A Simple Solution to Bayesian Mixture Labeling

Weixin Yao

Department of Statistics, Kansas State University, Manhattan, Kansas 66506, U.S.A.

wxyao@ksu.edu

Abstract

The label switching problem is one of the fundamental problems in Bayesian mixture analysis. Using all the Markov chain Monte Carlo samples as the initials for the EM algorithm, we propose to label the samples based on the modes they converge to. Our method is based on the assumption that the samples converged to the same mode have the same labels. If a relative noninformative prior is used or the sample size is large, the posterior will be close to the likelihood and then the posterior modes can be located approximately by the EM algorithm for mixture likelihood, without assuming the availability of the closed form of the posterior. In order to speed up the computation of this labeling method, we also propose to first cluster the samples by K-means with large number of clusters K. Then, by assuming that the samples within each cluster have the same labels, we only need to find one converged mode for each cluster. Using a Monte Carlo simulation study and a real data set, we demonstrate the success of our new method in dealing with the label switching problem.

Key Words: Bayesian approach; EM algorithm; Label switching; Markov chain Monte Carlo; Mixture models

1 Introduction

The label switching problem is one of the fundamental problems in Bayesian mixture analysis. If the prior is symmetric for all components, the posterior distribution of Bayesian mixtures will be symmetric and thus invariant to all the permutations of the component parameters. Then the marginal posterior distributions for the parameters will be identical for each mixture component. Therefore the posterior means of all the components are the same and thus are poor estimates of these parameters. Similar problems will occur if we want to estimate the quantities relating to

individual components, such as predictive component densities and marginal classification probabilities. It is then meaningless to draw inference directly from Markov chain Monte Carlo (MCMC) samples using ergodic averaging before solving the label switching problem. For the illustrative examples of label switching, see Stephens (2000) and Jasra et al. (2005), among others.

To deal with the labeling problem in Bayesian analysis, many methods have been proposed. One simple way is to use an explicit parameter constraint so that only one permutation can satisfy it. See Diebolt and Robert (1994); Dellaportas et al. (1996); Richardson and Green (1997). Stephens (2000) and Celeux (1998) proposed a relabeling algorithm, 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. Yao and Lindsay (2009) proposed to label the samples based on the highest posterior region and posterior modes. Yao (2012a) established the equivalence between the labeling and clustering and proposed two clustering objective functions to label the samples. Other labeling methods include, for example, Celeux et al. (2000); Fruhwirth (2001); Hurn et al. (2003); Chung et al. (2004); Marin et al. (2005); Geweke (2007); Grun and Leisch (2009); Yao (2012b). Jasra et al. (2005) provided a good review about the existing methods to solve the label switching problem in Bayesian mixture modelling.

In this article, we propose a new modal labeling method, which labels the samples based on the *likelihood modes* they are associated with when they are used as the starting points for the *EM algorithm* of mixture models (Dempster et al., 1977). The modal labeling method assumes that if the two samples converge to the same likelihood modes based on the EM algorithm, then they have the same labels. If the prior is relative noninformative, which is the case in general, or the sample size is large, the likelihood will be close to the posterior density. Then this labeling method is almost equivalent to labeling the samples based on the posterior modes. One important nice feature of our new method is that it avoids finding the posterior modes directly. Therefore, our method doesn't depend on the specific priors used. In order to speed up our labeling process, we also propose to first cluster the samples by a method like K-means with large number of clusters K. Then, by assuming that the samples within each cluster have the same labels (this assumption will hold if K is large enough), we only need to find one converged mode for each cluster.

Compared to many existing labeling methods, such as KL algorithm (Stephens, 2000), and Yao (2010a)'s clustering algorithm, our proposed new method have the following main advantages:

1. It does not depend on any objective/loss function;

2. It is an online algorithm and is computationally much faster when the number of components is large;

3. It does not depend on the initial labels, which saves much computation time and provides more stable labeling results.

4. It is easy to implement since it only involves running a traditional EM algorithm for mixtures starting from different initial values. Therefore, we can easily use existing mixture models package in R, such as "mixtools", "FlexMix", "MCLUST", and "MIX", to implement our labeling methods.

The rest of the paper is organized as follows. Section 2 introduces our new labeling method. In Section 3, we use two simulation examples and a real data set to compare the new labeling method with two popular existing methods. We summarize our proposed labeling method in Section 4.

2 Introduction of New Labeling Method

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be independent observations from a 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 $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_m, \lambda_1, \ldots, \lambda_m)$, $f(\cdot)$ is a density function called the component density, λ_j is the component specific parameter, and π_j is the proportion of j^{th} subpopulation in the whole population with $\sum_{j=1}^{m} \pi_j = 1$. The likelihood for $\mathbf{x}$ is

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

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, by maximizing (1), is straightforward using the EM algorithm (Dempster et al., 1977). For a general introduction to mixture models, see Lindsay (1995); Bohning (1999); McLachlan and Peel (2000); Fruhwirth (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

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

A special feature of mixture model is that the likelihood function $L(\boldsymbol{\theta}^{\boldsymbol{\omega}}; \mathbf{x})$ is numerically the same as $L(\boldsymbol{\theta}; \mathbf{x})$ for any permutation $\boldsymbol{\omega}$. This is the so-called *label switching* problem.

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 $\mathbf{x} = (x_1 \cdots, x_n)$. If $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}^{\boldsymbol{\omega}})$ 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.

For simplicity of explanation of our new labeling method, let us consider the situation when $m = 2$. The following explanation can be easily extended to the situation when $m > 2$. When $m = 2$, the posterior distribution will have two symmetric modal regions, around each of the two symmetric maximal models. We can consider each of them as the "true" labeled parameter space. The aim of labeling is to recover one of the modal regions. Assuming that the prior is relative noninformative, which is the case in general, or the sample size is large, the posterior will be very close to the likelihood function. Hence the posterior modes can be found closely by the EM algorithm of the mixture models.

Ideally, suppose that there are only two symmetric modes with one in each modal region. Because of the monotone increasing property of the EM algorithm (Dempster et al., 1977), the points in one modal region will most likely converge to the same mode using the EM algorithm and the points in the other modal region will converge to the other permuted mode. Therefore, using the converged modes to label the samples, we can easily recover the modal regions. Practically, we can use one of the modes, say by order constraint labeling on some parameter, as the reference

mode. Denote by $\hat{\boldsymbol{\theta}}$ the chosen reference mode and thus $\hat{\boldsymbol{\theta}}$ has the identical label $(1, 2)$. Suppose we have N raw unlabeled MCMC samples $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ (see, for example, Diebolt and Robert (1994) and Richardson and Green (1997) for more detail about how to draw MCMC samples for mixture models), the aim of labeling is to find the labels $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$ such that $\{\boldsymbol{\theta}_1^{\boldsymbol{\omega}_1}, \ldots, \boldsymbol{\theta}_N^{\boldsymbol{\omega}_N}\}$ are in the same modal region (i.e. have the same label meaning) as the reference mode $\hat{\boldsymbol{\theta}}$. If a sample converges to $\hat{\boldsymbol{\theta}}$, then it is most likely in the same modal region as $\hat{\boldsymbol{\theta}}$ and thus has the label $(1, 2)$; if a sample converges to the permuted mode of the reference mode, which is $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$, then it is in the other permuted modal region and thus has the label $(2, 1)$.

However, if there are other minor modes besides the two permuted major modes, we simply label these minor modes by minimizing their euclidian distance to the reference major mode. Here the major mode is defined to be the one that most samples are converged to among all the posterior modes. Empirically, the major mode is usually the mode having the largest posterior.

If one wishes to use this algorithm in a way that does not require storage of all the MCMC samples (then the proposed algorithm is an on-line algorithm), one needs to find the reference major mode $\hat{\boldsymbol{\theta}}$ in advance of processing. The reference major mode can be found by the EM algorithm. We propose to run the EM algorithm from different initial values and choose the mode to which most initial values converge. Practically, the initial values can be chosen equally spaced from the burn-in samples of the MCMC sampling. Based on our experience, twenty initial values will be very reassuring to find the major mode.

Suppose the found reference mode is $\hat{\boldsymbol{\theta}}$ and it is labeled by order constraint on some parameter. The aim of labeling is to find the labels $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$ such that $\{\theta_1^{\boldsymbol{\omega}_1}, \ldots, \theta_N^{\boldsymbol{\omega}_N}\}$ have the same label meaning as $\hat{\boldsymbol{\theta}}$. Based on the above explanation, for general m, the algorithm of our proposed labeling method is as follows.

Algorithm 1. *Labeling based on EM algorithm (EMLAB)*

Step 1: Taking the MCMC sample $\{\boldsymbol{\theta}_t, t = 1, \ldots, N\}$ as the initial value, find the corresponding converged mode $\{\mathbf{m}_t, t = 1, \ldots, N\}$ using the EM algorithm of the mixture models.

Step 2: Apply to $\mathbf{m}_t$ the order constraint labeling used to define $\hat{\boldsymbol{\theta}}$, denoted by $\boldsymbol{\omega}_t^*$ (hence $\mathbf{m}_t^{\boldsymbol{\omega}_t^*}$ has

the same order constraint as $\hat{\boldsymbol{\theta}}$) and find the label $\boldsymbol{\omega}_t$ of $\boldsymbol{\theta}_t$ based on the following situations.

a) If $\mathbf{m}_t^{\boldsymbol{\omega}_t^*}$ is $\hat{\boldsymbol{\theta}}$, up to numerical error, then $\boldsymbol{\omega}_t = \boldsymbol{\omega}_t^*$.

b) If $\mathbf{m}_t^{\boldsymbol{\omega}_t^*}$ is not $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\omega}_t$ is found by a risk based criterion such as least squares:

$$\boldsymbol{\omega}_t = \arg\min_{\boldsymbol{\omega}} (\mathbf{m}_t^{\boldsymbol{\omega}} - \hat{\boldsymbol{\theta}})^T (\mathbf{m}_t^{\boldsymbol{\omega}} - \hat{\boldsymbol{\theta}}) \,. \tag{2}$$

For example, when $m = 2$ and there are only two symmetric modes, suppose the reference mode is

$$\hat{\boldsymbol{\theta}} = (\hat{\pi}_1, \hat{\pi}_2, \hat{\lambda}_{11}, \hat{\lambda}_{21}, \ldots, \hat{\lambda}_{1p}, \hat{\lambda}_{2p}),$$

where $(\lambda_{j1}, \ldots, \lambda_{jp})$ is the p-dimensional parameter vector for jth component. Suppose $\hat{\lambda}_{1k}$ and $\hat{\lambda}_{2k}$ are different for some index k and $\hat{\boldsymbol{\theta}}$ is chosen such that $\hat{\lambda}_{1k} < \hat{\lambda}_{2k}$. (If such k doesn't exist when $m > 2$, which is very unlikely in practice, we can also use a set of component parameters, say $(\lambda_{jk_1}, \ldots, \lambda_{jk_l})$, that can differentiate all the m components, which always exists since at least $(\lambda_{j1}, \ldots, \lambda_{jp})$ are different for all the m components. Then the order of components can be done by applying the "alphabetical order" to $(\lambda_{jk_1}, \ldots, \lambda_{jk_l})$ for all js, i.e., we first order the components based on λ_{jk_1} and if there is a tie, we order the components based on λ_{jk_2}, etc.)

Based on Algorithm 1, for any given sample $\boldsymbol{\theta}_t$, in order to find the label $\boldsymbol{\omega}_t$ such that $\boldsymbol{\theta}_t^{\boldsymbol{\omega}_t}$ has the same label meaning as the reference mode $\hat{\boldsymbol{\theta}}$, we first run the ascending EM algorithm using $\boldsymbol{\theta}_t$ as the initial value and denote by m_t the final converged mode after the EM algorithm converges. If there is only two symmetric modes, then m_t is either equal to the reference mode $\hat{\boldsymbol{\theta}}$ or equal to its permutation $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$. Therefore, if m_t has the same order constraint, $\lambda_{1k} < \lambda_{2k}$, as $\hat{\boldsymbol{\theta}}$, m_t must be equal to $\hat{\boldsymbol{\theta}}$ (otherwise, it is equal to $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$). Therefore, we can simply apply order constraint labeling method to m_t to find the label $\boldsymbol{\omega}_t$ such that $m_t^{\boldsymbol{\omega}_t}$ is equal to $\hat{\boldsymbol{\theta}}$. Therefore, one nice feature of EMLAB is that it does not require to compare $m!$ permutations to find $\boldsymbol{\omega}_t$ in step 2 if $\mathbf{m}_t$ is one of the major modes. In our experience, most of the samples will converge to one of the $m!$ major modes. This computation strategy makes EMLAB much faster, when m is larger, than the relabeling algorithm, such as KL algorithm (Stephens, 2000), and Yao (2010a)'s clustering algorithm, which requires $m!$

comparisons in each iteration. In addition, note that m_t is the converged mode starting from $\boldsymbol{\theta}_t$ based on the ascending EM algorithm. Therefore, graphically, we can consider $\boldsymbol{\theta}_t$ and m_t are in the same modal region and have the same labels. Therefore if m_t has the label $\boldsymbol{\omega}_t$, $\boldsymbol{\theta}_t$ also has the label $\boldsymbol{\omega}_t$, i.e., $\boldsymbol{\theta}_t^{\boldsymbol{\omega}_t}$ is in the same modal region as $\hat{\boldsymbol{\theta}}$.

The expression (2) is used to find the label $\boldsymbol{\omega}_t$ for a minor mode m_t, such that $m_t^{\boldsymbol{\omega}_t}$ is in the same modal region as the reference mode $\hat{\boldsymbol{\theta}}$. The expression (2) assumes that the distance between $m_t^{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\theta}}$ for any permutation $\boldsymbol{\omega}$ will be more likely smaller when they are in the same modal region than when they are in different modal regions. Therefore, for a minor mode m_t, we need to try all $m!$ possible permutations and find $\boldsymbol{\omega}_t$ such that $m_t^{\boldsymbol{\omega}_t}$ is closest to $\hat{\boldsymbol{\theta}}$.

Notice that the EMLAB method does not depend on the initial labels, which can save much computation time compared to the relabeling algorithm. In addition, it is an online algorithm, which does not use batch processing and thus reduces the amounts of storage.

However, note that the EMLAB method requires to run EM algorithm for each MCMC sample to find their labels. If $m = 2$ and there are only two symmetric modes, the posterior will have two symmetric modal regions around each mode. Then it is natural to think if the two points are close enough, they are most likely to be in the same modal region and thus have the same labels. Therefore, one might expect that if one sample, say $\boldsymbol{\theta}$, has label $\boldsymbol{\omega}$, then the samples in a small neighborhood around $\boldsymbol{\theta}$ will be most likely in the same modal region as $\boldsymbol{\theta}$ and thus have the same label $\boldsymbol{\omega}$. Therefore, in order to speed up the computation, we can also first cluster the samples by a method like K-means with large number of clusters K. Then, by assuming that the samples within each cluster have the same labels, we only need to run the EM algorithm once and find one converged mode for each cluster. We will denote this labeling method by EMLAB-KM.

If the likelihood is unbounded, it is possible for some points to converge to the parameter values with infinity likelihood. When this happens, we might consider running the constrained EM algorithm which puts some constraint on parameter space to make the likelihood function bounded. For example, the univariate normal mixture with unequal variance has an unbounded likelihood function. When the variance of one component goes to zero and the corresponding mean is equal to any observation, the likelihood value will go to infinity. In order to avoid this kind of unboundness,

one might run the EM algorithm over the constrained parameter space

$$\Omega_C = \{\boldsymbol{\theta} \in \Omega : \sigma_h^2/\sigma_j^2 \geq C > 0, 1 \leq h \neq j \leq m\}, \tag{3}$$

where Ω denotes the unconstrained parameter space and σ_j is the standard deviation of the j^{th} normal component. See Hathaway (1983, 1986); Bezdek, Hathaway, and Huggins (1985) for more detail. Hathaway (1985) showed that the global maximizer $\hat{\boldsymbol{\theta}}$ of likelihood over Ω_C exists and that $\hat{\boldsymbol{\theta}}$ is strongly consistent for the true value $\boldsymbol{\theta}$ provided that the true value of $\boldsymbol{\theta}$ lies in Ω_C. For multivariate normal mixture with unequal covariance matrix, $\boldsymbol{\Sigma}_i$ $(i = 1 \ldots, m)$, the likelihood function is also unbounded. Similarly to the univariate case, we can put some constraint on the covariance matrix to get the constrained global maximizer. For example, we constrain all the eigenvalues of $\boldsymbol{\Sigma}_h \boldsymbol{\Sigma}_j^{-1}$ $(1 \leq h \neq j \leq m)$ to be greater than or equal to some minimum value $C > 0$ or $|\boldsymbol{\Sigma}_h|/|\boldsymbol{\Sigma}_j| \geq C > 0$ $(1 \leq h \neq j \leq m)$.

3 Examples

In this section, we use a simulation study and a real data application to compare our proposed labeling methods EMLAB-KM and EMLAB with order constraint (OC) labeling and Stephens' KL algorithm (KL). The OC method refers to ordering on the mean parameters. For EMLAB-KM and EMLAB, we used twenty equally spaced samples from the burn-in samples as the initial values to find the reference major mode. In all of our examples, we successfully found the major modes. For EMLAB-KM, we used 100 clusters for K-means.

All the computations were done in Matlab 7.0 using a personal desktop with Intel Core 2 Quad CPU 2.40GHz. For comparison, we reported the number of different labels for each method that differed from EMLAB. It is known that the OC method is the fastest one and it took no more than several seconds in our examples. Hence, we only reported the runtime for KL, EMLAB-KM, and EMLAB. We here have used EMLAB-KM and EMLAB in batch mode so that we can determine its runtime in direct comparison with the others. Since the runtime for the KL algorithm depends on the number of starting points (i.e., the initial labels for all samples), we only reported the runtime

of KL when using the EMLAB labels as the initial labels. (*The real runtime for KL could be much longer. If one used ten different initializations for the algorithm, it might take about ten times as long* .) Using these starts also ensures that KL is as similar to EMLAB as possible.

3.1 Simulation Studies

Example 3.1: 400 data points were generated from 0.3N(0,1)+0.7N(0.5,2). Based on this data set, we generated 20,000 MCMC samples (after initial burn-in) of component means, component proportions, and the unequal component variance. The priors used for MCMC samples are

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

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 two examples.

In this example, all the samples converged to the major modes and thus can be labeled naturally by the corresponding converged modes. The runtime for KL, EMLAB-KM, and EMLAB were 63, 26, and 235 seconds, respectively. Note that EMLAB-KM was much faster than EMLAB. The total numbers of different labels between (OC, KL, EMLAB-KM) and EMLAB were: 861, 110, and 0, respectively. So, EMLAM-KM and EMLAB had the same labeling results in this example and KL had closer results than OC when compared to EMLAB.

Since there are only two components and the two symmetric modal regions are separate, we can easily make use of some parameter plots to see where the labeling differences occurred. Figure 1 is the plot of $\sigma_1 - \sigma_2$ vs. $\mu_1 - \mu_2$ and Figure 2 is the plot of $\sigma_1 - \sigma_2$ vs. π_1. For better visual results, we also add the permuted samples to the plots. From these plots, we can see that there are two separated modal regions, and that OC and KL did not accurately recover these two regions. Based on Figure 2, it appears that KL used the component proportions more heavily than the other methods. In addition, OC also mixed the two separated regions together. The EMLAB-

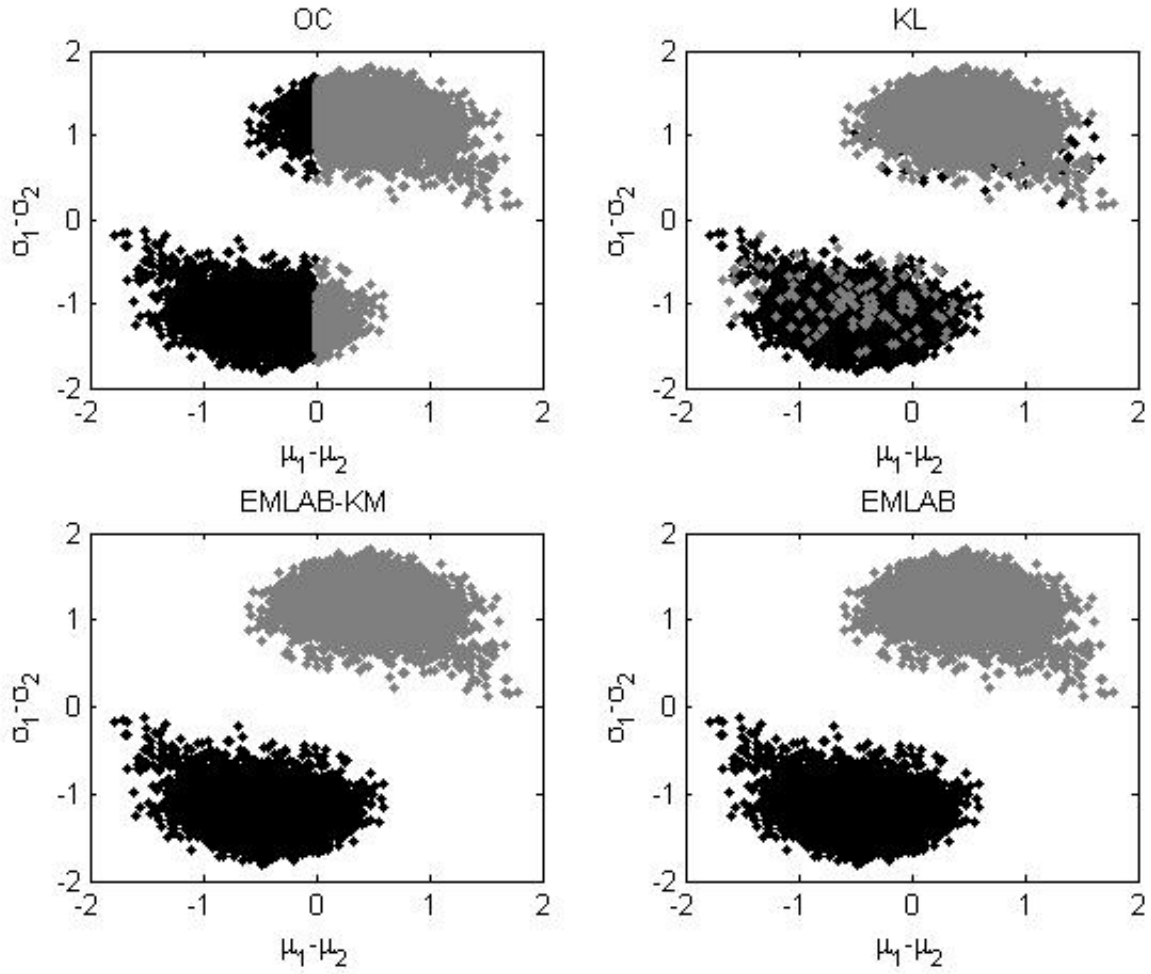


Figure 1: Plots of $\sigma_1 - \sigma_2$ vs. $\mu_1 - \mu_2$ for the four labeling methods in Example 3.1. The black points represent one set of labels and the gray points are the permuted samples.

KM/EMLAB methods clustered the two groups more naturally.

Example 3.2: We generated 400 data points from the 8-component normal mixture $\sum_{j=1}^{8} 0.125N(\mu_j, \sigma_j^2)$, where $\mu_j = 3(j-1), \sigma_j = 0.5j,\ j = 1, \ldots, 8$. Based on this data set, we generated 5,000 MCMC samples of component means, component proportions, and the unequal component variance. (Our personal computer does not have enough memory for KL algorithm when we tried to label 10,000 samples. Stephens (2000) did provide some alternative on-line versions for KL algorithm, which can reduce the storage requirements.)

In this example, 91% of samples converged to the major modes and thus can be directly labeled

by the converged modes. The other 9% of samples converged to minor modes. The runtime for KL, EMLAB-KM, and EMLAB were 5.12×10^4, 21, and 265 seconds, respectively. We can see that both EMLAB and EMLAB-KM were much faster than KL method since KL required one to compare $8! = 40320$ permutations in each iteration. From this example, one can see that if the number of components is large EMLAB and EMLAB-KM will be much faster than KL. The total numbers of different labels between (OC, KL, EMLAB-KM) and EMLAB were: 198, 812, and 144, respectively. So, in this case both OC and EMLAB-KM had much closer results than KL when compared to EMLAB. The difference between EMLAB-KM and EMLAB mainly occurred for the points converged to the degenerate modes.

It is difficult to graphically compare different labeling methods when the number of components is large. Here, we provided the trace plots and the marginal density plots to illustrate the success of EMLAB. (The OC, KL, and EMLAB-KM methods had similar visual results for those plots.) Figure 3 provides the trace plots for the original Gibbs samples and the labeled samples by EMLAB and Figure 4 provides the estimated marginal posterior density plots. From these figures, we can see that the raw samples jump around between different symmetric regions in the trace plots, which explains the multi-modalities of the marginal densities for raw samples. However, after applying EMLAB, the labeled samples have a relatively flat band in the center for the trace plots and have uni-modal densities. Therefore, EMLAB successfully removed the label switching in the raw output of the Gibbs sampler at a considerably lower computational expense than KL.

3.2 Real Data Application

Acidity Data: We consider the acidity data set (Crawford et al., 1992; Crawford, 1994). The observations are the logarithms of an acidity index measured in a sample of 155 lakes in north-central Wisconsin. More details on the data, including a histogram of its distribution can be found in Yao (2010a). Crawford et al. (1992), Crawford (1994), and Richardson and Green (1997) have used a mixture of Gaussian distributions to analyze this data set. Based on the result of Richardson and Green (1997), the posterior for 3 components was largest. Hence, we fit this data set by a 3-component normal mixture. We post processed the 20,000 Gibbs samples by the OC, KL, EMLAB,

and EMLAB-KM labeling methods.

In this example around 91% of samples converged to major modes. The runtime for KL, EMLAB-KM, and EMLAB were 45, 46, and 130 seconds, respectively. The total numbers of different labels between (OC, KL, EMLAB-KM) and EMLAB were: 445, 649, and 185, respectively. We can see that EMLAB-KM had much closer results to EMLAB than OC and KL.

Here we mainly provided the trace plots and the marginal density plots to illustrate the success of EMLAB method. (The OC, KL, and EMLAB-KM methods had similar visual results for those plots.) Figure 5 and 6 are the trace plots and the estimated marginal posterior density plots, respectively, for the original samples and the labeled samples by EMLAB. From the above two figures, we can see that the raw samples jump around between different symmetric modal regions in the trace plots and their marginal densities also have multiple modes. However, the labeled samples by EMLAB have a relatively flat band in the center for the trace plots and their marginal densities are uni-modal.

4 Summary

In this article, we proposed a new labeling method, called EMLAB, based on the converged modes when running the EM algorithm of mixtures starting from each MCMC sample. The assumption we used is that the samples converged to the same likelihood mode have the same labels. The simulation study and real data application demonstrate the success of the proposed labeling method.

We did not extensively study the effect of the choice of K on labeling results. Usually, when K increases, the EMLAB-KM method will be closer to the EMLAB. Specially, when $K = N$ (the number of MCMC samples), the EMLAB-KM will be exactly the same as EMLAB. When K is larger, the labeling results will be more accurate, but save less computation time compared with EMLAB. Hence there is a trade-off between the computation time and the accuracy of labeling when choosing K. Empirically, if K is large enough, the choice of K will not affect the labeling results too much. Based on our limited empirical studies, we recommend to choose K such that on average there are no more than 200 samples within each cluster, i.e. $K \geq N/200$, where N is the

number of MCMC samples. Similar idea can be also applied to Yao and Lindsay (2009)'s posterior modes based labeling method to speed up the computation.

References

Bezdek, J. C., Hathaway, R. M., and Huggins, V. J. (1985). Parametric estimation for normal mixtures. *Pattern Recognition*, 3, 79-84.

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 Compstat 98-*Proc. in Computational Statistics* (eds. R. Payne and P.J. Green), 227-232. Physica, Heidelberg.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Ass.*, 95, 957-970.

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

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

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

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

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

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

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

−−(2006). *Finite Mixture and Markov Switching Models*, Springer, 2006.

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

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

Hathaway, R. J. (1983). Constrained maximum likelihood estimation for a mixture of m univariate normal distribbutions. *Technical report No. 92* Columbia, South Carolina: University of Carolina.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for nomral mixture distributions. *Ann. Statist.*, 13, 795-800.

Hathaway, R. J. (1986). A constrained EM algorithm for univariate mixutres. *Journal of Statistical Computation and Simulation*, 23, 211-230.

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

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

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

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

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

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

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

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

−− (2012b). Model based labeling for mixture models. *Statistics and Computing*. 22, 337-347.

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

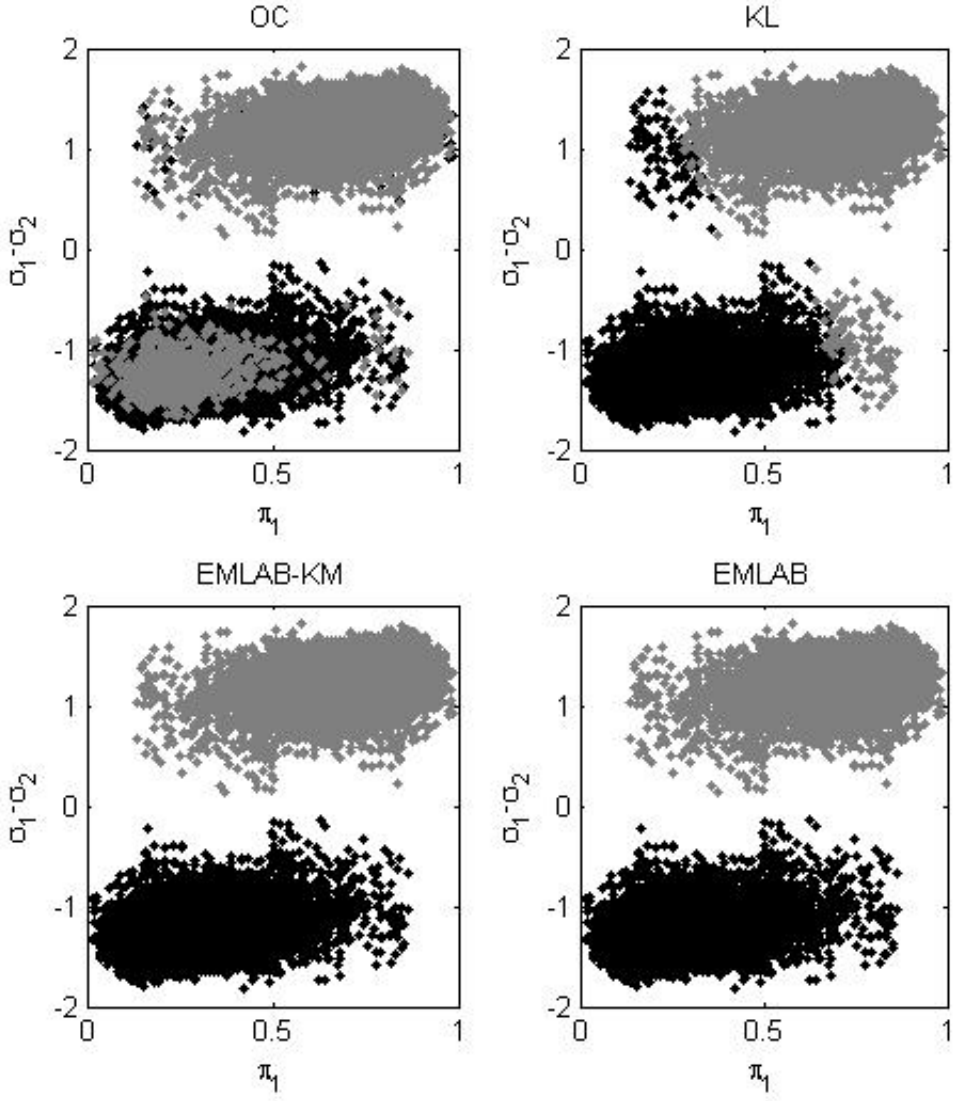
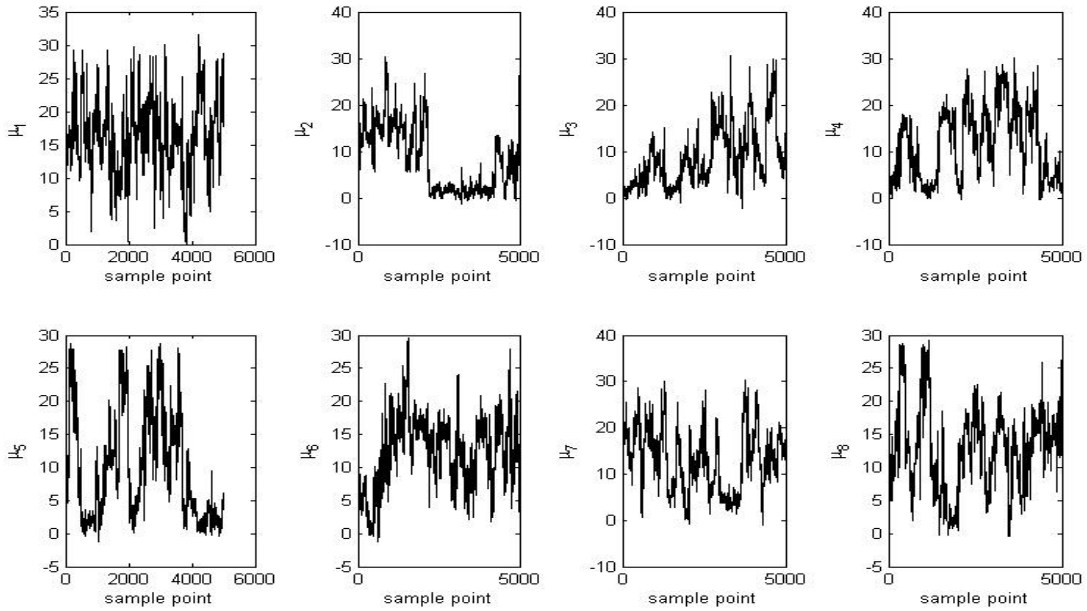
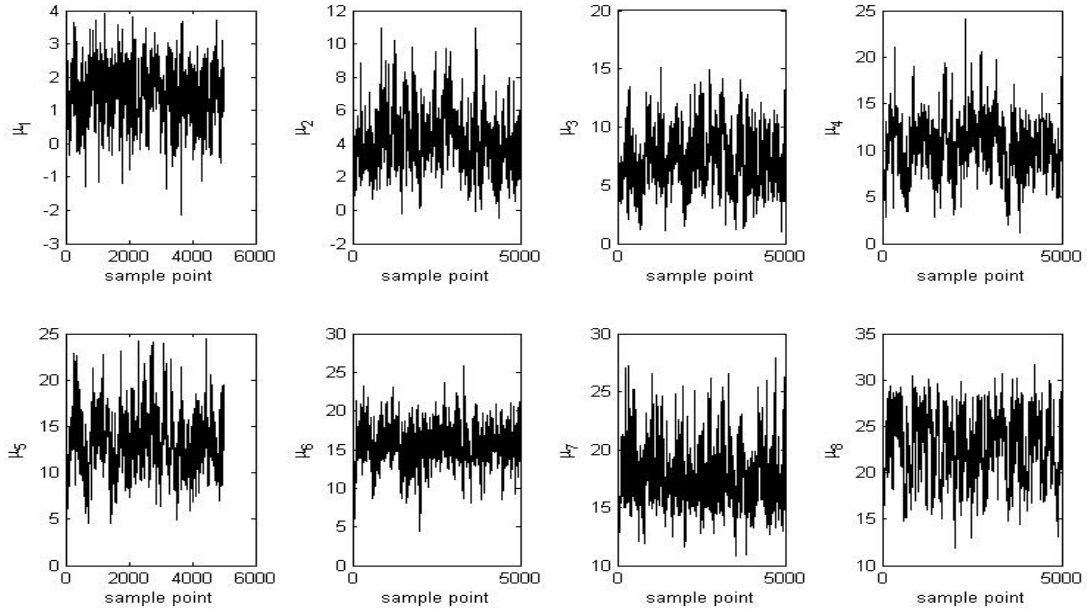


Figure 2: Plots of $\sigma_1 - \sigma_2$ vs. π_1 for the four labeling methods in Example 3.1.

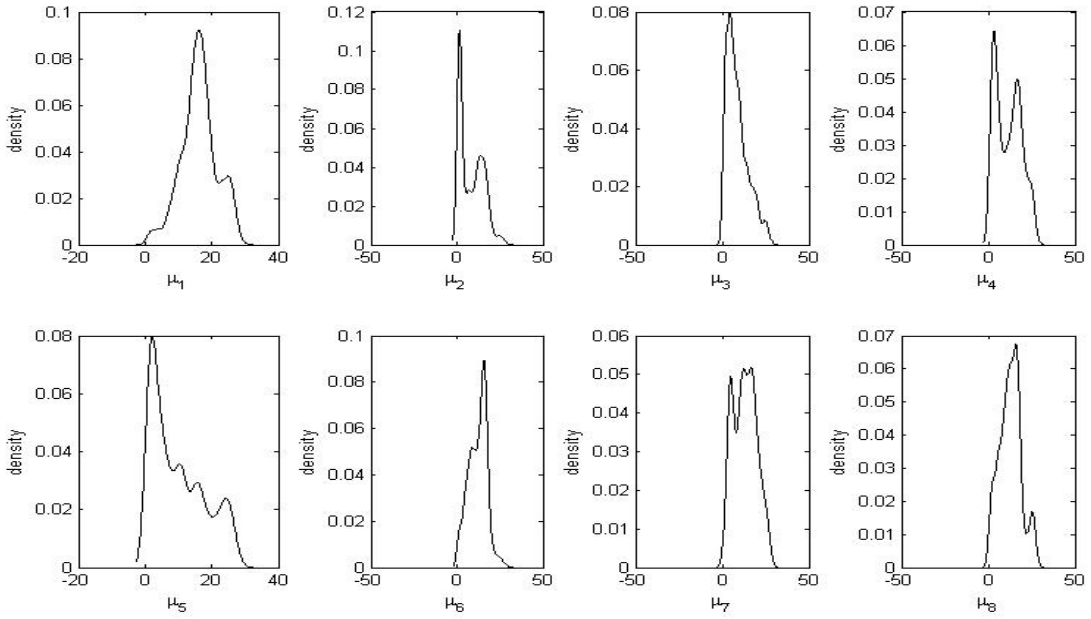


(a)

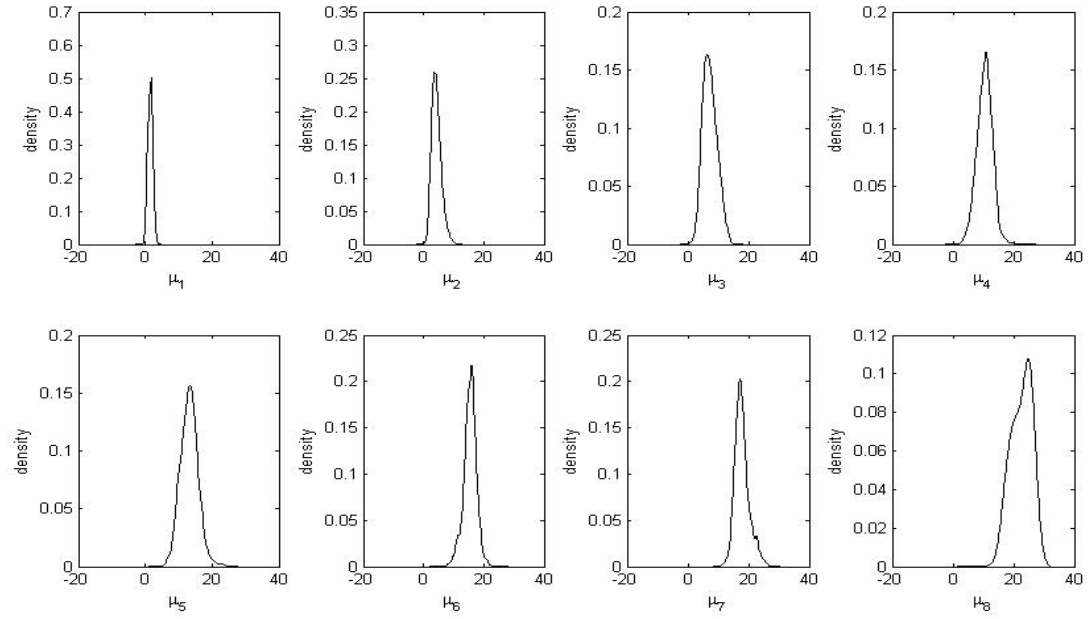


(b)

Figure 3: Trace plots of the Gibbs samples of component means for Example 3.2: (a) original Gibbs samples; (b) labeled samples by EMLAB.

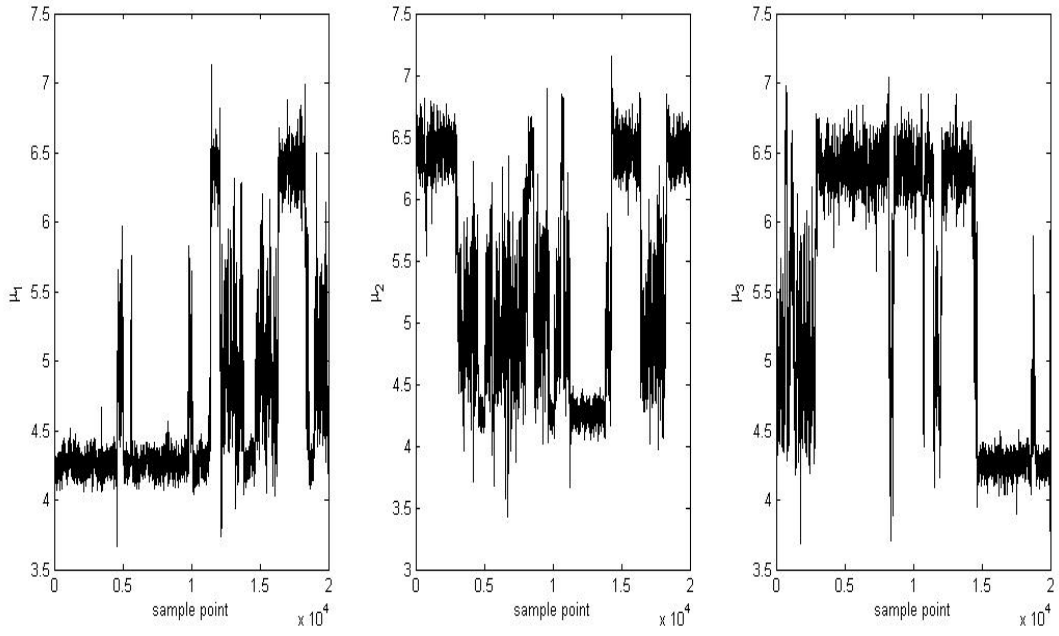


(a)

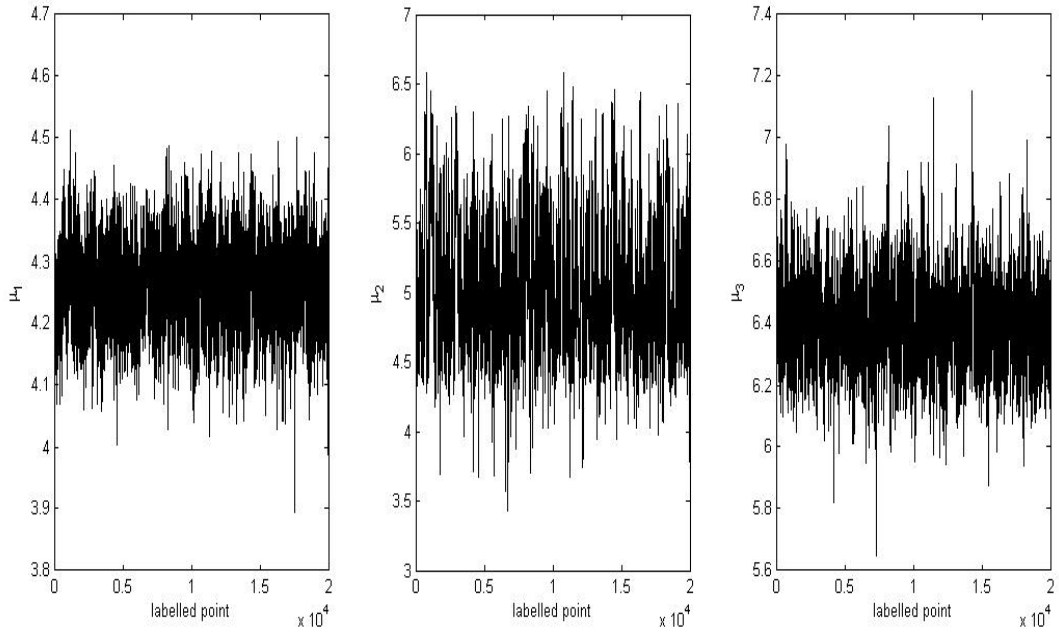


(b)

Figure 4: Plots of estimated marginal posterior densities of component means for Example 3.2 based on: (a) original Gibbs samples; (b) labeled samples by EMLAB.

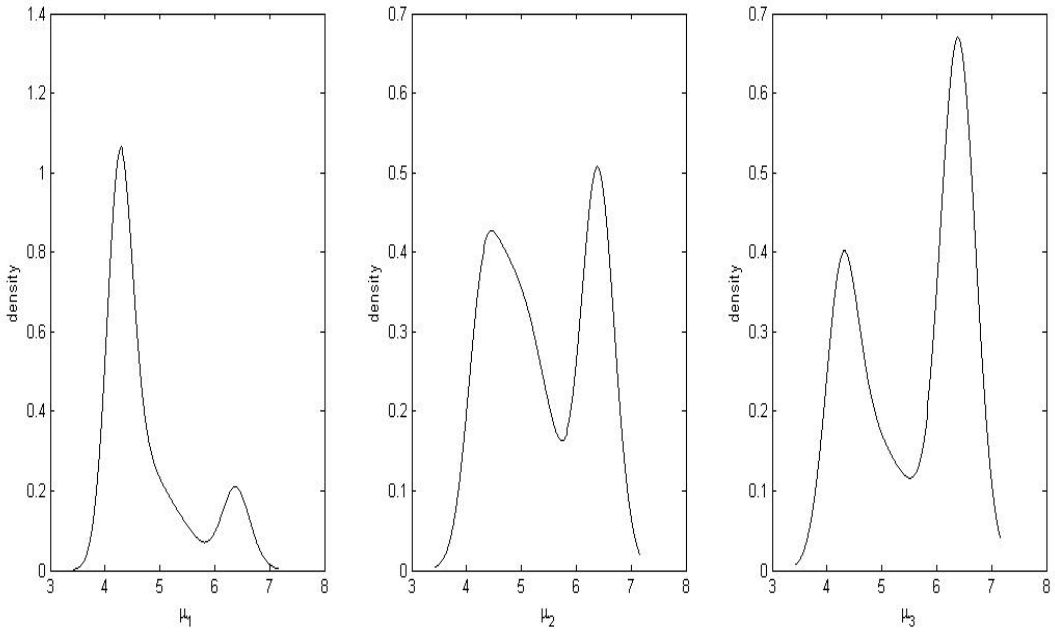


(a)

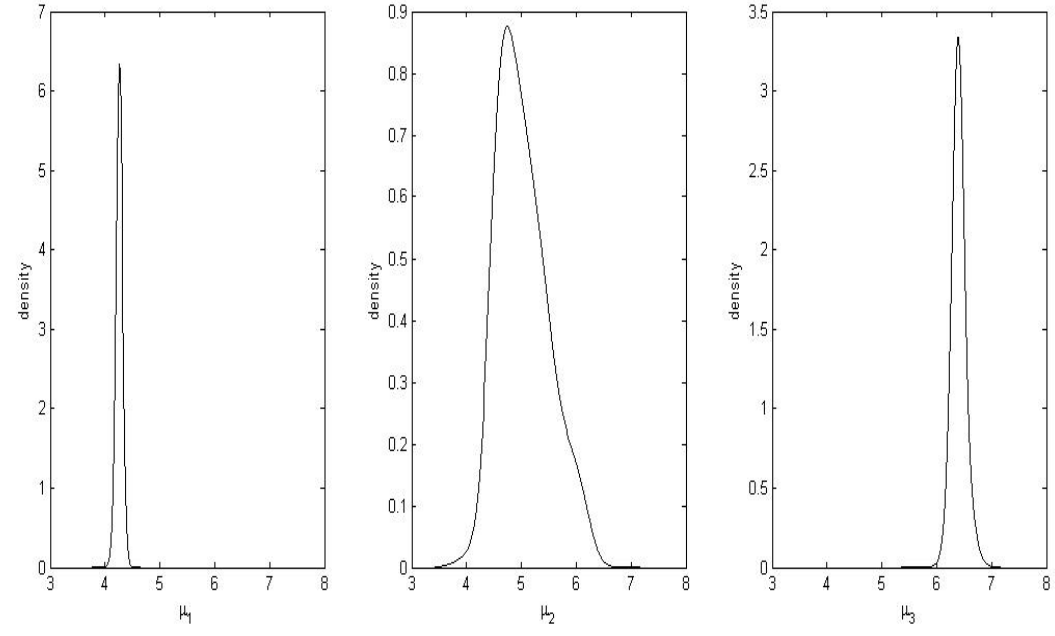


(b)

Figure 5: Trace plots of the Gibbs samples of component means for acidity data: (a) original Gibbs samples; (b) labeled samples by EMLAB.



(a)



(b)

Figure 6: Plots of estimated marginal posterior densities of component means for acidity data based on: (a) original Gibbs samples; (b) labeled samples by EMLAB.