistical Association* 104:758–767.