Stat Comput (2012) 22:337–347
DOI 10.1007/s11222-010-9226-8

Model based labeling for mixture models

Weixin Yao

Received: 11 January 2010 / Accepted: 28 December 2010 / Published online: 19 January 2011

Abstract Label switching is one of the fundamental problems for Bayesian mixture model analysis. Due to the permutation invariance of the mixture posterior, we can consider that the posterior of a m-component mixture model is a mixture distribution with $m!$ symmetric components and therefore the object of labeling is to recover one of the components. In order to do labeling, we propose to first fit a symmetric $m!$-component mixture model to the Markov chain Monte Carlo (MCMC) samples and then choose the label for each sample by maximizing the corresponding classification probabilities, which are the probabilities of all possible labels for each sample. Both parametric and semi-parametric ways are proposed to fit the symmetric mixture model for the posterior. Compared to the existing labeling methods, our proposed method aims to approximate the posterior directly and provides the labeling probabilities for all possible labels and thus has a model explanation and theoretical support. In addition, we introduce a situation in which the "ideally" labeled samples are available and thus can be used to compare different labeling methods. We demonstrate the success of our new method in dealing with the label switching problem using two examples.

Keywords Bayesian mixtures · Labeling probabilities · Label switching · Markov chain Monte Carlo · Mixture model

W. Yao (✉)
Department of Statistics, Kansas State University, Manhattan, KS 66506, USA
e-mail: wxyao@ksu.edu

1 Introduction

Suppose $\mathbf{x} = (x_1, \ldots, x_n)$ are independent observations from an m-component mixture density

$$p(x; \boldsymbol{\theta}) = \pi_1 f(x; \lambda_1) + \pi_2 f(x; \lambda_2) + \cdots + \pi_m f(x; \lambda_m),$$

where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_m, \lambda_1, \ldots, \lambda_m)^T$, $f(\cdot)$ is some parametric component density or mass function, λ_j is the component specific parameter, and π_j is the proportion of j^{th} component with $\sum_{j=1}^{m} \pi_j = 1$. The likelihood for $\mathbf{x}$ is

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} \{\pi_1 f(x_i; \lambda_1) + \pi_2 f(x_i; \lambda_2) + \cdots + \pi_m f(x_i; \lambda_m)\}. \quad (1)$$

For a general introduction to mixture models, see Lindsay (1995), Böhning (1999), McLachlan and Peel (2000), and Frühwirth-Schnatter (2006).

For any permutation $\boldsymbol{\omega} = (\boldsymbol{\omega}(1), \ldots, \boldsymbol{\omega}(m))$ of the identity permutation $(1, \ldots, m)$, define the corresponding permutation of the parameter vector $\boldsymbol{\theta}$ by

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

Noticing that $L(\boldsymbol{\theta}^{\boldsymbol{\omega}}; \mathbf{x})$ is numerically the same as $L(\boldsymbol{\theta}; \mathbf{x})$ for any permutation $\boldsymbol{\omega}$, hence if $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator (MLE), $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$ is the MLE for any permutation $\boldsymbol{\omega}$. This is so-called label switching problem.

The label switching problem also occurs in Bayesian mixtures. Let $\pi(\boldsymbol{\theta})$ be the prior for mixture model, the posterior distribution of $\boldsymbol{\theta}$ is equal to $p(\boldsymbol{\theta} \mid \mathbf{x}) = \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}; \mathbf{x}) / p(\mathbf{x})$, where $p(\mathbf{x})$ is the marginal density for $\mathbf{x} = (x_1, \ldots, x_n)$ and $L(\boldsymbol{\theta}; \mathbf{x})$ is defined in (1). If we do not have prior information that distinguishes between the components of a mixture model, i.e., $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}^{\boldsymbol{\omega}})$ for any permutation $\boldsymbol{\omega}$, then

$p(\boldsymbol{\theta} \mid \mathbf{x}) = p(\boldsymbol{\theta}^{\boldsymbol{\omega}} \mid \mathbf{x})$ for any permutation $\boldsymbol{\omega}$ and thus the posterior $p(\boldsymbol{\theta} \mid \mathbf{x})$ has $m!$ permutation symmetric maximal modes, with each of them associated with a modal region such that each modal region is a permutation image of the other and thus can be considered as one well labeled parameter space. Among each modal region there is a well defined highest poster density region such that their posterior is greater or equal to a fixed value c, say. Therefore, we also have $m!$ permutation symmetric highest poster density regions for any fixed cut point c. See Yao and Lindsay (2009) for more detail about the modal region and highest posterior density region for mixture model. Due to the permutation symmetry, the marginal posterior distributions for the parameters are identical for each mixture component. Hence, it is meaningless to draw inference, relating to individual components, directly from Markov chain Monte Carlo (MCMC) samples using ergodic averaging before solving the label switching problem.

Given the MCMC samples $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$, the latent "true" labels $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$ are defined such that $\boldsymbol{\theta}_1^{\boldsymbol{\omega}_1}, \ldots, \boldsymbol{\theta}_N^{\boldsymbol{\omega}_N}$ are all in the same modal region and therefore have the same label meaning. The aim of labeling is to recover the latent labels $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$. Since each modal region defines a set of latent labels, there are essentially $m!$ sets of latent "true" labels and they are identifiable up to the same permutation. Therefore, one only needs to find one of the modal regions and the corresponding set of latent "true" labels. Note that, in fact, there is no unique way to write the posterior as a mixture of distributions with $m!$ components and each component is a permutation version of the other (Papastamoulis and Iliopoulos 2010). However, as Papastamoulis and Iliopoulos (2010) stated, an efficient solution to label switching is to make each of these components correspond to one of the $m!$ symmetric modal regions (highest posterior density areas). Here, we also use such modal regions to define the latent "true" labels. Most of the existing labeling methods try to directly find the labels such that the labeled samples are as similar as possible based on some clustering criteria. Usually, different clustering criteria will give different labeling results and there is lack of widely accepted criteria.

In this article, we propose a novel model based labeling method by approximating the posterior using a symmetric mixture model. In our labeling procedure, we propose to first fit a symmetric $m!$-component mixture model to the MCMC samples and estimate the classification probabilities of all $m!$ possible labels for each sample. Then, the labels for each sample are chosen by maximizing the corresponding labeling/classification probabilities. Considering the label $\boldsymbol{\omega}_t$ as a missing modal region indicator for $\boldsymbol{\theta}_t$, one knows that the posterior distribution can be considered as a mixture distribution with $m!$ components, each component corresponding to one of the $m!$ symmetric modal regions. Therefore, to solve the label switching, one only needs to find one of the $m!$ components/modal regions (all other components can be derived by permutations), i.e., determine which component/modal region each sample belongs to. Compared to the existing labeling methods, our proposed method aims to approximate the posterior directly and provide the labeling probabilities for all possible labels. In addition, our labeling method has model explanation and some theoretical support. We propose both parametric and semi-parametric ways to fit the symmetric mixture model to the MCMC samples.

Comparing different labeling methods is always a difficult issue since practically one never knows the true labels. In this article, we introduce a situation in which the "ideally" labeled samples are available in some sense and thus can be used to compare different labeling methods. Using two examples, we demonstrate the effectiveness of our proposed model based labeling method in removing the label switching in the raw MCMC samples.

Many methods have been proposed to deal with the label switching problem in Bayesian analysis. The easiest way to solve the label switching is to use an explicit parameter constraint so that only one permutation can satisfy it. See Diebolt and Robert (1994) and Richardson and Green (1997). Another popular labeling method is relabeling algorithm (Celeux 1998; Stephens 2000), which is based on minimizing a Monte Carlo risk. Stephens (2000) suggested a particular choice of loss function based on the Kullback-Liebler (KL) divergence. We will refer to this particular relabeling algorithm as *KL algorithm*. Yao and Lindsay (2009) proposed the PM(ALG) method to label the samples based on the posterior modes they are associated with when they are used as the starting points for an ascending algorithm of the posterior. The PM(ALG) method is an online algorithm and does not require one to compare $m!$ permutations when doing labeling, which makes it much faster than some other relabeling algorithms. Sperrin et al. (2010) developed several probabilistic relabeling algorithms by extending the probabilistic relabeling of Jasra (2005). Similar to our method, Sperrin et al. (2010) assigned labeling probabilities for all possible labels to account for the uncertainty in the relabeling process. However, their method assigns the labels to the allocation vectors directly instead of the MCMC samples, which makes the method depend on the assumption that there is no label switching between the allocation vectors and the corresponding MCMC samples, which is not necessarily true for all the samples. Papastamoulis and Iliopoulos (2010) proposed an artificial allocations based solution to the label switching problem. One of the advantages of their method is that it requires small computational effort compared to many other sophisticated solutions. However, similar to Sperrin et al. (2010), their method also assigns the labels to the allocation vectors directly. Our proposed methods deal with the MCMC samples directly and come from different statistical perspective.

Other labeling methods include, for example, Celeux et al. (2000), Frühwirth-Schnatter (2001), Hurn et al. (2003), Chung et al. (2004), Marin et al. (2005), Geweke (2007), and Grun and Leisch (2009). Jasra et al. (2005) provided a good review about the existing methods to solve the label switching problem in Bayesian mixture modelling.

The rest of the paper is organized as follows. Section 2 introduces our new model based labeling method. Both parametric and semi-parametric ways are introduced to fit the mixture model to the MCMC samples. In Sect. 3, we use a simulation example and a real data set to compare our new labeling method with some of other existing methods. We summarize our proposed labeling method and discuss some future research work in Sect. 4.

2 Introduction of model based labeling

For simplicity of explanation of our new model based labeling method, let us first consider the situation when $m = 2$. When $m = 2$, there will be two symmetric modal regions of the posterior density. We can consider each of them as the region for the "true" labeled parameter space and having the labeled posterior. The aim of labeling is to recover one of the modal regions, which will be called reference modal region.

Suppose that the reference modal region has the well labeled posterior density $g(\boldsymbol{\theta} \mid \mathbf{x})$. Then, if any $\boldsymbol{\theta}$ comes from the reference modal region (i.e., $\boldsymbol{\theta}$ has identical label $(1, 2)$), it has the density $g(\boldsymbol{\theta} \mid \mathbf{x})$; if $\boldsymbol{\theta}$ comes from the other model region, i.e., $\boldsymbol{\theta}^{\boldsymbol{\omega}^*}$ comes from the reference modal region, where $\boldsymbol{\omega}^* = (2, 1)$, then $\boldsymbol{\theta}$ has the density $g(\boldsymbol{\theta}^{\boldsymbol{\omega}^*} \mid \mathbf{x})$. Note that marginally the probabilities for the two possible labels are equal and both are $1/2$. Hence, we have

$$p\{\boldsymbol{\theta}_t \mid \mathbf{x}, \boldsymbol{\omega}_t = (1, 2)\} = g(\boldsymbol{\theta}_t \mid \mathbf{x});$$

$$p\{\boldsymbol{\theta}_t \mid \mathbf{x}, \boldsymbol{\omega}_t = (2, 1)\} = g(\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*} \mid \mathbf{x});$$

$$P\{\boldsymbol{\omega}_t = (1, 2)\} = P\{\boldsymbol{\omega}_t = (2, 1)\} = 1/2,$$

where $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ are all MCMC samples. So, the original full posterior density for $\boldsymbol{\theta}_t$, without knowing its true label $\boldsymbol{\omega}_t$, has a mixture form

$$p(\boldsymbol{\theta}_t \mid \mathbf{x}) = \frac{1}{2} g(\boldsymbol{\theta}_t \mid \mathbf{x}) + \frac{1}{2} g(\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*} \mid \mathbf{x}), \tag{2}$$

where $\boldsymbol{\omega}^* = (2, 1)$. The model (2) will be called symmetric mixture model, due to the permutation symmetry of the mixture components. The ideal way to solve the label switching is to find $g(\boldsymbol{\theta} \mid \mathbf{x})$ and do Bayesian inference based on it instead of the original unlabeled posterior $p(\boldsymbol{\theta} \mid \mathbf{x})$.

Finding the labels $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$, corresponding to the reference model region, is equivalent to determining whether $\boldsymbol{\theta}_t$ or $\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*}$ is from the first component that has the density $g(\boldsymbol{\theta} \mid \mathbf{x})$. Let $\boldsymbol{\Delta} = \{\boldsymbol{\theta}_t, \boldsymbol{\theta}_t^{\boldsymbol{\omega}^*},\ t = 1, \ldots, N\}$, which includes both of the original samples and their permutations. In order to do labeling, we first fit the mixture model (2) to the "dataset" $\boldsymbol{\Delta}$ and estimate $g(\boldsymbol{\theta} \mid \mathbf{x})$ by $\hat{g}(\boldsymbol{\theta} \mid \mathbf{x})$, say. Then, we find the classification/labeling probabilities $(\hat{p}_{t1}, \hat{p}_{t2})$ for each $\boldsymbol{\theta}_t$, where

$$\hat{p}_{t1} = \frac{\hat{g}(\boldsymbol{\theta}_t \mid \mathbf{x})}{\hat{g}(\boldsymbol{\theta}_t \mid \mathbf{x}) + \hat{g}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*} \mid \mathbf{x})},$$

$$\hat{p}_{t2} = 1 - \hat{p}_{t1}, \quad t = 1, \ldots, N.$$

The estimate $\hat{p}_{t1}$ can be considered as the probability that $\boldsymbol{\theta}_t$ comes from the reference modal region (i.e., the label of $\boldsymbol{\theta}_t$ is $\boldsymbol{\omega}_t = (1, 2)$) and $\hat{p}_{t2}$ can be considered as the probability that $\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*}$ comes from the reference modal region (i.e., $\boldsymbol{\omega}_t = (2, 1)$).

We can then choose $\boldsymbol{\omega}_t$ by maximizing the labeling probabilities $\{\hat{p}_{t1}, \hat{p}_{t2}\}$, i.e., assign the identity permutation label $\hat{\boldsymbol{\omega}}_t = (1, 2)$ if $\hat{p}_{t1} \geq \hat{p}_{t2}$ and assign the permutation label $\hat{\boldsymbol{\omega}}_t = (2, 1)$ if $\hat{p}_{t1} < \hat{p}_{t2}$. The Bayesian inference can then be done based on the labeled samples $\{\boldsymbol{\theta}_t^{\hat{\boldsymbol{\omega}}_t}, t = 1, \ldots, N\}$.

Next, we will provide both parametric and semi-parametric ways to fit the symmetric mixture model (2) based on the MCMC samples $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$.

2.1 Parametric labeling

From the asymptotic theory for the posterior distribution, see Walker (1969) and Frühwirth-Schnatter (2006, Sects. 1.3, 2.4.3, 3.3), one knows that when sample size is large, the suitably labeled MCMC samples should, approximately, follow the normal distribution, i.e., there exist permutations $\{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N\}$ such that $\{\boldsymbol{\theta}_1^{\boldsymbol{\omega}_1}, \ldots, \boldsymbol{\theta}_N^{\boldsymbol{\omega}_N}\}$ follows approximately a normal distribution. Therefore, $g(\boldsymbol{\theta} \mid \mathbf{x})$ in (2) can be approximated by a normal density and the model (2) can be approximated by a normal mixture

$$p(\boldsymbol{\theta} \mid \mathbf{x}) \approx \frac{1}{2} N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2} N(\boldsymbol{\theta}^{\boldsymbol{\omega}^*}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\omega}^* = (2, 1)$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the center and covariance matrix for the reference modal region, and $N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density function for $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The following paragraph describes how to fit the symmetric normal mixture model (3) using an EM algorithm.

Algorithm 1 (Model based labeling by normal mixture model (MBLNM)) Starting with the initial value $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$, in the $(k+1)^{\text{th}}$ step,

E step: compute the labeling probabilities

$$p_{t1}^{(k+1)} = \frac{N(\boldsymbol{\theta}_t; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})}{N(\boldsymbol{\theta}_t; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}) + N(\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})}, \tag{4}$$

$$p_{t2}^{(k+1)} = 1 - p_{t1}^{(k+1)}, \quad t = 1, \ldots, N;$$

M step: update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{N}\sum_{t=1}^{N}\left\{p_{t1}^{(k+1)}\boldsymbol{\theta}_t + p_{t2}^{(k+1)}\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*}\right\},$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{N}\sum_{t=1}^{N}\left\{p_{t1}^{(k+1)}(\boldsymbol{\theta}_t - \boldsymbol{\mu}^{(k+1)})(\boldsymbol{\theta}_t - \boldsymbol{\mu}^{(k+1)})^T + p_{t2}^{(k+1)}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*} - \boldsymbol{\mu}^{(k+1)})(\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*} - \boldsymbol{\mu}^{(k+1)})^T\right\}.$$

In the E step of Algorithm 1, if we use the hard label for $\boldsymbol{\theta}_t$, i.e., $p_{t1}^{(k+1)} = 0$ or 1, $t = 1, \ldots, N$, depending on whether the original classification probability is less or greater than 0.5, then the Algorithm 1 provides the same labeling results as the NORMLH method proposed by Yao and Lindsay (2009), which minimizes the following negative log-normal likelihood over $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ along with the missing labels $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$,

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\omega}) = N\log(|\boldsymbol{\Sigma}|) + \sum_{t=1}^{N}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_t} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_t} - \boldsymbol{\mu}). \tag{5}$$

The NORMLH method is computationally easy and fast and will be used to create the initial labels for other methods in our examples in Sect. 3.

Denote by $\{\hat{p}_{t1}, \hat{p}_{t2}\}$ the converged labeling probabilities from the last E-step of Algorithm 1. Based on the above labeling probabilities, we can then choose the label $\boldsymbol{\omega}_t$ by maximizing the labeling probabilities $\{\hat{p}_{t1}, \hat{p}_{t2}\}$.

In practice, sometimes, one might need to do some transformation of the original samples to make the labeled samples more close to the normal distribution before fitting the symmetric normal mixture model (3). For example, for the standard error parameters, one might take the log transformation. For the mixing proportion parameters, one might take log odds transformation.

2.2 Semi-parametric labeling

In many cases, especially when sample size is small, the labeled samples may not be approximated by normal distribution very well. In this section, we propose a way to fit the mixture model (2) without any parametric assumption about $g(\boldsymbol{\theta} \mid \mathbf{x})$. We will call such mixture model a semi-parametric mixture model due to the symmetric restriction (equal mixing proportions and permutation symmetric component density functions). By extending the semi-parametric EM algorithm proposed by Bordes et al. (2007) and Benaglia et al. (2009), we propose the following EM-like algorithm to fit the model (2).

Algorithm 2 (Model based labeling by semi-parametric mixture model (MBLSP)) Starting with the initial density estimate $g^{(0)}(\boldsymbol{\theta} \mid \mathbf{x})$, in the $(k+1)^{\text{th}}$ step,

E step: compute the labeling probabilities

$$p_{t1}^{(k+1)} = \frac{g^{(k)}(\boldsymbol{\theta}_t \mid \mathbf{x})}{g^{(k)}(\boldsymbol{\theta}_t \mid \mathbf{x}) + g^{(k)}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*} \mid \mathbf{x})}, \tag{6}$$
$$p_{t2}^{(k+1)} = 1 - p_{t1}^{(k+1)}, \quad t = 1, \ldots, N.$$

Nonparametric step: update $g(\boldsymbol{\theta} \mid \mathbf{x})$

$$g^{(k+1)}(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{1}{N}\sum_{t=1}^{N}\left\{p_{t1}^{(k+1)}K_{\mathbf{H}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}) + p_{t2}^{(k+1)}K_{\mathbf{H}}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}^*} - \boldsymbol{\theta})\right\}, \tag{7}$$

where $K_{\mathbf{H}}(\boldsymbol{\theta}) = \det(\mathbf{H})^{-1}K(\mathbf{H}^{-1}\boldsymbol{\theta})$, $\mathbf{H}$ is a bandwidth matrix (nonsingular), and $K(\boldsymbol{\theta})$ is a multivariate kernel density.

It is well known that the choice of the multivariate kernel $K(\cdot)$ is not very critical for kernel density estimate (see, for example, (Scott 1992)). In this article, we will simply use multivariate Gaussian kernel for $K(\cdot)$. The choice of the bandwidth $\mathbf{H}$ will be discussed in Sect. 3.

The initial density estimate for $g(\boldsymbol{\theta} \mid \mathbf{x})$ can be kernel density estimate based on some initial labels (such as order constraint labels or NORMLH labels (Yao and Lindsay 2009)) or the density estimate by the MBLNM method using Algorithm 1. The stopping rule for this algorithm can be based on the difference of labeling probabilities for two consecutive EM iterations.

2.3 More than two components

More generally, when there are m components, there will be $m!$ symmetric modal regions. Suppose $g(\boldsymbol{\theta} \mid \mathbf{x})$ is the posterior density for the reference modal region. Then, the posterior distribution $p(\boldsymbol{\theta} \mid x)$ is a mixture with $m!$ components

$$p(\boldsymbol{\theta} \mid x) = \frac{1}{m!}\sum_{j=1}^{m!}g(\boldsymbol{\theta}^{\boldsymbol{\omega}_{(j)}} \mid \mathbf{x}), \tag{8}$$

where $\{\boldsymbol{\omega}_{(1)}, \ldots, \boldsymbol{\omega}_{(m!)}\}$ are the $m!$ permutations of $(1, \ldots, m)$ with $\boldsymbol{\omega}_{(1)} = (1, 2, \ldots, m)$, the identity permutation. Note that we only need to estimate one component density, say $g(\boldsymbol{\theta} \mid \mathbf{x})$ of the first component, for model (8) since all others are just its permuted versions.

The model based labeling algorithms introduced in Sect. 2.1 and 2.2 can be easily extended to the situation when the number of components is larger than two. For example, for the symmetric normal mixture model, the E step

of Algorithm 1 is now

$$p_{tj}^{(k+1)} = \frac{N(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})}{\sum_{l=1}^{m!} N(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(l)}}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})},$$
$$t = 1, \ldots, N,\ j = 1, \ldots, m!,$$

and the M step is now

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{m!} p_{tj}^{(k+1)} \boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}},$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{m!} p_{tj}^{(k+1)} (\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}} - \boldsymbol{\mu}^{(k+1)}) \times (\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}} - \boldsymbol{\mu}^{(k+1)})^T. \quad (9)$$

In addition, for the symmetric semi-parametric mixture model, the E step in Algorithm 2 is now

$$p_{tj}^{(k+1)} = \frac{g^{(k)}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}} \mid \mathbf{x})}{\sum_{l=1}^{m!} g^{(k)}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(l)}} \mid \mathbf{x})},$$
$$t = 1, \ldots, N,\ j = 1, \ldots, m!,$$

and the Nonparametric step is

$$g^{(k+1)}(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{m!} p_{tj}^{(k+1)} K_{\mathbf{H}}(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}} - \boldsymbol{\theta}).$$

Let $\boldsymbol{\Delta} = \{\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}}, t = 1, \ldots, N, j = 1, \ldots, m!\}$. The above two EM algorithms are equivalent to fitting the symmetric mixture model (8) to the data set $\boldsymbol{\Delta}$. The estimate $\hat{p}_{tj}$, derived from the converged E step, can be considered as the probability that $\boldsymbol{\theta}_t$ has the label $\boldsymbol{\omega}_{(j)}$, i.e., the probability that $\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}}$ is from the reference modal region.

Based on the labeling probabilities $\{\hat{p}_{tj}, t = 1, \ldots, N, j = 1, \ldots, m!\}$, we can choose the label $\boldsymbol{\omega}_t$ for $\boldsymbol{\theta}_t$ by maximizing $\{\hat{p}_{t1}, \ldots, \hat{p}_{tm!}\}$. For example, if $\hat{p}_{tk}$ maximizes $\{\hat{p}_{t1}, \ldots, \hat{p}_{tm!}\}$ for some k, then $\hat{\boldsymbol{\omega}}_t = \boldsymbol{\omega}_{(k)}$, where $\{\boldsymbol{\omega}_{(1)}, \ldots, \boldsymbol{\omega}_{(m!)}\}$ are the $m!$ possible permutations of $(1, \ldots, m)$.

3 Examples

In this section, we use a simulation example and a real data set to illustrate the effectiveness of our proposed two model based labeling methods (MBLNM and MBLSP) and compare them with order constraint labeling (OC) and Stephens' KL algorithm (KL). The OC method refers to the ordering constraint labeling on the mean parameters. For KL algorithm, we used the transportation algorithm to maximize over the permutations. We used the NORMLH labels as the initial labels for KL, MBLNM, and MBLSP. For comparison, we reported the number of different labels for each method that differed from MBLSP. In addition, we also introduce a situation in which the "ideally" labeled samples are available and thus can be used to compare different labeling methods.

To use the semi-parametric labeling MBLSP, we need to choose the bandwidth matrix $\mathbf{H}$ first. A good rule of thumb is to use a bandwidth matrix proportional to $\hat{\boldsymbol{\Sigma}}^{1/2}$, i.e., $\mathbf{H} = h\hat{\boldsymbol{\Sigma}}^{1/2}$, where $\hat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix based on the initial labeled samples. Using such a bandwidth corresponds to a transformation of the initial labeled samples, so that they have an identity covariance matrix. By assuming a multivariate normal distribution for the labeled samples, we can get the rule of thumb for the bandwidth matrix $\mathbf{H}$

$$\hat{\mathbf{H}} = N^{-1/(d+4)} \hat{\boldsymbol{\Sigma}}^{1/2}, \quad (10)$$

where d is the dimension of $\boldsymbol{\theta}$ (see Scott 1992, p. 152). As suggested by Benaglia et al. (2009), one might also use an iterative procedure in which the value of $\mathbf{H}$ is modified after the new labels. For simplicity, in all of our examples in this section, we used the bandwidth (10) for the MBLSP method. (We also tried some other bandwidths such as $0.5\hat{\mathbf{H}}$ and $1.5\hat{\mathbf{H}}$ and the labeling results were almost the same. Hence, empirically, the labeling results by MBLSP are not very sensitive to the choice of bandwidth $\hat{\mathbf{H}}$, which is sensible since our objective is the classification probabilities instead of the component density function itself.)

All the computations were done in Matlab 7.0 using a personal desktop with Intel Core 2 Quad CPU 2.40 GHz. It is known that the OC method is the fastest one and it takes no more than several seconds in our examples. Hence, we only reported the runtime for KL, MBLNM, and MBLSP.

Example 3.1 We generated 400 data points from $0.3N(0, 1) + 0.7N(0.5, 2^2)$. Based on this data set, we generated 5,000 MCMC samples (after initial burn-in) of component means, component proportions, and the unequal component variance. The MCMC samples were generated by Gibbs sampler with the priors given by Richardson and Green (1997). That is to assume

$$\boldsymbol{\pi} \sim D(\delta, \delta), \qquad \mu_j \sim N(\xi, \kappa^{-1}),$$
$$\sigma_j^{-2} \sim \Gamma(\alpha, \beta), \qquad \beta \sim \Gamma(g, h), \quad j = 1, 2,$$

where $D(\cdot)$ is Dirichlet distribution and $\Gamma(\alpha, \beta)$ is gamma distribution with mean α/β and variance α/β^2. Following the suggestion of Richardson and Green (1997), we let $\delta = 1$, ξ equal the sample mean of the observations, $\kappa = 1/R^2$, $\alpha = 2$, $g = 0.2$, and $h = 10/R^2$, where R is the range of the observations. Similar priors are used for the other example.

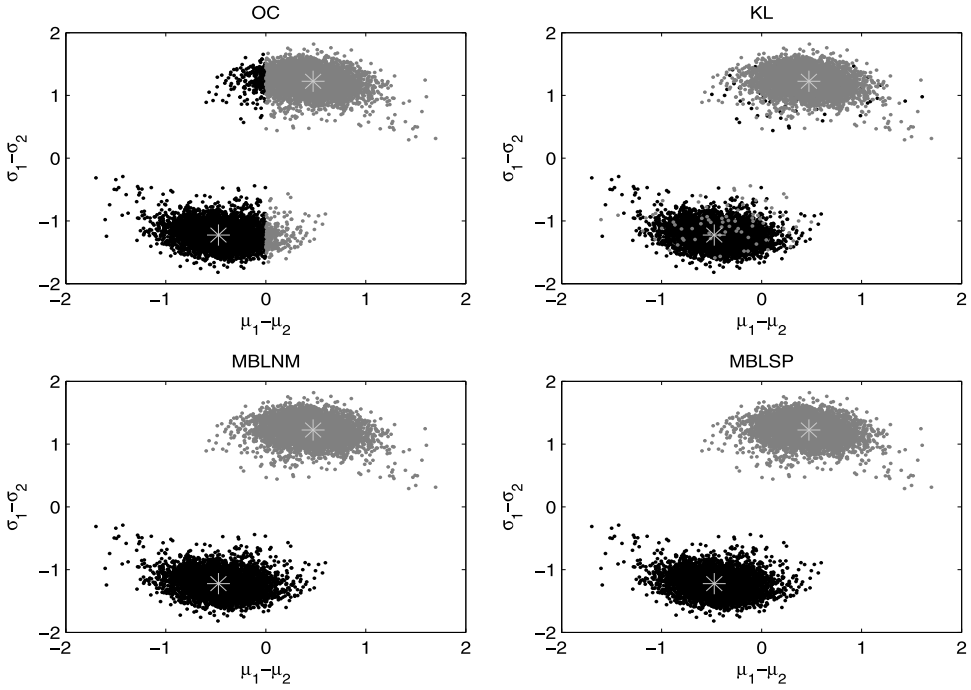

Fig. 1 Plots of $\sigma_1 - \sigma_2$ vs. $\mu_1 - \mu_2$ for different labeling methods in Example 3.1. The *black points* represent one set of labels and the *gray points* are the permuted samples. The *star points* are the posterior modes

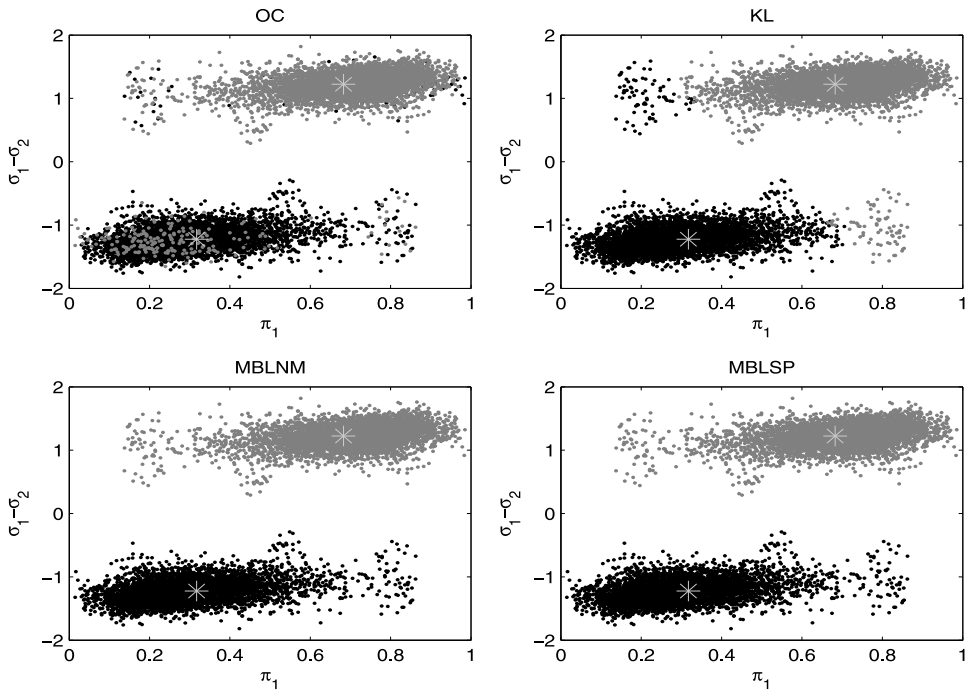

Fig. 2 Plots of $\sigma_1 - \sigma_2$ vs. π_1 for different labeling methods in Example 3.1

The total numbers of different labels between (OC, KL, MBLNM) and MBLSP were: 212, 76, and 0, respectively. Hence, in this example, MBLNM and MBLSP provided the same labeling results, and KL had closer labeling results to the MBLSP method than OC. The runtime for KL, MBLNM, and MBLSP were 10.2, 0.5, and 63.4 seconds, respectively. Hence, MBLNM was much quicker than the KL method. However, MBLSP was slower than the other three methods due to the nonparametric multivariate density estimation. (In fact, MBLNM can be computed much faster if one has a computer with a larger memory. When using our personal computer for calculation, we did not store the kernel function calculations between the samples and their permutations in (7), and instead, we recalculated them in every iteration, due to the large memory storage requirement.)

Since there are only two components and the two symmetric modal regions are separate (see Figs. 1 and 2), we can easily make use of some parameter plots to check where the

labeling differences occurred. Figure 1 is the plot of $\sigma_1 - \sigma_2$ vs. $\mu_1 - \mu_2$. Figure 2 is the plot of $\sigma_1 - \sigma_2$ vs. π_1. For better visual results, we also added the permuted samples to the plots. The star points are the posterior modes. From Figs.1 and 2, one knows that OC and KL did not cluster the parameter points in a natural manner. The MBLNM and MBLSP methods clustered the two groups more naturally.

In this example, the labeled samples provided by the MBLNM and MBLSP methods in fact were the same as the original raw samples. Hence, it is highly likely that the label switching did not happen in the raw samples. This can also be seen by noting that if there is label switching in the raw samples, the sampled modal regions will most likely be connected together, which is not the case in Figs. 1 and 2. Note, however, since there is no label switching in the raw samples, they have not yet explored the whole posterior region, and the label switching would happen if we were to continue to run the sampler. It can be seen from Figs. 1 and 2 that the raw samples are around the posterior modes and thus are from the separate highest posterior density regions. Note that for any converged samples, we can always pick part of them (usually in highest posterior density region) such that there is no label switching among them (so they must have the same labels). Similar techniques of using part of the MCMC samples without label switching have also been used by many other researchers to get some initial estimates for labeling.

The above finding further demonstrated that MBLNM and MBLSP provided the right labels but OC and KL did not, since the raw samples themselves are the ideally labeled samples. In addition, it can be seen that the raw samples without label switching provide an ideal situation to compare different labeling methods, since only in such situation one knows the ideally labeled samples, which are the raw samples. Note also that in practice one will apply their favorite labeling method to the MCMC samples without checking whether the label switching has happened yet or not in the raw samples. (Theoretically, if the sequence is long enough, the label switching must happen in the converged raw samples.)

Using the above strategy, we also tried several other cases when the ideally labeled samples were available and found similar results to this example. That is MBLNM and MBLSP had similar labeling results, and usually had closer labeled samples to the ideally labeled samples (the raw samples in those situations) than the OC and KL methods. In addition, KL usually also had closer labeled samples to the ideally labeled samples than OC.

Example 3.2 We consider the acidity data set (Crawford et al. 1992; Crawford 1994). The data are shown in Fig. 3. The observations are the logarithms of an acidity index measured in a sample of 155 lakes in north-central Wisconsin.

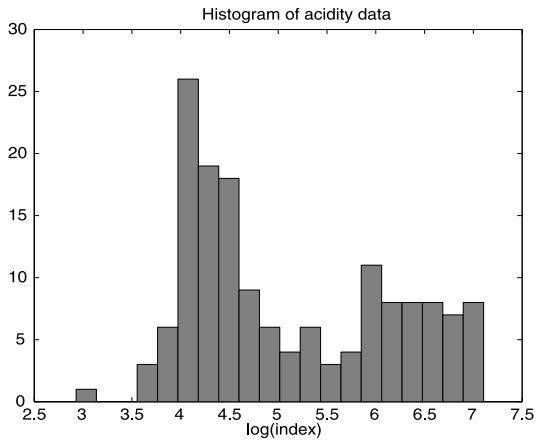

Fig. 3 Histogram of acidity data. The number of bins used is 20

This data set has been analyzed as a mixture of Gaussian distributions by Crawford et al. (1992), Crawford (1994), and Richardson and Green (1997). Based on the result of Richardson and Green (1997), the posterior for 3 components is largest. Hence, we fit this data set by a 3-component normal mixture. We post processed the 5,000 Gibbs samples by different labeling methods.

The total numbers of different labels between (OC, KL, MBLNM) and MBLSP were: 470, 325, and 53, respectively. Hence, MBLNM had much closer labeling results to MBLSP than OC and KL. The runtime for KL, MBLNM, and MBLSP were 8.9, 2.8, and 397.1 seconds, respectively. Similar to Example 3.1, one sees that MBLNM is faster than KL, but MBLSP is slower than KL.

In this example, the number of components is larger than two and the modal regions are not separated. Hence, it is difficult to use the parameter plots used in Example 3.1 to compare different labeling methods. Here, we mainly provided the trace plots and the marginal density plots of component means, shown in Figs. 4 and 5, respectively, to illustrate the success of the MBLSP method and compare it with the OC method. The KL and MBLNM methods had similar visual results as MBLSP for the above two plots. From Figs. 4 and 5, it can be seen that the MBLSP method successfully removed the label switching in the raw output of the Gibbs sampler. Based on Fig. 5, one sees that the multi-modality of the marginal posterior densities of the component means in the raw output has been removed by MBLSP, however the OC method did not remove the multi-modality very well for the second component mean μ_2.

4 Discussion

Due to the label switching, the posterior of an m-component mixture can be considered as a symmetric mixture distribution with $m!$ components. In this article, we proposed to

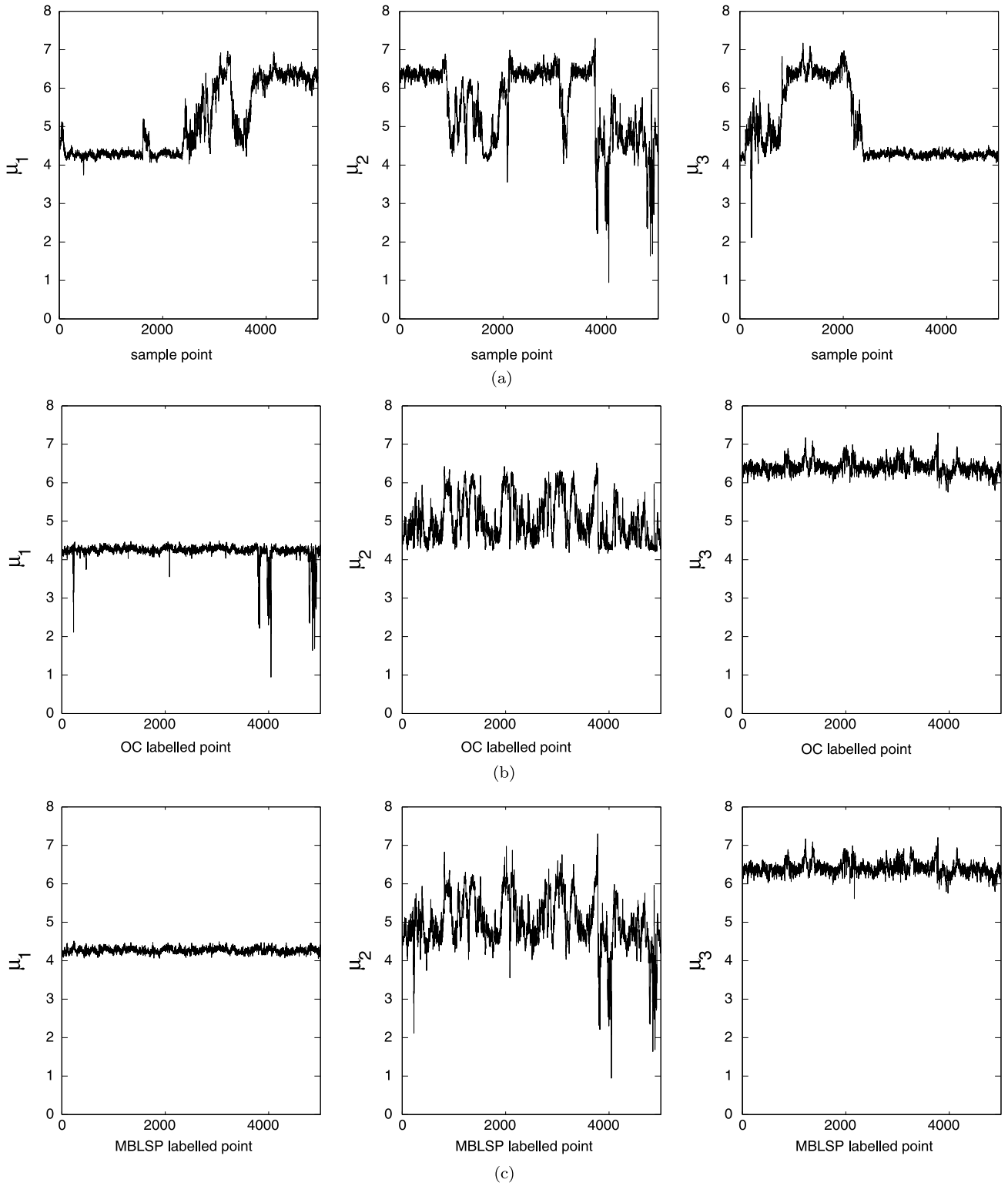

Fig. 4 Trace plots of the Gibbs samples of component means for acidity data: (**a**) original Gibbs samples; (**b**) labeled samples by OC; (**c**) labeled samples by MBLSP

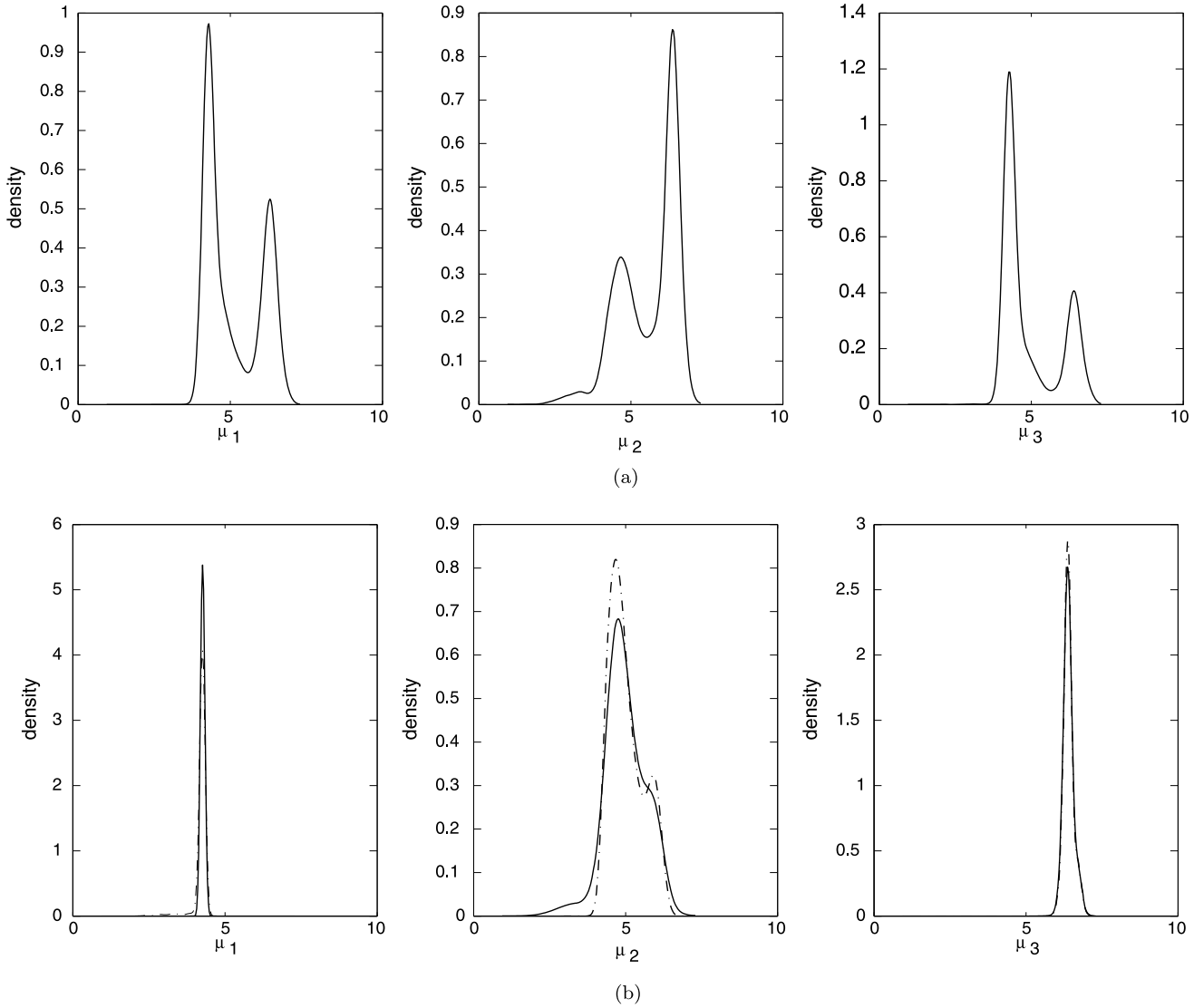

Fig. 5 Plots of estimated marginal posterior densities of component means for acidity data based on: (**a**) original Gibbs samples; (**b**) labeled samples by MBLSP (*line*) and labeled samples by OC (*dash-dot*)

solve the label switching by fitting an $m!$-component mixture model to the MCMC samples. The label for each sample can then be chosen by maximizing the corresponding labeling probabilities from the fitted mixture model. We proposed both parametric and semi-parametric mixture models to approximate the posterior. In the examples in Sect. 3, we can see that such defined labels worked quite well. In addition, it can be seen that MBLNM is much faster than MBLSP (and KL), but provides close results to MBLSP. Therefore, in practice, we prefer MBLNM method except when the sample size is too small (in such cases, the asymptotic normality might not hold).

However, as one referee pointed out that the proposed two algorithms can solve successfully the problem in ordinary cases but not when the probability of empty components existence is nonnegligible, since in such situations the generated MCMC samples from the empty components have large variations and tend to be outliers in many cases. Therefore, it will be desirable to come out some robust estimation method/algorithm, which is not sensitive to the outliers in the MCMC samples. This will be the future research.

Note that usually the posterior density has explicit form, for example when there is no hyper prior, or it can be evaluated approximately by some numerical method. Therefore, one alternative way to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (3) is to directly minimize the distance between the posterior and the symmetric normal mixture model (3). The distance can be L_2 loss or Kullback-Liebler divergence.

For the semi-parametric model, similar to Bordes et al. (2007) and Benaglia et al. (2009), the convergence property

of the Algorithm 2 has not been established and needs further research, although empirically, the Algorithm 2 worked quite well and did converge for all the data sets we tried. One might also use one-step of Algorithm 2 to speed up the MBLSP method if one starts the Algorithm 2 from some good labels (such as NORMLH labels (Yao and Lindsay 2009)). In addition, the selection of an appropriate bandwidth is another area in which further work needs to be done.

Given the labeling probabilities, one might also use them to do weighted averaging when doing Bayesian inference. Let G be the cumulative distribution function (CDF) corresponding to $g(\cdot \mid \mathbf{x})$, the labeled posterior density. Usually, the quantity of interest can be expressed as

$$T(G) = \int T(\boldsymbol{\theta}) dG(\boldsymbol{\theta}), \tag{11}$$

where $T(\boldsymbol{\theta})$ is any integrable function with respect to $G(\boldsymbol{\theta})$ and $T(\boldsymbol{\theta}^{\boldsymbol{\omega}_{(j)}}) \neq T(\boldsymbol{\theta}^{\boldsymbol{\omega}_{(k)}})$, $j \neq k$ ($T(\boldsymbol{\theta}) = \boldsymbol{\theta}$, for example). The traditional labeling methods, given the found labels $(\hat{\boldsymbol{\omega}}_1, \ldots, \hat{\boldsymbol{\omega}}_N)$, estimate G by

$$\hat{G}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^{N} I(\boldsymbol{\theta}_t^{\hat{\boldsymbol{\omega}}_t} \leq \boldsymbol{\theta})$$

and $T(G)$ by

$$T(\hat{G}) = \frac{1}{N} \sum_{t=1}^{N} T(\boldsymbol{\theta}_t^{\hat{\boldsymbol{\omega}}_t}), \tag{12}$$

where $I(\cdot)$ is the index function and $\boldsymbol{\theta}_t \leq \boldsymbol{\theta}$ is evaluated element wise. For our model based labeling method, one might also estimate G by

$$\hat{G}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{m!} \{\hat{p}_{tj} I(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}} \leq \boldsymbol{\theta})\}, \tag{13}$$

which is the distribution function that puts the point mass $\hat{p}_{tj}/N$ on $\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}}$, $t = 1, \ldots, N$, $j = 1, \ldots, m!$. Then the quantity $T(G)$ can be estimated by

$$T(\hat{G}) = \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{m!} \hat{p}_{tj} T(\boldsymbol{\theta}_t^{\boldsymbol{\omega}_{(j)}}). \tag{14}$$

Note that the weighted average (14) mimics the idea of the M step in (9).

One might expect that the weighted average (14) could reduce the bias of Bayesian inference compared to the traditional method (12), which uses the single best label for each sample. (Note that, in finite mixture model theory, estimating the missing component-indicator variables directly as unknown parameters could lead to inconsistent inference (McLachlan and Peel 2000, Sect. 2.21)). However, it is very difficult to verify this even using the simulation study. In fact, the bias or mean square errors are even undefined for the labeling problems in simulation study. For example, supposing one wants to estimate the bias or mean square errors of $T(\hat{G})$ with $T(\boldsymbol{\theta}) = \boldsymbol{\theta}$ for a certain labeling method, one needs to find $T(\hat{G})$ for many replicates. For each replicate, one uses certain labeling method to label the samples and then uses the labeled samples to estimate $T(G)$. Although $T(\hat{G})$'s are well defined for each replicate, they might have different label meaning for different replicates, because the labeled samples across different replicates need not have the same label meaning, i.e., the modal regions recovered might not be the same for different replicates. Hence, there is an issue about how to align/label the $T(\hat{G})$'s across different replicates when comparing the bias or mean square errors of $T(\hat{G})$'s for different labeling methods. Without knowing the true alignment of $T(\hat{G})$'s, some arbitrary alignment might even give misleading comparison results. This requires further research.

Acknowledgements The authors are grateful to the editor Dr. Gilles Celeux, the associate editor, and the two referees for their insightful comments and suggestions, which greatly improved this article. In addition, I am also indebted to my dissertation advisor, Bruce G. Lindsay, for his assistance and counsel in this research.

References

Benaglia, T., Chauveau, D., Hunter, D.R.: An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. J. Comput. Graph. Stat. **18**, 505–526 (2009)

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

Bordes, L., Chauveau, D., Vandekerkhove, P.: A stochastic EM algorithm for a semiparametric mixture model. Comput. Stat. Data Anal. **51**, 5429–5443 (2007)

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

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

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

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

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

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

Frühwirth-Schnatter, S.: Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. J. Am. Stat. Assoc. **96**, 194–209 (2001)

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

Geweke, J.: Interpretation and inference in mixture models: Simple MCMC works. Comput. Stat. Data Anal. **51**, 3529–3550 (2007)

Grun, B., Leisch, F.: Dealing with label switching in mixture models under genuine multimodality. J. Multivar. Anal. **100**, 851–861 (2009)

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

Jasra, A.: Bayesian inference for mixture models via Monte Carlo. Ph.D. Thesis, Imperial College London (2005)

Jasra, A., Holmes, C.C., Stephens, D.A.: Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Stat. Sci. **20**, 50–67 (2005)

Lindsay, B.G.: Mixture Models: Theory, Geometry, and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 5. Institute of Mathematical Statistics, Hayward (1995)

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

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

Papastamoulis, P., Iliopoulos, G.: An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. J. Comput. Graph. Stat. **19**, 313–331 (2010)

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

Scott, D.W.: Multivariate Density Estimation: Theory, Practice and Visualization. Wiley, New York (1992)

Sperrin, M., Jaki, T., Wit, E.: Probabilistic relabeling strategies for the label switching problem in Bayesian mixture models. Stat. Comput. **20**, 357–366 (2010)

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

Walker, A.M.: On the asymptotic behaviour of posterior distributions. J. R. Stat. Soc. B **31**, 80–88, (1969)

Yao, W., Lindsay, B.G.: Bayesian mixture labeling by highest posterior density. J. Am. Stat. Assoc. **104**, 758–767 (2009)