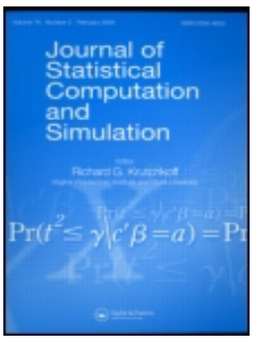

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

Label switching and its solutions for frequentist mixture models

Weixin Yao

To cite this article: Weixin Yao (2015) Label switching and its solutions for frequentist mixture models, Journal of Statistical Computation and Simulation, 85:5, 1000-1012, DOI: 10.1080/00949655.2013.859259

To link to this article: https://doi.org/10.1080/00949655.2013.859259

Journal of Statistical Computation and Simulation, 2015
Vol. 85, No. 5, 1000–1012, http://dx.doi.org/10.1080/00949655.2013.859259

Taylor & Francis
Taylor & Francis Group

Label switching and its solutions for frequentist mixture models

Weixin Yao*

Department of Statistics, Kansas State University, Manhattan, KS 66506, USA

(*Received 10 September 2013; accepted 22 October 2013*)

In this article, the label switching problem and the importance of solving it are discussed for frequentist mixture models if a simulation study is used to evaluate the performance of mixture model estimators. Two effective labelling methods are proposed by using true label for each observation. The empirical studies demonstrate that the new proposed methods work well and provide better results than the rule of thumb method of order constraint labelling. In addition, a Monte Carlo study also demonstrates that simple order constraint labelling can sometimes produce severely biased, and possibly meaningless, estimated bias and standard errors.

Keywords: complete likelihood; EM algorithm; label switching; mixture models

1. Introduction

Label switching has long been known to be a challenging problem for Bayesian mixture modelling. However, there has been much less attention on the label switching issue for frequentist mixture models. In fact, as far as we know, all of label switching papers are devoted to Bayesian mixture models. See, for example, Stephens,[1] Celeux et al.,[2] Chung et al.,[3] Geweke,[4] Yao and Lindsay,[5] Grün and Leisch,[6] Sperrin et al.,[7] Papastamoulis and Iliopoulos,[8] Cron and West,[9] and Yao.[10,11] In this article, we discuss the label switching problem for frequentist mixture models and propose its possible solutions if a simulation study is used to evaluate the performance of mixture model estimators in terms of variance and bias.

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

$$p(x; \boldsymbol{\theta}) = \pi_1 f(x; \lambda_1) + \pi_2 f(x; \lambda_2) + \cdots + \pi_m f(x; \lambda_m), \tag{1}$$

where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{m-1}, \lambda_1, \ldots, \lambda_m)$, $\pi_j > 0$ for all js, and $\sum_{j=1}^{m} \pi_j = 1$. Then, the likelihood function for $\mathbf{x}$ is

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \{\pi_1 f(x_i; \lambda_1) + \pi_2 f(x_i; \lambda_2) + \cdots + \pi_m f(x_i; \lambda_m)\}. \tag{2}$$

*Email: wxyao@ksu.edu

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

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

A special feature of the mixture model is that the likelihood function $L(\boldsymbol{\theta}^{\boldsymbol{\omega}}; \mathbf{x})$ is the same as $L(\boldsymbol{\theta}; \mathbf{x})$ for any permutation $\boldsymbol{\omega}$, i.e. $L(\boldsymbol{\theta}; \mathbf{x})$ is invariant to the permutation of component label of $\boldsymbol{\theta}$. Hence, if $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (MLE), $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$ is the MLE for any permutation $\boldsymbol{\omega}$. This is the so-called *label switching* for mixture models.

It is well known that label switching results in much difficulty for the Bayesian mixture analysis. In this article, we focus on the label switching problem for frequentist mixture models if a simulation study is used to evaluate the variation and the biases of the MLEs of component specific parameters. For example, for a univariate two-component normal mixture with parameter $\boldsymbol{\theta} = (\pi_1, \pi_2, \mu_1, \mu_2)$, let us assume that the jth raw simulated parameter estimate is $\hat{\boldsymbol{\theta}}_j = (\hat{\pi}_{j1}, \hat{\pi}_{j2}, \hat{\mu}_{j1}, \hat{\mu}_{j2})$, $j = 1, \ldots, N$. The label in $(\hat{\pi}_{j1}, \hat{\mu}_{j1})$ is meaningless and $(\hat{\pi}_{j1}, \hat{\mu}_{j1})$ can be the estimate of either (π_1, μ_1) or (π_2, μ_2), due to the invariance of the likelihood to the component labels. The true labels of $(\hat{\pi}_{j1}, \hat{\mu}_{j1})$s might be different for different js. Therefore, one cannot use $\{(\hat{\pi}_{j1}, \hat{\mu}_{j1}),\ j = 1, \ldots, N\}$ directly to estimate the bias and the variation for the MLE of (π_1, μ_1).

Given a sequence of raw unlabelled simulated estimates $(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_N)$ of $\boldsymbol{\theta}$, in order to measure their bias and variation, one must first label these samples, i.e. find the labels $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$ such that $(\hat{\boldsymbol{\theta}}_1^{\boldsymbol{\omega}_1}, \ldots, \hat{\boldsymbol{\theta}}_N^{\boldsymbol{\omega}_N})$ have the same label meaning. Then one can use the labelled estimates to estimate the variation and the bias. Without 'correct' labels, the estimates tend to have serious bias and the estimated variation might also be misleading.

The rule of thumb to solve the label switching is to simply put an explicit parameter constraint on all of the estimates such that only one permutation can satisfy the parameter constraint for each estimate (for example, set $\hat{\mu}_1 < \hat{\mu}_2$). However, Celeux,[12] Celeux et al.,[2] and Stephens [1] have all expressed their concerns about imposing an identifiability constraint for the Bayesian mixture model analysis. One main problem with the order constraint labelling is that it can only use the label information of one component parameter at a time, which is not desirable when many component parameters could simultaneously provide label information. Another main problem with identifiability constraint labelling is the choice of constraint, especially for multivariate problems. Different order constraints might generate significantly different results (for example, the constraint put on component means and the constraint put on component proportions might give different results). Therefore, it is difficult to anticipate the overall effect.

In this article, it is demonstrated that the simple order constraint labelling leads to similar undesirable results in the frequentist mixture models as in the Bayesian mixture models. Additionally, two effective labelling methods are proposed for frequentist mixture models if a simulation study is used to evaluate the variation and the biases of the MLEs of component specific parameters. For a simulation study, the component label for each observation is usually generated first and then the observations are generated from the corresponding components. Although the true component labels of the observations are not used to find the MLE of the mixture model parameters, they provide valuable information for the labels of the MLEs obtained. Based on the true component label information for each observation, two labelling methods are proposed. The first method is to label the parameter estimates by maximizing the complete likelihood. The second method is to label the parameter estimates by minimizing the Euclidean distance between the classification probabilities and the latent true labels. The empirical studies demonstrate that the newly proposed methods work well and provide better labelling results than the rule of thumb method of order constraint labelling.

Note that label switching creates issues if one focuses on the inference for the component specific parameters, such as component parameters, component densities, or classification

probabilities. However, label switching might not be an issue for some mixture problems. For example, if one uses a simulation study to evaluate the accuracy of the estimated mixture density, then label switching does not create issues, because the mixture density does not depend on the labels.

The paper is organized as follows. In Section 2, the new labelling methods are introduced. In Section 3, a simulation study is used to compare the proposed labelling methods to the traditional order constraint labelling. In Section 4, further discussions are provided on how to apply the proposed labelling methods to bootstrapped parameter estimates.

2. New labelling methods

For a simulation study, in order to generate the sample $(x_1, \ldots, x_n)$ from the model (1), one can first generate the multinomial random variable J_i, which is called the latent component label, with $P(J_i = j) = \pi_j,\ j = 1, 2, \ldots, m$, and next generate x_i from the density $f(x; \lambda_{J_i})$. The generated random variable x_i will have the mixture density (1). Based on $(x_1, \ldots, x_n)$, one can then find the MLE of $\boldsymbol{\theta}$ without using the latent labels. By repeating the above procedures, one can get a sequence of the simulated estimates, say $(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_N)$.

Note that the labels in the raw estimates $\hat{\boldsymbol{\theta}}_j$s are meaningless. However, one would probably assume that the latent labels, used to generate the sample, should carry some label information of the MLE. In this article, two new labelling methods are proposed by using the latent labels J_is.

Complete likelihood-based labelling (COMPLH): Let $\mathbf{z} = \{z_{ij},\ i = 1, \ldots, n, j = 1, \ldots, m\}$, where

$$z_{ij} = \begin{cases} 1 & \text{if } J_i = j, \\ 0 & \text{otherwise.} \end{cases}$$

The first proposed method is to find the label $\boldsymbol{\omega}$ for $\hat{\boldsymbol{\theta}}$ by maximizing the complete likelihood of $(\mathbf{x}, \mathbf{z})$ over $\boldsymbol{\omega}$

$$L(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}; \mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \{\hat{\pi}_j^{\boldsymbol{\omega}} f(x_i; \hat{\lambda}_j^{\boldsymbol{\omega}})\}^{z_{ij}}, \tag{3}$$

where $\hat{\pi}_j^{\boldsymbol{\omega}} = \hat{\pi}_{\boldsymbol{\omega}(j)}$ and $\hat{\lambda}_j^{\boldsymbol{\omega}} = \hat{\lambda}_{\boldsymbol{\omega}(j)}$. Unlike the mixture likelihood, the complete likelihood $L(\hat{\boldsymbol{\theta}}; \mathbf{x}, \mathbf{z})$ is not invariant to the permutation of component labels, since the variable $\mathbf{z}$ carries the label information. Therefore, $L(\hat{\boldsymbol{\theta}}; \mathbf{x}, \mathbf{z})$ carries the label information of parameter $\hat{\boldsymbol{\theta}}$ and can be used to do labelling. Here, the information of latent variable $\mathbf{z}$ is used to break the permutation symmetry of mixture likelihood.

Note that

$$\begin{aligned} \log\{L(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}; \mathbf{x}, \mathbf{z})\} &= \sum_{i=1}^{n} \sum_{j=1}^{m} \left[z_{ij} \log \left\{ \frac{\hat{\pi}_j^{\boldsymbol{\omega}} f(x_i; \hat{\lambda}_j^{\boldsymbol{\omega}})}{p(x_i; \hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}})} \right\} \right] + \sum_{i=1}^{n} \sum_{j=1}^{m} [z_{ij} \log\{p(x_i; \hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}})\}] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}) + \sum_{i=1}^{n} \log p(x_i; \hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}) \end{aligned} \tag{4}$$

where

$$p_{ij}(\boldsymbol{\theta}^{\boldsymbol{\omega}}) = \frac{\pi_j^{\boldsymbol{\omega}} f(x_i; \lambda_j^{\boldsymbol{\omega}})}{p(x_i; \boldsymbol{\theta}^{\boldsymbol{\omega}})}$$

and $p(x_i; \boldsymbol{\theta}^{\boldsymbol{\omega}}) = \sum_{j=1}^{m} \pi_j^{\boldsymbol{\omega}} f(x_i; \lambda_j^{\boldsymbol{\omega}})$. Note that the second term of Equation (4) is log mixture likelihood and thus is invariant to the permutation of component labels of $\hat{\boldsymbol{\theta}}$.

THEOREM 2.1 *Maximizing* $L(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}; \boldsymbol{x}, \boldsymbol{z})$ *with respect to* $\boldsymbol{\omega}$ *in Equation* (3) *is equivalent to maximizing*

$$\ell_1(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}; \boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \log p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}). \tag{5}$$

Note that Theorem 2.1 is equivalent to minimizing the Kullback–Leibler divergence if z_{ij} is considered the true classification probability and $p_{ij}(\hat{\boldsymbol{\theta}})$ is considered the estimated classification for x_i. In practice, it is usually easier to work on Equation (5) than Equation (3), since the classification probabilities $p_{ij}(\boldsymbol{\theta})$ is a byproduct of an expectation–maximization (EM) algorithm. In addition, note that $p_{ij}(\boldsymbol{\theta}^{\boldsymbol{\omega}}) = p_{i\boldsymbol{\omega}(j)}(\boldsymbol{\theta})$. Therefore, the computation of the complete likelihood labelling is usually very fast.

Distance-based labelling (DISTLAT): The second method is to do labelling by minimizing the distance between the classification probabilities and the true latent labels over $\boldsymbol{\omega}$

$$\ell_2(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \{p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}) - z_{ij}\}^2. \tag{6}$$

Therefore, distance-based labelling (DISTLAT) tries to find the labels such that the estimated classification probabilities are as similar to latent labels as possible based on Euclidian distance.

Note that

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \{p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}) - z_{ij}\}^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} [\{p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}})\}^2 + \{z_{ij}\}^2] - 2\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}).$$

The first part of above formula is invariant to the label. Thus, minimizing the above Euclidian distance is equivalent to maximizing the second part.

THEOREM 2.2 *Minimizing* $\ell_2(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}; \boldsymbol{x}, \boldsymbol{z})$ *with respect to* $\boldsymbol{\omega}$ *in Equation* (6) *is equivalent to maximizing*

$$\ell_3(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}). \tag{7}$$

Note that the objective functions (5) and (7) are very similar, except that Equation (5) uses log transformation for $p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}})$ but Equation (7) does not. One could also find the labels by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \rho(z_{ij}, p_{ij}(\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}})), \tag{8}$$

where ρ is a loss function. However, it requires more research to find the optimal loss function ρ and to investigate how sensitive the choice of ρ is to the labelling results. The choices of $\rho(x, y) = x \log y$ and $\rho(x, y) = xy$ used in Theorems 2.1 and 2.2, respectively, have some nice interpretations. For example, the choice of $\rho(x, y) = x \log y$ has the complete likelihood and Kullback–Leibler distance interpretations. The simulation studies in Section 3 demonstrate that these two choices give similar labelling results. It is expected that any reasonable choice of loss function ρ should give similar relabelling results.

Similar to the relabelling algorithm proposed by Stephens,[1] the optimization of Equation (8) (and thus Equations (5) and (7)) can be considered as a special version of the assignment problem [13, p.195] and its computation can be done by the Hungarian algorithm.[14] Since the computational cost of the Hungarian algorithm scales cubicly with the number of mixture components, the computation of Equation (8) requires a lot of time or parallel computing, when the number of components is large, for example, larger than 100.

3. Simulation study

In this section, a simulation study is used to compare the proposed complete likelihood-based labelling method (COMPLH) and the Euclidean distance-based labelling method (DISTLAT) with rule of thumb method of order constraint labelling and the normal likelihood-based labelling method (NORMLH) [5] for both univariate and multivariate normal mixture models. Yao and Lindsay [5] proposed to use normal likelihood criterion to do labelling for Bayesian mixtures based on asymptotic normality assumption of posterior distribution. Based on the asymptotic normality of the 'corrected' labelled MLE,[15,16] one can similarly do labelling for estimates $(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_N)$ by minimizing the following negative log normal likelihood over $(\bar{\boldsymbol{\theta}}, \boldsymbol{\Omega}, \boldsymbol{\Sigma})$,

$$\ell_4(\bar{\boldsymbol{\theta}}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) = N \log(|\boldsymbol{\Sigma}|) + \sum_{t=1}^{N} (\hat{\boldsymbol{\theta}}_t^{\boldsymbol{\omega}_t} - \bar{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\theta}}_t^{\boldsymbol{\omega}_t} - \bar{\boldsymbol{\theta}}), \tag{9}$$

where $\bar{\boldsymbol{\theta}}$ is the centre value for the normal distribution, $\boldsymbol{\Sigma}$ is the covariance structure, and $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$. The minimization of Equation (9) can be done by repeating the following two steps:

Step 1: Update $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}$ by minimizing Equation (9)

$$\bar{\boldsymbol{\theta}} = \frac{1}{N} \sum_{t=1}^{N} \hat{\boldsymbol{\theta}}_t^{\boldsymbol{\omega}_t},$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{t=1}^{N} (\hat{\boldsymbol{\theta}}_t^{\boldsymbol{\omega}_t} - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_t^{\boldsymbol{\omega}_t} - \bar{\boldsymbol{\theta}})^{\mathrm{T}}.$$

Step 2: For $t = 1, \ldots, N$, choose $\boldsymbol{\omega}_t$ by

$$\boldsymbol{\omega}_t = \arg\min_{\boldsymbol{\omega}} (\hat{\boldsymbol{\theta}}_t^{\boldsymbol{\omega}} - \bar{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\theta}}_t^{\boldsymbol{\omega}} - \bar{\boldsymbol{\theta}}).$$

It is well known that normal mixture models with unequal variance have unbounded likelihood function. Therefore, the MLE is not well defined. Similar unboundness issue also exists for multivariate normal mixture models with unequal covariance. Considerable research has been performed on the unbounded mixture likelihood issue. See, for example, Hathaway,[17,18] Chen et al.,[19] Chen and Tan,[20] and Yao.[21] Lindsey [22] also recommended using the Fisher's [23] original likelihood definition to avoid the unboundedness of the mixture likelihood.

Note that the purpose in this article is not on parameter estimation but on labelling the parameter estimates after they are found. For the simplicity of computations of MLE and explanations, it is assumed that the component variance is equal to avoid the unboundedness of mixture likelihood. However, the proposed labelling methods do apply to cases of unequal variances. The EM algorithm is run based on 20 randomly chosen initial values and stops until the maximum difference between the updated parameter estimates of two consecutive iterations is less than 10^{-5}.

To compare different labelling results, the estimated biases and standard errors of labelled MLEs from different labelling methods are reported. It is expected that the 'ideally' labelled estimates should have small bias. Therefore, the bias is a good indicator of how well each labelling method performs. Note that the estimated standard errors or mean squared error (MSE) by each labelling method cannot be used easily to compare different labelling methods. One of the reasons is that no labelling methods can have unanimous smaller simulated standard errors or MSE for all parameters based on the empirical studies. For example, the order constraint labelling methods usually have smaller estimated standard errors for the used constrained parameters but larger estimated standard errors for other parameters. More research is required to find other criteria to compare different labelling methods. However, it is expected that this will not be an easy task since the true labels of parameter estimates are unknown, even in a simulation study.

Example 1 Consider the mixture model $\pi_1 N(\mu_1, 1) + (1 - \pi_1)N(\mu_2, 1)$, where $\pi_1 = 0.3$ and $\mu_1 = 0$. The following four cases are considered for μ_2: (I) $\mu_2 = 0.5$; (II) $\mu_2 = 1$; III) $\mu_2 = 1.5$; (IV) $\mu_2 = 2$. The above four cases have equal component variance, unequal component proportions, and the separation of two mixture components increases from Cases I to IV. For each case, 500 replicates are run for both sample size 50 and 200 and the MLE is found for each replicate by EM algorithm assuming equal variance. Five labelling methods are considered: order constraint labelling based on component means (OC-μ), order constraint labelling based on component proportions (OC-π), normal likelihood-based labelling (NORMLH), complete likelihood-based labelling (COMPLH), and the Euclidean distance-based labelling method (DISTLAT).

Tables 1 and 2 report the estimated biases and standard deviations (Stds) of the MLEs based on different labelling methods for $n = 50$ and 200, respectively. Since equal variance is assumed, the variance estimates do not carry the label information and are the same under different labelling methods. Therefore, the variance estimates are not reported in the tables for the simplicity of comparison. From Tables 1 and 2, it can be seen that COMPLH and DISTLAT have similar results and both have smaller estimated biases than any other labelling method, especially when two components are close, and NORMLH also has slightly smaller estimated biases than OC-μ and OC-π, which have large estimated biases. In addition, as sample size increases, all labelling methods perform better and have smaller estimated biases and standard errors. Therefore, when sample size is large enough, it is expected that the five labelling methods will provide similar labelled estimates for all four cases considered in this example.

Table 1. Bias (Std) of point estimates when $n = 50$ for Example 1.

Case	TRUE	OC-μ	OC-π	NORMLH	COMPLH	DISTLAT
I	$\mu_1 : 0$	−0.547 (0.705)	0.250 (1.456)	−0.246 (1.240)	−0.040 (1.356)	−0.017 (1.371)
	$\mu_2 : 0.5$	0.652 (0.640)	−0.145 (0.476)	0.351 (0.457)	0.145 (0.529)	0.122 (0.518)
	$\pi_1 : 0.3$	0.171 (0.284)	−0.043 (0.149)	0.053 (0.245)	0.004 (0.207)	−0.002 (0.201)
II	$\mu_1 : 0$	−0.292 (0.722)	0.479 (1.497)	−0.030 (1.22)	0.075 (1.309)	0.121 (1.348)
	$\mu_2 : 1$	0.518 (0.617)	−0.252 (0.516)	0.257 (0.470)	0.152 (0.499)	0.106 (0.491)
	$\pi_1 : 0.3$	0.162 (0.277)	−0.034 (0.151)	0.059 (0.241)	0.028 (0.220)	0.012 (0.207)
III	$\mu_1 : 0$	−0.131 (0.711)	0.441 (1.436)	0.026 (1.110)	0.031 (1.112)	0.071 (1.167)
	$\mu_1 : 1.5$	0.300 (0.575)	−0.271 (0.576)	0.144 (0.412)	0.138 (0.415)	0.098 (0.416)
	$\pi_1 : 0.3$	0.107 (0.238)	−0.011 (0.144)	0.051 (0.208)	0.051 (0.208)	0.036 (0.196)
IV	$\mu_1 : 0$	−0.083 (0.709)	0.288 (1.322)	−0.030 (0.901)	−0.013 (0.940)	0.002 (0.975)
	$\mu_2 : 2$	0.152 (0.461)	−0.218 (0.594)	0.100 (0.366)	0.083 (0.366)	0.067 (0.365)
	$\pi_1 : 0.3$	0.049 (0.191)	−0.006 (0.129)	0.033 (0.177)	0.026 (0.170)	0.020 (0.163)

Example 2 Consider the multivariate normal mixture model

$$\pi_1 N\left(\begin{pmatrix}\mu_{11}\\ \mu_{12}\end{pmatrix}, \quad \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right) + (1-\pi_1) N\left(\begin{pmatrix}\mu_{21}\\ \mu_{22}\end{pmatrix}, \quad \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right),$$

where $\mu_{11} = 0$ and $\mu_{12} = 0$. The following four cases are considered for π_1, μ_{21}, and μ_{22}:

I. $\pi_1 = 0.3, \mu_{21} = 0.5, \mu_{22} = 0.5$;
II. $\pi_1 = 0.3, \mu_{21} = 1, \mu_{22} = 1$;
III. $\pi_1 = 0.3, \mu_{21} = 1.5, \mu_{22} = 1.5$;
IV. $\pi_1 = 0.5, \mu_{21} = 3, \mu_{22} = 0$.

For each case, 500 replicates are run for both sample sizes 100 and 400 and the MLE is found for each replicate by EM algorithm assuming equal covariance. The following six labelling methods

Table 2. Bias (Std) of point estimates when $n = 200$ for Example 1.

Case	TRUE	OC-μ	OC-π	NORMLH	COMPLH	DISTLAT
I	$\mu_1 : 0$	−0.479 (0.718)	0.122 (1.345)	−0.479 (0.718)	−0.065 (1.282)	−0.047 (1.29)
	$\mu_2 : 0.5$	0.503 (0.613)	−0.099 (0.381)	0.220 (0.366)	0.088 (0.367)	0.071 (0.369)
	$\pi_1 : 0.3$	0.151 (0.293)	−0.049 (0.162)	0.018 (0.235)	−0.027 (0.192)	−0.031 (0.187)
II	$\mu_1 : 0$	−0.212 (0.728)	0.342 (1.332)	−0.004 (1.130)	0.050 (1.178)	0.077 (1.199)
	$\mu_2 : 1$	0.341 (0.559)	−0.214 (0.424)	0.133 (0.353)	0.078 (0.366)	0.052 (0.371)
	$\pi_1 : 0.3$	0.132 (0.271)	−0.030 (0.159)	0.039 (0.229)	0.017 (0.211)	0.007 (0.202)
III	$\mu_1 : 0$	−0.132 (0.621)	0.113 (1.029)	−0.076 (0.792)	−0.073 (0.797)	−0.073 (0.797)
	$\mu_1 : 1.5$	0.096 (0.358)	−0.149 (0.429)	0.040 (0.273)	0.037 (0.275)	0.037 (0.275)
	$\pi_1 : 0.3$	0.032 (0.190)	−0.013 (0.139)	0.007 (0.165)	0.006 (0.163)	0.006 (0.163)
IV	$\mu_1 : 0$	−0.028 (0.367)	0.017 (0.532)	−0.021 (0.418)	−0.019 (0.428)	−0.019 (0.428)
	$\mu_1 : 2$	0.009 (0.220)	−0.036 (0.277)	0.002 (0.187)	0.000 (0.189)	0.000 (0.189)
	$\pi_1 : 0.3$	0.005 (0.103)	0.000 (0.092)	0.003 (0.099)	0.002 (0.097)	0.002 (0.097)

Table 3. Bias (Std) of point estimates when $n = 100$ for Example 2.

Case	TRUE	OC-μ_1	OC-μ_2	OC-π	NORMLH	COMPLH	DISTLAT
I	$\mu_{11} : 0$	−0.246 (0.633)	0.306 (0.819)	0.308 (1.136)	0.252 (1.130)	0.145 (1.091)	0.164 (1.099)
	$\mu_{12} : 0$	0.324 (0.92)	−0.242 (0.604)	0.395 (1.175)	0.408 (1.180)	0.241 (1.153)	0.244 (1.155)
	$\mu_{21} : 0.5$	0.414 (0.593)	−0.137 (0.868)	−0.140 (0.366)	−0.083 (0.358)	0.024 (0.404)	0.004 (0.399)
	$\mu_{22} : 0.5$	−0.090 (0.816)	0.476 (0.639)	−0.162 (0.363)	−0.175 (0.359)	−0.007 (0.392)	0.010 (0.388)
	$\pi_1 : 0.3$	0.181 (0.278)	0.212 (0.278)	−0.031 (0.155)	−0.027 (0.160)	0.009 (0.202)	0.003 (0.196)
II	$\mu_{11} : 0$	−0.017 (0.598)	0.444 (0.937)	0.435 (1.101)	0.281 (1.040)	0.258 (1.019)	0.264 (1.024)
	$\mu_{12} : 0$	0.435 (0.916)	−0.046 (0.631)	0.392 (1.123)	0.272 (1.070)	0.213 (1.027)	0.222 (1.036)
	$\mu_{21} : 1$	0.238 (0.536)	−0.225 (0.708)	−0.216 (0.400)	−0.061 (0.367)	−0.038 (0.383)	−0.044 (0.382)
	$\mu_{22} : 1$	−0.250 (0.782)	0.232 (0.534)	−0.207 (0.399)	−0.087 (0.378)	−0.028 (0.396)	−0.037 (0.390)
	$\pi_1 : 0.3$	0.141 (0.258)	0.131 (0.256)	−0.020 (0.148)	0.000 (0.174)	0.022 (0.196)	0.017 (0.191)
III	$\mu_{11} : 0$	0.033 (0.563)	0.23 (0.841)	0.233 (0.902)	0.150 (0.814)	0.136 (0.789)	0.138 (0.797)
	$\mu_{12} : 0$	0.238 (0.823)	0.036 (0.540)	0.215 (0.859)	0.107 (0.727)	0.115 (0.736)	0.120 (0.743)
	$\mu_{21} : 1.5$	0.051 (0.354)	−0.145 (0.505)	−0.148 (0.391)	−0.065 (0.329)	−0.052 (0.331)	−0.053 (0.319)
	$\mu_{22} : 1.5$	−0.166 (0.507)	0.036 (0.344)	−0.143 (0.380)	−0.035 (0.293)	−0.042 (0.303)	−0.047 (0.308)
	$\pi_1 : 0.3$	0.046 (0.177)	0.043 (0.175)	−0.004 (0.117)	0.009 (0.137)	0.011 (0.139)	0.009 (0.137)
IV	$\mu_{11} : 0$	0.030 (0.257)	1.536 (1.510)	1.629 (1.568)	0.030 (0.257)	0.030 (0.256)	0.030 (0.256)
	$\mu_{12} : 0$	−0.008 (0.228)	−0.132 (0.207)	−0.012 (0.266)	−0.008 (0.228)	−0.008 (0.227)	−0.008 (0.227)
	$\mu_{21} : 3$	−0.002 (0.248)	−1.507 (1.510)	−1.601 (1.434)	−0.002 (0.248)	−0.002 (0.248)	−0.002 (0.248)
	$\mu_{22} : 0$	0.004 (0.225)	0.129 (0.159)	0.008 (0.177)	0.004 (0.225)	0.004 (0.224)	0.004 (0.225)
	$\pi_1 : 0.5$	0.004 (0.075)	−0.007 (0.075)	−0.056 (0.050)	0.004 (0.075)	0.004 (0.075)	0.004 (0.075)

Table 4. Bias (Std) of point estimates when $n = 400$ for Example 2.

Case	TRUE	OC-μ_1	OC-μ_2	OC-π	NORMLH	COMPLH	DISTLAT
I	μ_{11} : 0	−0.193 (0.544)	0.263 (0.767)	0.259 (0.991)	0.118 (0.949)	0.124 (0.952)	0.125 (0.954)
	μ_{12} : 0	0.325 (0.820)	−0.195 (0.580)	0.292 (1.060)	0.144 (1.030)	0.172 (1.034)	0.173 (1.035)
	μ_{21} : 0.5	0.326 (0.519)	−0.130 (0.700)	−0.126 (0.310)	0.015 (0.325)	0.009 (0.320)	0.008 (0.320)
	μ_{22} : 0.5	−0.163 (0.739)	0.357 (0.573)	−0.131 (0.304)	0.017 (0.312)	−0.010 (0.315)	−0.012 (0.313)
	π_1 : 0.3	0.171 (0.286)	0.174 (0.286)	−0.034 (0.167)	0.001 (0.207)	−0.006 (0.200)	−0.007 (0.199)
II	μ_{11} : 0	−0.008 (0.488)	0.258 (0.712)	0.210 (0.798)	0.102 (0.706)	0.104 (0.708)	0.104 (0.708)
	μ_{12} : 0	0.303 (0.800)	−0.011 (0.494)	0.246 (0.860)	0.132 (0.776)	0.130 (0.773)	0.130 (0.773)
	μ_{21} : 1	0.077 (0.345)	−0.191 (0.537)	−0.143 (0.316)	−0.035 (0.275)	−0.037 (0.277)	−0.037 (0.277)
	μ_{22} : 1	−0.194 (0.518)	0.121 (0.388)	−0.137 (0.322)	−0.023 (0.278)	−0.021 (0.276)	−0.021 (0.276)
	π_1 : 0.3	0.064 (0.211)	0.079 (0.221)	−0.011 (0.136)	0.009 (0.164)	0.010 (0.164)	0.010 (0.164)
III	μ_{11} : 0	−0.007 (0.217)	0.005 (0.252)	−0.005 (0.224)	−0.004 (0.232)	−0.004 (0.232)	−0.004 (0.232)
	μ_{12} : 0	0.022 (0.287)	0.007 (0.208)	0.022 (0.294)	0.013 (0.259)	0.013 (0.259)	0.013 (0.259)
	μ_{21} : 1.5	0.005 (0.112)	−0.008 (0.188)	0.003 (0.117)	0.002 (0.122)	0.002 (0.122)	0.002 (0.122)
	μ_{22} : 1.5	−0.007 (0.158)	0.008 (0.135)	−0.007 (0.143)	0.001 (0.120)	0.001 (0.119)	0.001 (0.119)
	π_1 : 0.3	0.003 (0.067)	0.005 (0.071)	0.002 (0.063)	0.002 (0.063)	0.002 (0.063)	0.002 (0.063)
IV	μ_{11} : 0	0.001 (0.097)	1.401 (1.51)	1.474 (1.533)	0.001 (0.097)	0.001 (0.097)	0.001 (0.097)
	μ_{12} : 0	−0.004 (0.082)	−0.052 (0.063)	0.001 (0.086)	−0.004 (0.082)	−0.004 (0.082)	−0.004 (0.082)
	μ_{21} : 3	0.001 (0.095)	−1.399 (1.500)	−1.473 (1.475)	0.001 (0.095)	0.001 (0.095)	0.001 (0.095)
	μ_{22} : 0	0.005 (0.082)	0.053 (0.063)	0.002 (0.078)	0.005 (0.082)	0.005 (0.082)	0.005 (0.082)
	π_1 : 0.5	0.000 (0.031)	−0.002 (0.031)	−0.025 (0.019)	0.000 (0.031)	−0.001 (0.031)	0.000 (0.031)

are considered: order constraint labelling based on the component means of first dimension (OC-μ_1), order constraint labelling based on the component means of second dimension (OC-μ_2), order constraint labelling based on the component proportions (OC-π), NORMLH, COMPLH, and DISTLAT.

Tables 3 and 4 report the estimated biases and Stds of the MLEs based on different labelling methods for $n = 100$ and 400, respectively. From the tables, it can be seen that, for Cases I, II, and III, COMPLH and DISTLAT have similar results and provide smaller estimated biases of component parameters than OC-μ_1, OC-μ_2, OC-π, and NORMLH, especially when the components are close and the sample size is not large. In addition, NORMLH also provides smaller estimated biases than OC-μ_1, OC-μ_2, and, OC-π.

In Case IV, the component means of the first dimension are very separate, but the component means of the second dimension and the component proportions are the same. In this case, the component means of the second dimension and component proportions do not have any label information. Based on the tables, OC-μ_2 and OC-π provide unreasonable estimates and have large estimated biases for both $n = 100$ and $n = 400$. Note that in this case, OC-μ_2 and OC-π do not work well even when the sample size is larger than 400 because the wrong component parameters are used for order constraint labelling. However, OC-μ_1, NORMLH, COMPLH, and DISTLAT all work well for both $n = 100$ and $n = 400$.

Example 3 Consider the multivariate normal mixture model

$$\sum_{j=1}^{3} \pi_j N\left(\begin{pmatrix}\mu_{j1}\\ \mu_{j2}\end{pmatrix}, \quad \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right),$$

where $\pi_1 = 0.2$, $\pi_2 = 0.3$, $\pi_3 = 0.5$, $\mu_{11} = 0$, and $\mu_{12} = 0$. The following three cases are considered for μ_{21}, μ_{22}, μ_{31}, and μ_{32}:

I. $\mu_{21} = 0.5$, $\mu_{22} = 0.5$, $\mu_{31} = 1$, $\mu_{32} = 1$;

Table 5. Bias (Std) of point estimates when $n = 100$ for Example 3.

Case	TRUE	OC-μ_1	OC-μ_2	OC-π	NORMLH	COMPLH	DISTLAT
I	$\mu_{11}:0$	−0.451 (0.719)	0.403 (1.110)	0.435 (1.504)	0.260 (1.480)	0.045 (1.255)	0.079 (1.225)
	$\mu_{12}:0$	0.408 (1.100)	−0.420 (0.695)	0.477 (1.465)	0.400 (1.440)	−0.001 (1.187)	0.045 (1.164)
	$\mu_{21}:0.5$	0.107 (0.353)	0.136 (0.853)	0.127 (0.863)	0.302 (0.837)	0.158 (0.911)	0.142 (0.988)
	$\mu_{22}:0.5$	0.137 (0.874)	0.120 (0.404)	0.163 (0.873)	0.218 (0.903)	0.246 (0.891)	0.235 (0.977)
	$\mu_{31}:1$	0.600 (0.666)	−0.285 (1.090)	−0.309 (0.389)	−0.308 (0.364)	0.052 (0.522)	0.034 (0.505)
	$\mu_{32}:1$	−0.221 (1.030)	0.625 (0.604)	−0.317 (0.414)	−0.295 (0.382)	0.080 (0.535)	0.043 (0.521)
	$\pi_1:0.2$	0.035 (0.159)	0.055 (0.187)	−0.077 (0.079)	−0.073 (0.084)	−0.018 (0.132)	−0.008 (0.137)
	$\pi_2:0.3$	0.171 (0.184)	0.160 (0.182)	0.014 (0.094)	0.013 (0.102)	0.052 (0.187)	0.032 (0.190)
II	$\mu_{11}:0$	−0.094 (0.671)	0.265 (0.922)	0.776 (1.530)	−0.025 (0.752)	0.089 (0.973)	0.180 (0.943)
	$\mu_{12}:0$	0.332 (1.070)	−0.108 (0.618)	0.905 (1.599)	0.039 (0.815)	0.149 (1.085)	0.145 (1.039)
	$\mu_{21}:1$	0.184 (0.449)	0.353 (0.948)	0.084 (0.995)	0.415 (0.625)	0.248 (0.807)	0.248 (1.061)
	$\mu_{22}:1$	0.298 (0.988)	0.228 (0.444)	0.058 (1.001)	0.144 (0.575)	0.256 (0.809)	0.352 (0.977)
	$\mu_{31}:2$	0.318 (0.566)	−0.212 (0.954)	−0.453 (0.506)	0.018 (0.852)	0.070 (0.555)	−0.021 (0.484)
	$\mu_{32}:2$	−0.124 (0.852)	0.387 (0.609)	−0.457 (0.541)	0.323 (0.683)	0.102 (0.566)	0.009 (0.498)
	$\pi_1:0.2$	0.026 (0.131)	0.033 (0.126)	−0.059 (0.076)	0.023 (0.123)	0.004 (0.118)	0.010 (0.113)
	$\pi_2:0.3$	0.120 (0.167)	0.132 (0.167)	0.025 (0.074)	0.150 (0.154)	0.080 (0.166)	0.037 (0.179)
III	$\mu_{11}:0$	−0.634 (0.595)	−0.032 (0.630)	−0.027 (1.120)	−0.457 (0.734)	−0.060 (0.819)	−0.040 (0.699)
	$\mu_{12}:0$	2.265 (1.890)	0.159 (0.763)	1.441 (2.138)	0.901 (1.770)	0.350 (1.255)	0.380 (1.219)
	$\mu_{21}:0$	−0.010 (0.193)	0.049 (0.754)	0.010 (0.483)	0.438 (0.726)	0.033 (0.781)	0.031 (0.958)
	$\mu_{22}:2$	0.108 (1.650)	0.415 (0.864)	−0.160 (1.444)	−0.070 (1.270)	0.331 (0.993)	0.400 (1.244)
	$\mu_{31}:0$	0.634 (0.616)	−0.027 (0.779)	0.006 (0.295)	0.009 (0.329)	0.015 (0.537)	−0.001 (0.407)
	$\mu_{32}:4$	−1.692 (1.780)	0.107 (0.551)	−0.600 (0.827)	−0.150 (0.508)	−0.001 (0.491)	−0.100 (0.455)
	$\pi_1:0.2$	0.084 (0.159)	0.032 (0.113)	−0.046 (0.076)	0.006 (0.106)	0.012 (0.107)	0.023 (0.105)
	$\pi_2:0.3$	0.116 (0.156)	0.069 (0.151)	0.025 (0.061)	0.003 (0.129)	0.041 (0.143)	−0.003 (0.146)

Table 6. Bias (Std) of point estimates when $n = 400$ for Example 3.

Case	TRUE	OC-μ_1	OC-μ_2	OC-π	NORMLH	COMPLH	DISTLAT
I	$\mu_{11}:0$	−0.369 (0.678)	0.233 (1.040)	0.485 (1.406)	0.463 (1.410)	−0.003 (1.185)	−0.001 (1.116)
	$\mu_{12}:0$	0.206 (1.020)	−0.364 (0.658)	0.406 (1.320)	0.392 (1.320)	−0.014 (1.11)	−0.019 (1.059)
	$\mu_{21}:0.5$	0.111 (0.325)	0.169 (0.784)	0.037 (0.751)	−0.009 (0.735)	0.182 (0.718)	0.241 (0.861)
	$\mu_{22}:0.5$	0.173 (0.678)	0.088 (0.294)	0.081 (0.766)	0.024 (0.761)	0.164 (0.705)	0.229 (0.819)
	$\mu_{31}:1$	0.499 (0.603)	−0.163 (0.886)	−0.283 (0.330)	−0.215 (0.318)	0.061 (0.418)	−0.001 (0.389)
	$\mu_{32}:1$	−0.194 (0.890)	0.461 (0.545)	−0.303 (0.339)	−0.231 (0.320)	0.036 (0.435)	−0.024 (0.394)
	$\pi_1:0.2$	0.028 (0.166)	0.031 (0.171)	−0.089 (0.077)	−0.089 (0.078)	−0.043 (0.124)	−0.032 (0.125)
	$\pi_2:0.3$	0.174 (0.190)	0.168 (0.201)	0.009 (0.092)	0.016 (0.105)	0.064 (0.187)	0.022 (0.189)
II	$\mu_{11}:0$	−0.063 (0.492)	0.132 (0.717)	0.733 (1.403)	0.063 (0.392)	0.062 (0.782)	0.041 (0.550)
	$\mu_{12}:0$	0.100 (0.638)	−0.076 (0.496)	0.701 (1.388)	0.074 (0.362)	0.012 (0.713)	0.007 (0.483)
	$\mu_{21}:1$	0.164 (0.427)	0.266 (0.850)	−0.012 (0.908)	0.318 (1.110)	0.220 (0.763)	0.318 (1.001)
	$\mu_{22}:1$	0.308 (0.916)	0.146 (0.438)	−0.001 (0.924)	0.265 (1.130)	0.229 (0.746)	0.324 (0.997)
	$\mu_{31}:2$	0.259 (0.512)	−0.040 (0.672)	−0.363 (0.405)	−0.023 (0.327)	0.076 (0.388)	−0.001 (0.327)
	$\mu_{32}:2$	−0.087 (0.661)	0.252 (0.481)	−0.379 (0.409)	−0.018 (0.339)	0.080 (0.398)	−0.009 (0.343)
	$\pi_1:0.2$	0.020 (0.102)	0.015 (0.104)	−0.057 (0.078)	0.051 (0.094)	0.003 (0.105)	0.026 (0.098)
	$\pi_2:0.3$	0.107 (0.164)	0.114 (0.157)	0.025 (0.062)	−0.033 (0.180)	0.062 (0.161)	−0.002 (0.174)
III	$\mu_{11}:0$	−0.308 (0.424)	−0.023 (0.320)	0.024 (0.739)	−0.014 (0.217)	0.001 (0.482)	0.009 (0.299)
	$\mu_{12}:0$	2.100 (1.790)	0.016 (0.441)	0.748 (1.609)	0.050 (0.398)	0.052 (0.598)	0.062 (0.533)
	$\mu_{21}:0$	−0.003 (0.092)	0.022 (0.568)	−0.009 (0.202)	0.033 (0.727)	0.004 (0.569)	0.013 (0.697)
	$\mu_{22}:2$	0.078 (1.630)	0.167 (0.718)	−0.265 (1.157)	0.192 (0.958)	0.162 (0.791)	0.178 (0.916)
	$\mu_{31}:0$	0.328 (0.471)	0.016 (0.433)	0.001 (0.160)	−0.002 (0.192)	0.011 (0.237)	−0.006 (0.194)
	$\mu_{32}:4$	−1.971 (1.730)	0.022 (0.289)	−0.278 (0.568)	−0.036 (0.277)	−0.008 (0.272)	−0.035 (0.275)
	$\pi_1:0.2$	0.110 (0.153)	0.014 (0.080)	−0.033 (0.071)	0.021 (0.077)	0.006 (0.083)	0.017 (0.078)
	$\pi_2:0.3$	0.089 (0.146)	0.024 (0.123)	0.023 (0.045)	−0.016 (0.122)	0.009 (0.111)	−0.012 (0.119)

II. $\mu_{21} = 1, \mu_{22} = 1, \mu_{31} = 2, \mu_{32} = 2$;
III. $\mu_{21} = 0, \mu_{22} = 2, \mu_{31} = 0, \mu_{32} = 4$.

In each case, 500 replicates are run for both sample size 100 and 400 and the MLE is found for each replicate by EM algorithm assuming equal covariance. The following six labelling methods are considered: OC-μ_1, OC-μ_2, OC-π, NORMLH, COMPLH, and DISTLAT.

Tables 5 and 6 report the estimated biases and Stds of the MLEs based on different labelling methods for $n = 100$ and 400, respectively. The findings are similar to those noted in Example 2. Based on the tables, COMPLH and DISTLAT produce similar results in most cases and have better overall performance than any other labelling method, especially when the sample size is small.

Example 4 Consider the mixture model with eight components $\sum_{j=1}^{8} \pi_j N(\mu_j, 1)$, where $\pi_1 = \pi_2 = 0.05, \pi_3 = \pi_4 = 0.1, \pi_5 = \pi_6 = 0.15, \pi_7 = \pi_8 = 0.2$, and $\mu_j = 2(j-1), j = 1, \ldots, 8$. Five hundred replicates are run for both sample sizes 200 and 400. The following five labelling methods are considered: OC-μ, OC-π, NORMLH, COMPLH, and DISTLAT.

Table 7 reports the estimated biases and Stds of the MLEs based on different labelling methods for both $n = 200$ and $n = 400$. In this example, since some of the component weights are equal, OC-π fails, which can be seen from the reported large estimated biases. COMPLH has smaller estimated biases than OC-μ and NORMLH for all parameter estimates except for μ_2. Note, however, DISTLAT fails to label the second component, which can be seen from the large estimated biases for μ_2. Based on our limited experience, although DISTLAT works well when the number of components is small, which can be seen in Examples 1–3, it is not stable when the number of components is large, especially when the parameter estimates and the estimated classification probabilities are not accurate. Therefore, we recommend COMPLH in practice.

To make the results reported in Tables 1–7 be easily understood, parts of absolute estimated biases versus different labelling methods have been plotted in Figure 1. Different symbols represent different cases of μ_2 in Figure 1(a) for Example 1, different cases of μ_{22} in Figure 1(b) for Example 2, different cases of μ_{32} in Figure 1(c) for Example 3 and different cases of component means in Figure 1(d) for Example 4. In view of Figures of 1(a)–(c), COMPLH and DISTLAT provide smaller estimated biases than any other method in Examples 1–3. In Example 4, OC-μ, NORMLH, and COMPLH provide similar results, while DISTLAT provides large estimated biases for one of the eight component means.

4. Discussion

Label switching issue has not received as much attention for frequentist mixture models as for Bayesian mixture models. In this article, we explained the importance of solving label switching issue for frequentist mixture models in the simulation study and proposed a new labelling method based on the latent component label for each observation. Based on the Monte Carlo study, the proposed complete likelihood-based labelling method (COMPLH) by incorporating the information of latent labels has overall better performance than any of the other method considered in the simulation study. In addition, it can be seen that the order constraint labelling methods only work well when the constrained component parameters have enough label information. These methods usually provide poor and even unreasonable estimates if the constrained component parameters are not very separate. Therefore, in practice, the choice of constraint is very sensitive and thus difficult, especially for multivariate data. Different order constraints may generate significantly different results and it is difficult to predict the overall effect. In addition, NORMLH has overall

better performance than order constraint labelling methods and can be easily applied to solve multivariate problems.

Note that the label switching problem also exists if one wants to use the bootstrap procedure to evaluate the accuracy of the mixture model parameter estimates. The parametric bootstrap setting is similar to the simulation study. Therefore, the proposed methods can be directly applied to the parametric bootstrap. The nonparametric bootstrap situation is a little more tricky. Note that for the nonparametric bootstrap method, the latent component labels for the observations are unknown. One possible solution is to estimate the latent labels based on the MLE and the corresponding classification probabilities $\{\hat{p}_{ij}, i = 1, \ldots, n, j = 1, \ldots, m\}$ obtained from the original samples $(x_1, \ldots, x_n)$. One can simply let $z_{ij} = \hat{p}_{ij}$ in Equations (5) and (7) or let

$$z_{ij} = \begin{cases} 1 & \text{if } \hat{p}_{ij} > \hat{p}_{il} \text{ for all } l \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the computations will be similar. However, more study is required to determine how well this labelling method is when compared to other existing labelling methods and whether there are better ways to estimate the latent labels z_{ij}s.

It will be also interesting to understand how to sensibly *define and construct* the confidence intervals for the unknown mixture parameters. One might simply use the sample percentile based on the labelled samples to construct the marginal confidence intervals. However, such constructed confidence intervals for different component parameters, for example μ_1 and μ_2, might overlap. It might make more sense to consider joint confidence regions for the unordered pair $\{\mu_1, \mu_2\}$.

Table 7. Bias (Std) of point estimates for Example 4.

n	TRUE	OC-μ	OC-π	NORMLH	COMPLH	DISTLAT
$n = 200$	$\mu_1 : 0$	−0.234 (0.687)	4.216 (6.816)	−0.234 (0.687)	−0.213 (0.868)	0.060 (1.873)
	$\mu_2 : 2$	0.346 (0.950)	1.657 (4.624)	0.346 (0.950)	0.710 (2.163)	3.143 (5.800)
	$\mu_3 : 4$	0.706 (1.039)	1.340 (4.128)	0.707 (1.039)	0.539 (0.965)	0.325 (2.156)
	$\mu_4 : 6$	0.849 (1.185)	1.178 (3.863)	0.850 (1.187)	0.850 (1.351)	0.743 (2.492)
	$\mu_5 : 8$	0.930 (1.200)	0.432 (3.490)	0.932 (1.203)	0.755 (1.037)	0.158 (0.996)
	$\mu_6 : 10$	0.830 (0.984)	−0.197 (3.091)	0.828 (0.986)	0.816 (0.980)	0.226 (0.855)
	$\mu_7 : 12$	0.793 (0.784)	−1.390 (2.604)	0.792 (0.783)	0.781 (0.794)	0.136 (0.622)
	$\mu_8 : 14$	0.713 (0.780)	−2.304 (1.859)	0.713 (0.780)	0.696 (0.758)	0.142 (0.470)
	$\pi_1 : 0.05$	0.001 (0.023)	−0.015 (0.019)	−0.001 (0.023)	−0.001 (0.024)	0.001 (0.023)
	$\pi_2 : 0.05$	0.033 (0.033)	0.014 (0.017)	0.033 (0.033)	0.030 (0.031)	0.013 (0.031)
	$\pi_3 : 0.1$	0.010 (0.036)	−0.013 (0.017)	0.010 (0.036)	0.009 (0.035)	0.003 (0.034)
	$\pi_4 : 0.1$	0.028 (0.047)	0.009 (0.017)	0.027 (0.047)	0.023 (0.044)	0.008 (0.039)
	$\pi_5 : 0.15$	−0.003 (0.044)	−0.019 (0.017)	−0.004 (0.045)	0.002 (0.044)	0.001 (0.044)
	$\pi_6 : 0.15$	0.015 (0.048)	0.007 (0.018)	0.016 (0.047)	0.015 (0.047)	0.008 (0.047)
	$\pi_7 : 0.2$	0.002 (0.049)	−0.013 (0.022)	0.002 (0.049)	0.001 (0.052)	0.000 (0.052)
$n = 400$	$\mu_1 : 0$	−0.235 (0.610)	3.116 (5.964)	−0.235 (0.610)	−0.212 (0.844)	−0.033 (1.356)
	$\mu_2 : 2$	0.198 (0.876)	1.069 (4.043)	0.198 (0.876)	0.316 (1.377)	2.571 (5.431)
	$\mu_3 : 4$	0.485 (0.970)	1.071 (3.722)	0.486 (0.970)	0.401 (0.887)	0.109 (1.732)
	$\mu_4 : 6$	0.640 (1.131)	0.830 (3.516)	0.641 (1.133)	0.637 (1.160)	0.187 (1.577)
	$\mu_5 : 8$	0.646 (1.182)	0.598 (3.347)	0.645 (1.183)	0.600 (1.123)	0.071 (0.633)
	$\mu_6 : 10$	0.610 (1.017)	−0.416 (2.828)	0.612 (1.020)	0.603 (1.006)	0.190 (0.683)
	$\mu_7 : 12$	0.589 (0.741)	−0.897 (2.464)	0.588 (0.741)	0.589 (0.741)	0.175 (0.513)
	$\mu_8 : 14$	0.483 (0.663)	−1.950 (1.608)	0.483 (0.663)	0.483 (0.662)	0.148 (0.373)
	$\pi_1 : 0.05$	−0.003 (0.021)	−0.014 (0.018)	−0.003 (0.021)	−0.003 (0.021)	−0.001 (0.020)
	$\pi_2 : 0.05$	0.024 (0.029)	0.012 (0.016)	0.024 (0.029)	0.023 (0.028)	0.007 (0.028)
	$\pi_3 : 0.1$	0.009 (0.030)	−0.012 (0.017)	0.009 (0.030)	0.008 (0.029)	0.003 (0.028)
	$\pi_4 : 0.1$	0.020 (0.042)	0.012 (0.016)	0.020 (0.042)	0.020 (0.042)	0.009 (0.034)
	$\pi_5 : 0.15$	−0.004 (0.038)	−0.015 (0.015)	−0.003 (0.037)	−0.003 (0.037)	−0.001 (0.035)
	$\pi_6 : 0.15$	0.013 (0.043)	0.008 (0.016)	0.014 (0.043)	0.013 (0.043)	0.012 (0.041)
	$\pi_7 : 0.2$	0.001 (0.046)	−0.015 (0.020)	−0.000 (0.047)	−0.001 (0.048)	−0.001 (0.048)

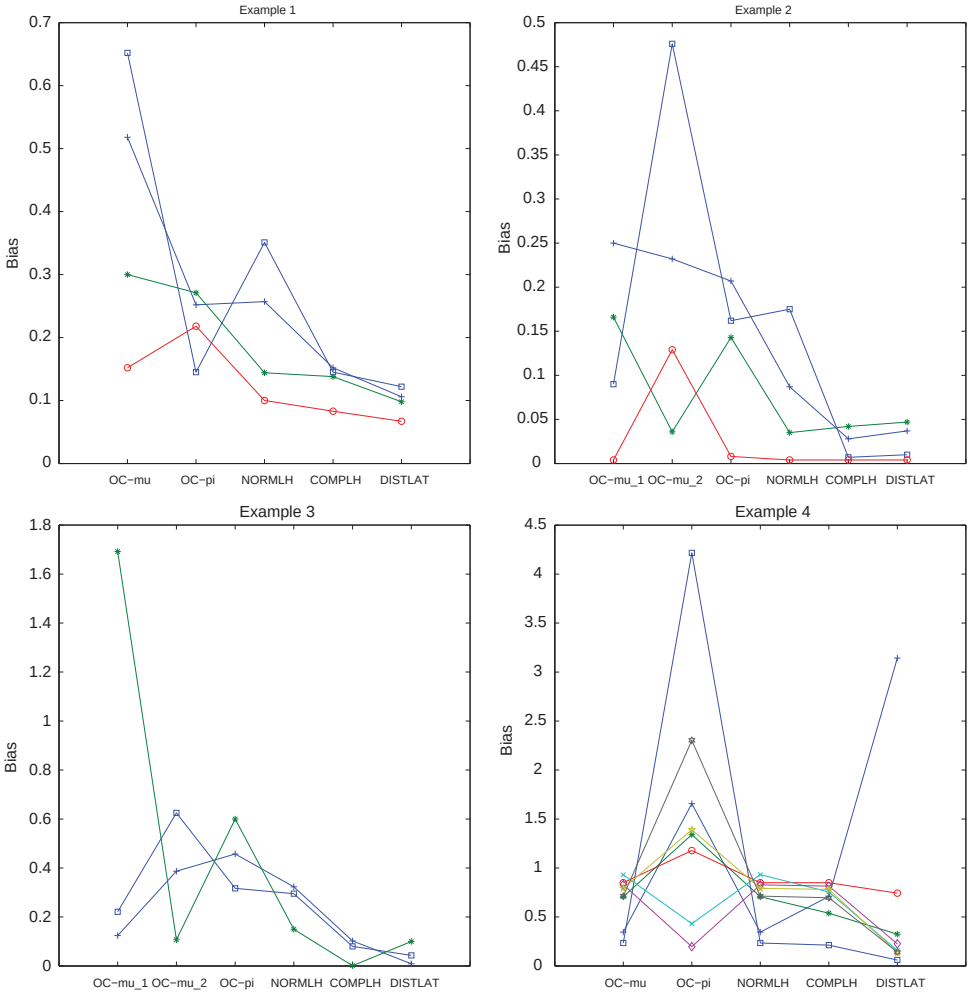

Figure 1. The plot of absolute bias for different labelling methods in (a) Example 1 with $n = 50$; (b) Example 2 with $n = 100$; (c) Example 3 with $n = 100$; (d) Example 4 with $n = 200$. Different symbols represent different cases for Examples 1–3 and represent different component means for Example 4.

However, it is not clear how to effectively construct such confidence regions and whether the relabelling would help to achieve this. One possible solution is to use the likelihood function to construct the confidence regions. See, for example, Daeyoung and Lindsay.[24] More research is required to determine how to apply the idea of the likelihood confidence regions to the mixture model setting.

References

[1] Stephens M. Dealing with label switching in mixture models. J R Stat Soc Ser B Stat Methodol. 2000;62:795–809.
[2] Celeux G, Hurn M, Robert CP. Computational and inferential difficulties with mixture posterior distributions. J Amer Statist Assoc. 2000;95:957–970.
[3] Chung H, Loken E, Schafer JL. Difficulties in drawing inferences with finite-mixture models: a simple example with a simple solution. Amer Statist. 2004;58:152–158.
[4] Geweke J. Interpretation and inference in mixture models: simple MCMC works. Comput Statist Data Anal. 2007;51:3529–3550.
[5] Yao W, Lindsay BG. Bayesian mixture labeling by highest posterior density. J Amer Statist Assoc. 2009;104:758–767.

[6] Grün B, Leisch F. Dealing with label switching in mixture models under genuine multimodality. J Multivariate Anal. 2009;100:851–861.
[7] Sperrin M, Jaki T, Wit E. Probabilistic relabeling strategies for the label switching problem in Bayesian mixture models. Stat Comput. 2010;20:357–366.
[8] Papastamoulis P, Iliopoulos G. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. J Comput Graph Statist. 2010;19:313–331.
[9] Cron AJ, West M. Efficient classification-based relabeling in mixture models. Amer Statist. 2011;65:16–20.
[10] Yao W. Model based labeling for mixture models. Stat Comput. 2012;22:337–347.
[11] Yao W. Bayesian mixture labeling and clustering. Comm Statist Theory Methods. 2012;41:403–421.
[12] Celeux G. Bayesian inference for mixtures: the label switching problem. In: Payne R, Green PJ, editors. Compstat 98-Proceedings in Computational Statistics. Physica: Heidelberg; 1998; p. 227–232.
[13] Taha HA. Operations research: an introduction. 4th ed. New York: Macmillan; 1989.
[14] Kuhn HW. The Hungarian method for the assignment problem. Naval Res Logist Quart. 1955;2:83–97.
[15] Feng Z, McCulloch CE. Using bootstrap likelihood ratios in finite mixture models. J R Stat Soc Ser B Stat Methodol. 1996;58:609–617.
[16] Chen R, Liu W. The consistency of estimators in finite mixture models. Scand J Statist. 2001;28:603–616.
[17] Hathaway RJ. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann Statist. 1985;13:795–800.
[18] Hathaway RJ. A constrained EM algorithm for univariate mixtures. J Stat Comput Simul. 1986;23:211–230.
[19] Chen J, Tan X, Zhang R. Inference for normal mixture in mean and variance. Statist Sincia. 2008;18:443–465.
[20] Chen J, Tan X. Inference for multivariate normal mixtures. J Multivariate Anal. 2009;100:1367–1383.
[21] Yao W. A profile likelihood method for normal mixture with unequal variance. J Statist Plann Inference. 2010;140:2089–2098.
[22] Lindsey JK. Some statistical heresies. The Statistician. 1999;48:1–40.
[23] Fisher RA. On the mathematical foundations of theoretical statistics. Philos Trans R Soc Lond Ser A Math Phys Eng Sci 1922;222:309–368.
[24] Kim D, Lindsay BG. Using confidence distribution sampling to visualize confidence sets. Statist Sinica. 2011;21: 923–948.