Statistica Sinica Preprint No: SS-2016-0483	
Title	A CLASSICAL INVARIANCE APPROACH TO THE NORMAL MIXTURE PROBLEM
Manuscript ID	SS-2016-0483
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0483
Complete List of Authors	Monia Ranalli Bruce G. Lindsay and David R. Hunter
Corresponding Author	Monia Ranalli
E-mail	mxr459@psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

A CLASSICAL INVARIANCE APPROACH TO THE NORMAL MIXTURE PROBLEM

Monia Ranalli, Bruce G. Lindsay, and David R. Hunter

Pennsylvania State University

Abstract:

Although normal mixture models have received great attention and are commonly used in different fields, they stand out for failing to have a finite maximum on the likelihood. In the univariate case, there are n solutions, corresponding to n distinct data points, along a parameter boundary, each with an infinite spike of the likelihood, and none making particular sense as a chosen solution. The multivariate case yields an even more complex likelihood surface. In this paper, we show that there is a marginal likelihood that is bounded and quite close to the full likelihood in information, as long as one is interested in the central part of the parameter space, away from its problematic boundaries. Our main goal is to show that the marginal likelihood solves the unboundedness problem in a manner competitive with other methods that were specifically designed for the normal mixture. To this end two algorithms have been developed. Their effectiveness is investigated through a simulation study. Finally, an application to real data is illustrated.

Key words: Mixture models, Marginal Likelihood, Monte Carlo Likelihood, EM algorithm

1 Introduction and background

Finite mixture models have played a central role in statistical modeling since they were turned into flexible tools to address the departures from the classical inferential assumption of normality by Pearson (1894). They have received increasing interest and are widely used in different fields, such as genetics, economics, marketing, engineering, and social sciences, among many others. According to their use, finite mixture models can

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

have different interpretations. They can be used in a clustering or in a classification context, as well as in a semiparametric or nonparametric framework. Their success is mainly the result of their simplicity when being fitted and interpreted. Such models arise naturally in contexts where the assumption of homogeneity is not reliable. Their mathematical structure implies that a population is a convex combination of a finite number of sub-populations represented by a finite number of densities.

The price to be paid for this flexibility is a challenging inference. Normal mixture models are the province of many likelihood anomalies, such as the failure of seemingly standard hypothesis testing problems to have a limiting chi-squared null distribution, and these models stand out for failing to have a finite maximum on the likelihood. In the univariate case, there are n solutions, corresponding to n distinct data points, along a parameter boundary, each with an infinite spike to the likelihood, and none making particular sense as a chosen solution: one can set μ_1 equal to any observation, and let σ_1^2 approach zero (Kiefer and Wolfowitz, 1956). In the multivariate d-dimensional case, there is an even more complex likelihood surface. The likelihood tends to infinity when the covariance matrix has one eigenvector that is parallel to one of the observations, and the corresponding eigenvalue goes to zero.

In the literature, it is known that there exists a sequence of roots of the likelihood equation that is consistent and asymptotically efficient (Kiefer, 1978; Peters and Walker, 1978). Nevertheless, for a given sample, multiple local maxima may exist; hence, the other major maximum-likelihood difficulty is determining when the correct one has been found. See, for example, Hathaway (1985), and the references therein.

Various methods have been proposed to avoid the unboundedness of the likelihood by constraining the parameter space. Hathaway (1985, 1986) suggests running the EM algorithm (the typical optimization tool used to maximize the likelihood of obtaining

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

parameter estimates) by constraining the ratio of the mixture component variances. Formally, in the univariate mixture case with G components, this means that $\sigma_h/\sigma_j \geq c > 0$, where $1 \leq h \neq j \leq G$. Similar solutions can be found in the literature; see, for example, Tanaka and Takemura (2006). However, the main issue is related to the choice of c: if it is too large, we might exclude the true parameters, and if it is not large enough, the maximum can occur at the boundary, where the ratio of variances is equal to c (i.e., σ_h/σ_j is close to zerp). To overcome this problem, constrained solutions based on the output of the EM algorithm have been proposed; see Ingrassia and Rocci (2007, 2011), and the references therein. Furthermore, other proposals use a penalty term on the scale parameter to avoid infinite spikes (Ciuperca et al., 2003; Chen and Tan, 2009; Chen et al., 2008). Once again, this approach requires some subjective choice: the penalized likelihood method works well with properly chosen penalty functions, but the choice of a penalty function in a finite sample is still a problem.

Furthermore, there exist solutions based on the conditional likelihood (Policello II, 1981), profile likelihood (Yao, 2010), or doubly smoothed maximum likelihood estimator (DS-MLE) (Seo and Lindsay, 2010). Although the DS-MLE still requires a subjective choice, that is, the bandwidth, it can be shown to be consistent with any fixed bandwidth. This implies that the DS-MLE is robust to the choice of bandwidths, even in small samples.

While not an exhaustive list, the works describe here indicate that all of the approaches share the same underlying intuition: researchers find the mixture likelihood quite acceptable for use, provided that they are looking for a local maximum that lies away from the chaotic boundary. One might conjecture that this likelihood is close to being quite stable, which is a fact we exploit in this study.

We show that there is a marginal likelihood that is bounded and quite close to the full likelihood in information, as long as one is interested in the central part of the parameter

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

space, away from its problematic boundaries. Our main goal is to show that the marginal likelihood solves the unboundedness problem in a manner competitive with other methods that were specifically designed for the normal mixture.

The remainder of the paper is organized as follows. We first briefly describe the intuition behind our proposal in Section 2. Then, we prove that the invariant likelihood for a univariate mixture is bounded in Section 3. In Sections 4 and 5, we introduce the calculations needed to obtain the Monte Carlo likelihood and the EM-like algorithm to get the parameter estimates, respectively. In Sections 7 and 8, the effectiveness of the proposals is investigated and proved through a comparative simulation study and an application to a data set on the acidity index of lakes, respectively.

2 Classical invariance and the normal model

The motivation for adopting a marginal likelihood approach can be drawn from a classic likelihood analysis. In the standard $N(\mu, \sigma^2)$ model for a univariate random sample $X_1, \ldots, X_n$, we have the sufficient statistics $\bar{X}$ and $S^2 = \sum(X_i - \bar{X})^2$. The marginal distribution for S^2 is $\sigma^2\chi^2_{(n-1)}$. This distribution can then be used to make inferences about σ^2, free of the parameter μ. This analysis can be constructed from first principles using a group invariance argument. If we consider data transformations of the form

$$x \to a + x = y,$$

for arbitrary $a \in \mathcal{R}$, then the new y sample has a $N(\mu + a, \sigma^2)$ density. If we need to make an inference about σ^2 that is free of the choice of a, and hence of the value of the parameter μ, we can focus our attention on the marginal distribution of S^2, which is the maximal invariant statistic (see Cox and Hinkley, 1979, Example 5.14). However,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

one might ask if information about the parameter of interest σ^2 is lost when using the marginal distribution of S^2. The answer, in a Fisherian sense, is yes, in that the Fisher information in S^2 about the parameter σ^2 is $(n-1)/2\sigma^4$, whereas the information in the full likelihood is $n/2\sigma^4$. The missing information, $1/2\sigma^4$, can be found in the conditional distribution of $\bar{X}$, given S^2, which here is also the marginal distribution of $\bar{X}$. At least on some intuitive grounds, we might consider the remaining information about σ^2 in $\bar{X}$ to be irretrievable, owing to the presence of the unknown μ. However, the relative information about σ^2 in S^2 compared to the full likelihood is $(n-1)/n$, and so the marginal likelihood certainly has a large-sample justification.

The d-dimensional multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has a similar analysis, where one may use the class of multivariate location transformations

$$\mathbf{x} \to \mathbf{a} + \mathbf{x} = \mathbf{y}.$$

Now, the maximal invariant is $S^2 = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})\ (\mathbf{X}_i - \bar{\mathbf{X}})^\top$, which has a Wishart distribution with $n-1$ degrees of freedom, depending only on the parameter $\boldsymbol{\Sigma}$.

2.1 The normal mixture likelihood

For the mixture framework, the previous argument can be extended as follows. Here, we are concerned with the two-component multivariate normal mixture density

$$\varphi(\mathbf{x}_i, \boldsymbol{\theta}) = p\phi(\mathbf{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma_1}) + (1-p)\phi(\mathbf{x}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \tag{1}$$

where the parameter space for $\boldsymbol{\theta} = (p, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ is $p \in (0,1)$, $\boldsymbol{\mu}_g \in R^d$, and $\boldsymbol{\Sigma}_g$ is in the set of nonnegative definite matrices. This parameterization has a labeling noniden-tifiability, in that one can interchange component 1 with component 2, and change p to

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

$(1-p)$, thereby achieving the same density. However, Equation (1) is otherwise a regular model, and the mapping is locally identifiable in the parameter space. We consider a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from this density.

Let us define $\mathbf{U}_i = \mathbf{X}_i - \mathbf{X}_n$ for $1 \leq i \leq n-1$. Because $\mathbf{U}_i = (\mathbf{X}_i - \mathbf{a}) - (\mathbf{X}_n - \mathbf{a})$ for any vector $\mathbf{a}$, letting $\mathbf{a} = \boldsymbol{\mu}_1$ we notice that the distribution of $\mathbf{U}_i$ must depend on the distribution of $\mathbf{X}_1 - \boldsymbol{\mu}_1, \ldots, \mathbf{X}_n - \boldsymbol{\mu}_1$, which is a sample from a mixture of normals with means 0 and $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. We conclude that the distribution of $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_{n-1})$ depends only on the parameters $\boldsymbol{\tau} = (p, \boldsymbol{\delta}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$. We therefore introduce the notation $h_{\boldsymbol{\tau}}(\mathbf{u})$ for the density of $\mathbf{U}$, and refer to $h_{\boldsymbol{\tau}}(\mathbf{u})$ as the marginal likelihood. Similarly, we call $f_{\boldsymbol{\theta}}(\mathbf{x})$ and $f_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})$ the full and conditional likelihoods, respectively, when they are viewed as functions of $\boldsymbol{\theta}$. We may derive an expression for $h_{\boldsymbol{\tau}}(\mathbf{u})$ by integrating out $\mathbf{x}_n$ from the joint density of $(\mathbf{U}, X_n)$:

$$h_{\boldsymbol{\tau}}(\mathbf{u}) = \int f_{\boldsymbol{\theta}}(\mathbf{u}_1 + \mathbf{x}) \cdots f_{\boldsymbol{\theta}}(\mathbf{u}_{n-1} + \mathbf{x}) f_{\boldsymbol{\theta}}(\mathbf{x})\, d\mathbf{x}. \tag{2}$$

Under the change of variables $\mathbf{t} = \mathbf{x} - \mathbf{x}_n$, and recalling that $\mathbf{x}_i = \mathbf{u}_i + \mathbf{x}_n$, we may rewrite Equation (2) as

$$h_{\boldsymbol{\tau}}(\mathbf{u}) = \int f_{\boldsymbol{\theta}}(\mathbf{x}_1 + \mathbf{t}) \cdots f_{\boldsymbol{\theta}}(\mathbf{x}_{n-1} + \mathbf{t}) f_{\boldsymbol{\theta}}(\mathbf{x}_n + \mathbf{t})\, d\mathbf{t}. \tag{3}$$

Equation (3) re-expresses the marginal likelihood in terms of the complete data, which is a computational trick we will use later in proving that the marginal likelihood is bounded in the case of univariate data. The form of Equation (3) also makes clear that the choice of $\mathbf{X}_n$ in the definition $\mathbf{U}_i = \mathbf{X}_i - \mathbf{X}_n$ is arbitrary, made here simply for convenience of notation.

This study proposes a two-stage estimation algorithm. The first stage estimates the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

parameters $\boldsymbol{\tau}$ using a maximum likelihood estimation, which we show is a well-behaved problem in the two-component unidimensional case. This latter fact is one of the main contributions of this study because, in general, the MLE is ill-behaved, theoretically, in this case. The second stage estimates the parameter $\boldsymbol{\mu}_1$, given our estimates of the $\boldsymbol{\tau}$ parameters. In the next section, we show that the likelihood for $\boldsymbol{\tau}$ is bounded in the important special case of univariate data.

3 The bounded marginal likelihood

In order to show that the marginal likelihood is bounded, we condition on $\mathbf{Z}$, the vector that indicates the group from which each observation is drawn, and write

$$h_{\boldsymbol{\tau}}(\mathbf{u}) = \sum_{\mathbf{z}} h_{\boldsymbol{\tau}}(\mathbf{u} \mid \mathbf{z}) \Pr(\mathbf{Z} = \mathbf{z}). \tag{4}$$

Each value of $\mathbf{z}$ effectively partitions the data into two sets, corresponding to observations drawn from each component. Let $n_1 = n_1(\mathbf{z})$ and $n_2 = n_2(\mathbf{z})$ be the sizes of groups 1 and 2, respectively. Given $\mathbf{Z} = \mathbf{z}$, define $\mathbf{v}_1, ..., \mathbf{v}_{n_1}$ as a relabeling of the n_1 observations among $\mathbf{x}_1, \ldots, \mathbf{x}_n$ drawn from the first component and, similarly, $\mathbf{w}_1, ..., \mathbf{w}_{n_2}$ as the observations from the second component. From Equation (3), we may write

$$\begin{aligned} h_{\boldsymbol{\tau}}(\mathbf{u} \mid \mathbf{z}) &= \int \prod_{i:z_i=1} f_1(\mathbf{x}_i + \mathbf{t}) \prod_{i:z_i=2} f_2(\mathbf{x}_i + \mathbf{t}) d\mathbf{t} \\ &= \int f_1(\mathbf{v}_1 + \mathbf{t}) \cdots f_1(\mathbf{v}_{n_1} + \mathbf{t}) f_2(\mathbf{w}_1 + \mathbf{t}) \cdots f_2(\mathbf{w}_{n_2} + \mathbf{t}) d\mathbf{t}, \end{aligned} \tag{5}$$

where f_1 and f_2 are normal densities with parameters $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively. We consider whether the summands of Equation (4) are bounded over the parameter space. In the one-dimensional case, we can show that the answer is yes.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

Proposition 1. *Let f_1 and f_2 be normal densities with parameters $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively, and let $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote a sample from a mixture density, placing probabilities p on f_1 and $1-p$ on f_2. Conditional on $\boldsymbol{Z} = \boldsymbol{z}$, let $\boldsymbol{v}_1, ..., \boldsymbol{v}_{n_1}$ denote the observations drawn from f_1, and $\boldsymbol{w}_1, ..., \boldsymbol{w}_{n_2}$ denote those drawn from f_2. Then, the marginal likelihood for data $\boldsymbol{u}_i = \boldsymbol{x}_i - \boldsymbol{x}_n$, for $1 \leq i \leq n-1$, may be expressed as a function of $\boldsymbol{\tau} = (p, \boldsymbol{\delta}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$, as follows:*

$$L(\boldsymbol{\tau}) = \sum_{\boldsymbol{z}} p^{n_1}(1-p)^{n_2} \int f_1(\boldsymbol{v}_1 + \boldsymbol{t}) \cdots f_1(\boldsymbol{v}_{n_1} + \boldsymbol{t}) f_2(\boldsymbol{w}_1 + \boldsymbol{t}) \cdots f_2(\boldsymbol{w}_{n_2} + \boldsymbol{t})\, d\boldsymbol{t}. \quad (6)$$

Furthermore, in the univariate case, $L(\boldsymbol{\tau})$ is bounded with probability one.

The first claim of Proposition 1 is merely a combination of Equations (4) and (5), yet it is a nontrivial statement in the sense that Expression (6), which clearly depends on μ_1 and μ_2, may depend on these parameters only through $\boldsymbol{\delta}$. The second claim, on the boundedness in the univariate case, relies on the following lemma, the proof of which is given in Supplementary Material (S1).

Lemma 1. *The integral in Equation (6) may be maximized over the parameter $\boldsymbol{\delta}$ in closed form, and the resulting maximum may be rewritten as*

$$K \left[(\det \boldsymbol{\Sigma}_1)^{n_1 - 1} (\det \boldsymbol{\Sigma}_2)^{n_2 - 1} \det(n_1 \boldsymbol{\Sigma}_2 + n_2 \boldsymbol{\Sigma}_1) \right]^{-1/2} \exp\left\{ -\frac{1}{2}(S_v^2 + S_w^2) \right\}, \quad (7)$$

where K is a constant, not depending on any parameters, $S_v^2 = \sum_{i=1}^{n_1} (\boldsymbol{v}_i - \overline{\boldsymbol{v}})^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{v}_i - \overline{\boldsymbol{v}})$, and $S_w^2 = \sum_{j=1}^{n_2} (\boldsymbol{w}_i - \overline{\boldsymbol{w}})^\top \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{w}_i - \overline{\boldsymbol{w}})$.

In the univariate case, Expression (7) may be rewritten as

$$\frac{K}{\sigma_1^{n_1 - 1} \sigma_2^{n_2 - 1} (n_1 \sigma_2^2 + n_2 \sigma_1^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^{n_1} (v_i - \overline{v})^2 - \frac{1}{2\sigma_2^2} \sum_{j=1}^{n_2} (w_j - \overline{w})^2 \right\}. \quad (8)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

The only way that Expression (8) can be unbounded as a function of the parameters σ_1 and σ_2 is if one of the sums disappears. With probability one, each sum will be nonzero, as long as it consists of at least two summands. Thus, the only cases that must be examined specifically are $n_1 = 0$ and $n_1 = 1$; because $n_2 = n - n_1$, the cases $n_1 = n$ and $n_1 = n - 1$ may be treated similarly. In addition, because σ_1 disappears from the denominator of (8) when $n_1 = 0$ or $n_1 = 1$, we conclude that (8) is bounded with probability one.

To see why (7) can be unbounded for multivariate data, observe that when $n_1 = 2$, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\overline{\boldsymbol{v}}$ must be collinear, and so the vectors $\boldsymbol{v}_i - \overline{\boldsymbol{v}}$ in S_v^2 both point in the same direction. We may construct a covariance matrix $\boldsymbol{\Sigma}_1$ with one of its eigenvectors orthogonal to this direction; thus, the corresponding eigenvalue disappears entirely from the expression S_v^2. Because this eigenvalue is arbitrary, the value of $\det \boldsymbol{\Sigma}_1$ can be made arbitrarily small, which means that (7) can be made arbitrarily large when $n_1 = 2$.

4 Monte Carlo calculation of the invariant likelihood

In order to obtain the parameter estimates, the marginal likelihood, which cannot be explicitly calculated, is approximated as a Monte Carlo likelihood (Geyer and Thompson, 1992; Geyer, 1994). The approximation is based on pseudodata obtained using simulated values for x_n, and the partition of the data into the two components via $\mathbf{z}$. Given this aim, we implement an innovative sampling importance scheme. Here, we obtain an importance sample from a seed distribution by running B independent Gibbs samplers.

One way to understand our approach is that our algorithm attempts to exactly maximize an approximate likelihood based on the importance sampling idea described below. In contrast, several existing stochastic EM algorithms in the literature attempt to approximately maximize an exact likelihood. As summarized in Celeux et al. (1995), the main stochastic versions of EM are the SEM algorithm (Broniatowski et al., 1983; Celeux and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

Diebolt, 1985), SAEM algorithm (Celeux and Diebolt, 1992), and MCEM algorithm (Wei and Tanner, 1990; Tanner, 1991). All of these algorithms introduce a simulation step that uses of pseudorandom draws at each iteration, although there are some differences between the SEM and MCEM. The SEM generates pseudo-complete samples by drawing potential unobserved samples from their conditional density, given the observed data. On the other hand, the MCEM replaces the analytic computation of the conditional expectation of the log-likelihood of the complete data, given the observations from a Monte Carlo approximation.

The main advantage of algorithms is that they eliminate the need for a simulation step at each iteration of the EM algorithm; the seed distribution generated initially is reused repeatedly. This saves computing time, and mitigates the tendency of Gibbs samplers to produce highly dependent observations.

We consider a general framework, in which $\mathbf{U}$ and $\mathbf{T}$ represent observed and hidden variables, respectively. Later, we will consider the specific case in which $\mathbf{U}$ is defined as in Section 2.1 and $\mathbf{T} = (X_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, where X_n and $\mathbf{Z}_i$ are as defined in Sections 2.1 and 3, respectively. In importance sampling, the parameter vector $\boldsymbol{\tau}_0$ is fixed, and samples are drawn independently from an importance density $g_{\text{imp};\boldsymbol{\tau}_0}$. We exploit the identity

$$\mathbb{E}_g\left[\frac{h_{\boldsymbol{\tau}}(\mathbf{U},\mathbf{T})}{g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{T} \mid \mathbf{U})} \mid \mathbf{U} = \mathbf{u}\right] = \int \frac{h_{\boldsymbol{\tau}}(\mathbf{u},\mathbf{t})}{g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t} \mid \mathbf{u})} g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t} \mid \mathbf{u})\mathrm{d}\mathbf{t} = h_{\boldsymbol{\tau}}(\mathbf{u}) \tag{9}$$

to construct the Monte Carlo likelihood, given by

$$M(\boldsymbol{\tau}) = \frac{1}{B}\sum_{b=1}^{B} \frac{h_{\boldsymbol{\tau}}(\mathbf{u},\mathbf{t}_b)}{g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t}_b \mid \mathbf{u})}, \tag{10}$$

where $\mathbf{t}_1, \ldots, \mathbf{t}_B$ is a sample from $g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t} \mid \mathbf{u})$.

Regardless of how the importance density $g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t} \mid \mathbf{u})$ is chosen, the law of large

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

numbers implies that $M(\boldsymbol{\tau}) \xrightarrow{p} h_{\boldsymbol{\tau}}(\mathbf{u})$ for all values of $\boldsymbol{\tau}$, assuming the expectation in Equation (9) exists. However, there are simple cases of Monte Carlo likelihoods where this convergence is so slow that it essentially never occurs in practice, particularly when $\boldsymbol{\tau}$ is far away from $\boldsymbol{\tau}_0$ (e.g., Hummel et al., 2012, Section 3). Thus, a sensible choice of $g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t} \mid \mathbf{u})$ is important, and to reduce the variance of the summands in Equation (10), we want this conditional density to be close to $h_{\boldsymbol{\tau}_0}(\mathbf{t} \mid \mathbf{u})$. The strategy we use to construct this importance density is elaborate, though it is not demanding computationally, and is very accurate. Furthermore, our proposed method, unlike methods such as the Monte Carlo EM, requires obtaining the importance density just once, before the estimation step.

We start by generating an initial estimate of $\boldsymbol{\tau} = (p, \delta, \sigma_1, \sigma_2)$ via a k-means clustering solution with two components. Then, we obtain sample estimates of p, $\delta = \mu_2 - \mu_1$, and the two variances. There are various means by which we might improve this initial $\boldsymbol{\tau}$ estimate, which we denote as $\boldsymbol{\tau}_0$, but we use a simple method for the analyses presented later. We then simulate a discrete seed distribution, which is a finite collection of realizations of hidden variables obtained after running B replicates of a Gibbs sampler for R steps, with randomly drawn initial values for each replicate. The realizations of the hidden variables $\mathbf{t}_b^{<1>}, \mathbf{t}_b^{<2>}, \ldots, \mathbf{t}_b^{<R>}$ are composed of $\mathbf{z}_{ib}^{<r>}$, indicating which group observation i belongs to, for $1 \leq i \leq n$, and the value of $x_{nb}^{<r>}$. For $1 \leq b \leq B$, $\mathbf{t}_b^{<R>} = (x_{nb}^{<R>}, \mathbf{z}_{1b}^{<R>}, \ldots, \mathbf{z}_{nb}^{<R>})$ is obtained as the Rth step in a Markov chain, the limiting distribution of which is given by $h_{\boldsymbol{\tau}_0}(\mathbf{t} \mid \mathbf{u})$.

Given our initial parameter values $\boldsymbol{\tau}_0 = (p_{01}, \delta_0, \sigma_{01}, \sigma_{02})$ and a replicate number b, the Gibbs sampler we use to approximate the target density begins by choosing a normally distributed random starting value $x_{nb}^{<0>}$. Then, at the the rth iteration, for $1 \leq r \leq R$, we define a distribution $k_{\boldsymbol{\tau}_0}(\mathbf{t} \mid x_{nb}^{<r-1>}, \mathbf{u})$ from which we sample $\mathbf{t}_b^{<r>} = (x_{nb}^{<r>}, \mathbf{z}_{1b}^{<r>}, \ldots, \mathbf{z}_{nb}^{<r>})$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

conditional on the observed data $\mathbf{u}$ and the value $x_{nb}^{<r-1>}$. This $k_{\boldsymbol{\tau}_0}(\mathbf{t} \mid x_{nb}^{<r-1>}, \mathbf{u})$ distribution, which we call the Gibbs kernel, is implicitly defined by the method for sampling $\mathbf{t}_b^{<r>}$, which is as follows:

- Sample $z_{ib1}^{<r>}$ as a Bernoulli random variable, with

$$p(z_{ib1}^{<r>} = 1) = \frac{p_{01}\phi(x_n^{<r-1>} + u_i, \mu_{01}, \sigma_{01}^2)}{\sum_{j=1}^{2} p_{0j}\phi(x_n^{<r-1>} + u_i, \mu_{0j}, \sigma_{0j}^2)},$$

  for $i = 1, \ldots, n-1$, and

$$p(z_{nb1}^{<r>} = 1 \mid x_n^{<r-1>}, \mathbf{u}) = \frac{p_{01}\phi(x_n^{<r-1>}, \mu_{01}, \sigma_1^{02})}{\sum_{j=1}^{2} p_{0j}\phi(x_n^{<r-1>}, \mu_{0j}, \sigma_{0j}^2)},$$

  where $p_{02} = 1 - p_{01}$, $\mu_{01} = 0$, $\mu_{02} = \delta_0$, and we let $z_{ib2}^{<r>} = 1 - z_{ib1}^{<r>}$.

- Sample $x_n^{<r>}$ from $N(c, d)$, where

$$\begin{aligned} d &= \left(\frac{n_1}{\sigma_{01}^2} + \frac{n_2}{\sigma_{02}^2}\right)^{-1}, \\ c &= d\left(\frac{\mu_{01}}{\sigma_{01}^2}n_1 + \frac{\mu_{02}}{\sigma_{02}^2}n_2 - \sum_{i=1}^{n} u_i\left[\frac{z_{ib1}^{<r>}}{\sigma_{01}^2} + \frac{z_{ib2}^{<r>}}{\sigma_{02}^2}\right]\right). \end{aligned}$$

Because the empirical distribution consisting of the seed values $\mathbf{t}_1^{<R>}, \ldots, \mathbf{t}_B^{<R>}$ is discrete, we propose creating from it a kernel density estimator that will serve as our seed distribution.

To this end, we let $x_{nb}^* = x_{nb}^{<R>}$, and define

$$g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t} \mid \mathbf{u}) = \frac{1}{B}\sum_{b=1}^{B} k_{\boldsymbol{\tau}_0}(\mathbf{t} \mid x_{nb}^*, \mathbf{u}),$$

where $k_{\boldsymbol{\tau}_0}(\cdot)$ is the Gibbs kernel defined above. We then run a stratified importance sam-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

pler (Owen and Zhou, 2000), as follows. For $1 \leq b \leq B$, we draw $\mathbf{t}_b = (x_{nb}, \mathbf{z}_{1b}, \ldots, \mathbf{z}_{nb})$ using the Gibbs sampling kernel $k_{\boldsymbol{\tau}_0}(\mathbf{t} \mid x^\star_{nb}, \mathbf{u})$. In the next section, we introduce an algorithm to maximize the simulated likelihood $M(\boldsymbol{\tau})$ defined in Equation (10) with respect to the invariant parameters $\boldsymbol{\tau} = \{\delta, \sigma_1, \sigma_2, p\}$.

5 An EM algorithm

Because $\log M(\boldsymbol{\tau})$ has the "log-of-sums" form discussed by Hunter et al. (2018), we may use their method to construct an iterative EM algorithm to help find a maximizer. With a denoting the iteration number of our algorithm, define

$$w_b^{<a>} = \frac{h_{\boldsymbol{\tau}^{<a>}}(\mathbf{u}, \mathbf{t}_b)}{g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t}_b \mid \mathbf{u})} \left[\sum_{j=1}^{B} \frac{h_{\boldsymbol{\tau}^{<a>}}(\mathbf{u}, \mathbf{t}_j)}{g_{\text{imp};\boldsymbol{\tau}_0}(\mathbf{t}_j \mid \mathbf{u})} \right]^{-1}$$

and

$$\mathrm{Q}(\boldsymbol{\tau} \mid \boldsymbol{\tau}^{<a>}) = \sum_{b=1}^{B} w_b^{<a>} \log h_{\boldsymbol{\tau}}(\mathbf{u}, \mathbf{t}_b).$$

As explained by Hunter et al. (2018), this definition ensures that $\mathrm{Q}(\boldsymbol{\tau} \mid \boldsymbol{\tau}^{<a>}) - \mathrm{Q}(\boldsymbol{\tau}^{<a>} \mid \boldsymbol{\tau}^{<a>})$ is a minorizer of $\log M(\boldsymbol{\tau}) - \log M(\boldsymbol{\tau}^{<a>})$. Thus, we maximize $\mathrm{Q}(\boldsymbol{\tau} \mid \boldsymbol{\tau}^{<a>})$ to get $\boldsymbol{\tau}^{<a+1>}$, which guarantees the familiar ascent property of an EM algorithm, namely, $M(\boldsymbol{\tau}^{<a+1>}) \geq M(\boldsymbol{\tau}^{<a>})$.

In the E-step, we update the importance weights $w_b^{<a>}$; the $\boldsymbol{z}_b$ are always the same, because the simulated partitions are kept fixed. This is the main difference from a standard EM algorithm. In the M-step, we maximize the previous surrogate function to obtain the estimates for the invariant model parameters. Letting $n_{bg} = \#\{i : z_{bi} = g\}$, for $g = 1, 2$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

we obtain

$$\begin{aligned}
\hat{p}^{<a+1>} &= \sum_{b=1}^{B} w_b^{<a>} \frac{n_{gb}}{n}, \\
\hat{\delta}^{<a+1>} &= \sum_{b=1}^{B} w_b^{<a>} \left[\frac{\sum_{i:z_{bi}=2}(u_i + x_{nb})}{n_{b2}} - \frac{\sum_{i:z_{bi}=1}(u_i + x_{nb})}{n_{b1}} \right], \\
\hat{\sigma}_g^{2<t+1>} &= \sum_{b=1}^{B} w_b^{<a>} \frac{\sum_{i:z_{bi}=g}(u_i + x_{nb} - \hat{\delta}^{<a>} I\{g=2\})^2}{n_{bg}}.
\end{aligned}$$

The two steps are repeated until the increase in the simulated likelihood between two consecutive steps is less than $\epsilon = 10^{-6}$.

6 An alternative DS-MLE approach

An additional novel estimation algorithm combines the importance sampling idea with a straightforward maximum likelihood estimation by selecting multiple samples of x_n. In effect, this creates an ensemble of B synthetic data sets when these values of x_{nb}, for $1 \leq b \leq B$, are added to the fixed and known $u_1, \ldots, u_{n-1}$. Because this method is based on the maximum likelihood using different samples, we refer to it as DS-MLE.

We first run 1000 k-means partitions on the original full data $x_1, \ldots, x_n$ using random starting points, keeping the one that gives the highest value of the original full-data likelihood, based on the sample estimates of the parameters p_0, μ_{01}, μ_{02}, σ_{01}^2, and σ_{02}^2. We then sample B independent values $X_{n1}, \ldots, X_{nB}$, where $X_{nb} \sim N(c, d)$, with

$$\begin{aligned}
d &= \left(\frac{n_1}{\sigma_{01}^2} + \frac{n_2}{\sigma_{02}^2} \right)^{-1}, \\
c &= d \left(\frac{\mu_{01}}{\sigma_{01}^2} n_1 + \frac{\mu_{02}}{\sigma_{02}^2} n_2 - \sum_{i=1}^{n} u_i \left[\frac{z_{i1}}{\sigma_{01}^2} + \frac{z_{i2}}{\sigma_{02}^2} \right] \right).
\end{aligned}$$

Given the B synthetic data sets consisting of the values $u_1 + x_{nb}, \ldots, u_{n-1} + x_{nb}, x_{nb}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

we may apply a standard EM algorithm to each of them. The final estimates are the means of the estimates obtained over the B runs. We obtain our estimates of δ as the difference of the estimates between μ_2 and μ_1.

7 Simulation study

In this section, we investigate the effectiveness of the proposed EM-like Monte Carlo (MC) algorithm using a simulation study. The proposed algorithm is compared to the constrained EM algorithm introduced by Ingrassia and Rocci (2007), doubly smoothed (DS) estimator proposed by Seo and Lindsay (2010), with tuning values $h = 0.01$ and $h = 0.1$, DS-MLE algorithm proposed in Section 6, and an unconstrained EM algorithm for the full likelihood, initialized using the final estimates of our main proposal (MC and Full).

7.1 Asymptotic standard errors

To derive the standard errors for the parameter estimates, we use the sandwich information matrix of Godambe (1960), $\mathbf{G} = (\mathbf{J}^{-1}\mathbf{V}\ \mathbf{J}^{-1})$, where $\mathbf{J}$ is minus the expectation of the second derivative of the log-likelihood, and $\mathbf{V}$ is the variance of the first derivative of the log likelihood (the score vector). When the model is correctly specified, $\mathbf{J}^{-1}\mathbf{V} = \mathbf{I}$ and $\mathbf{G}^{-1} = \mathbf{J}$ is the Fisher information matrix. Furthermore, as shown in Sung and Geyer (2007), $\mathbf{G}$ is also the asymptotic variance when the Monte Carlo sample size B is very large. On the other hand, we expect higher variance for the estimates when B is small, because in that case, the variability due to the stochastic method itself is nonnegligible. The $\mathbf{G}$ matrix can be used to find the standard errors of the estimates in this section and in Section 8. In the case of the simulation studies described in Section 7.2, we compare the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

asymptotically derived standard errors with the sample standard deviations of multiple estimates to assess the efficacy of the asymptotic approximations.

7.2 Simulation results

We simulated $R = 500$ samples of sizes $n = 100$ and $n = 500$ from a two-component mixture, with two sets of parameter values:

1. Model I: $p = 0.3$, $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 3$, and $\sigma_2^2 = 0.25$;

2. Model II: $p = 0.5$, $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 1$, and $\sigma_2^2 = 1$.

For each sample, we generate simulated values for x_n by applying either the algorithm in Section 4 or the algorithm in Section 6.

The constrained EM and the DS algorithms were initialized using a random partition. All algorithms were stopped when the increase in the log-likelihood was less than 10^{-6}. For algorithms that estimate the invariant parameters $\delta = \mu_2 - \mu_1$, σ_1^2, σ_2^2, and p, the estimates of μ_1 and μ_2 are obtained via a maximum likelihood estimation using a standard EM algorithm, after fixing the invariant parameters. This is a straightforward mixture problem involving only one unknown parameter. The results of these simulation studies for $n = 500$ are displayed in Figures 1 and 2; the remainders are provided in the Supplementary Material (S2).

Figures 1 and 2 and the results in the Supplementary Material (S2) show all algorithms exhibit smaller biases and mean square errors under Model I than they do under Model II. Although the proposed algorithm, labeled MC, performs well, the constrained EM algorithm appears to work better. This is not entirely surprising, because there is no simulation error or information loss. On the other hand, the MC algorithm followed by the full EM algorithm seems to improve the performance. In fact the good initialization

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

Figure 1: Box plots of 500 parameter estimates, each resulting from a sample of size $n = 500$ from Model I. The competitors are our main proposal, labeled MC, using both $B = 100$ and $B = 500$; the unconstrained EM algorithm initialized with estimates produced by MC, labeled MC and Full, again using both $B = 100$ and $B = 500$; the constrained EM algorithm of Ingrassia and Rocci (2007), labeled Constr Full; the doubly smoothed algorithm of Seo and Lindsay (2010), labeled DS, using both $h = 0.01$ and $h = 0.1$; and our alternative approach from Section 6, labeled DS-MLE.

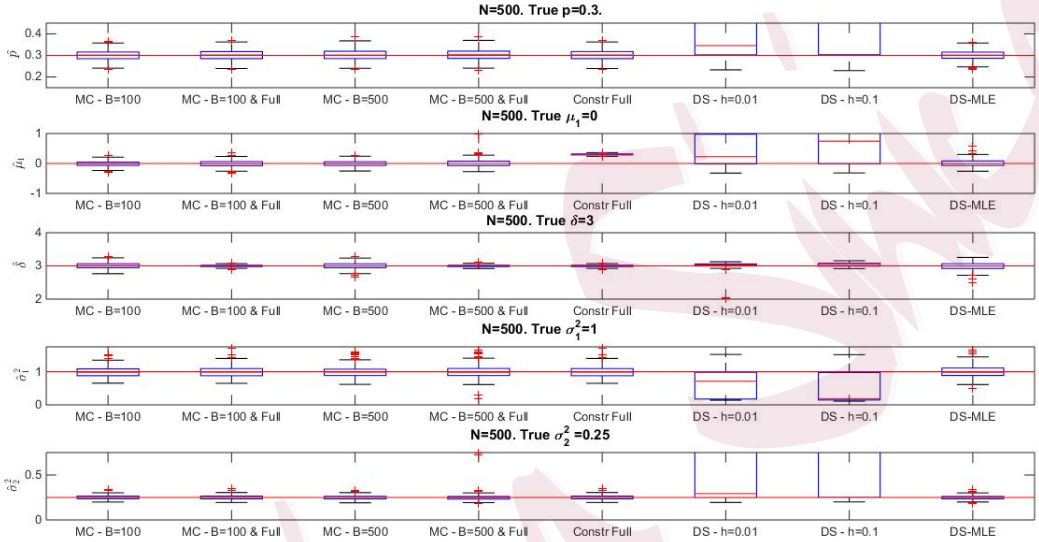
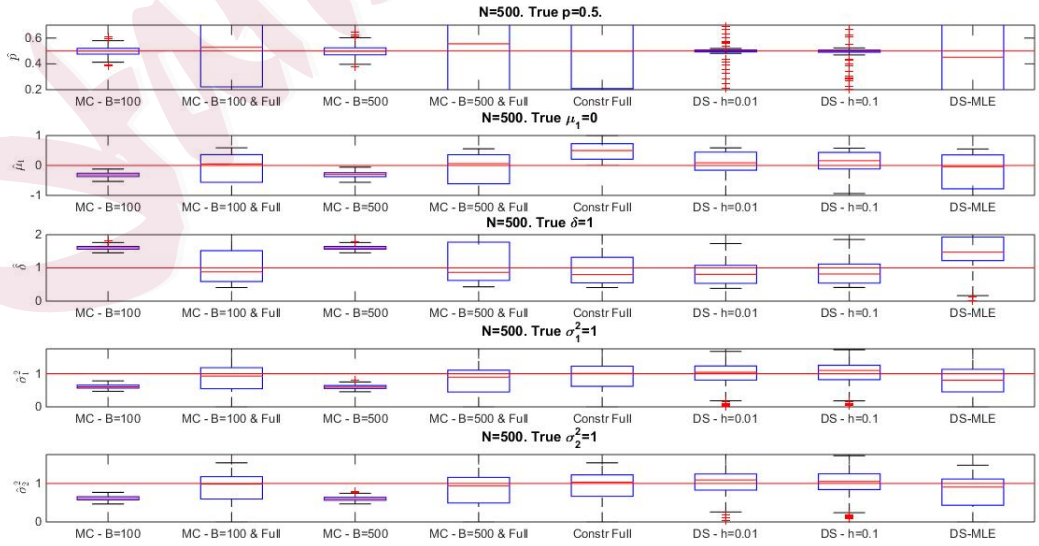


Figure 2: Box plots of 500 parameter estimates, each resulting from a sample of size $n = 500$ from Model II. The algorithm labels are explained in Figure 1.



Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

of the EM algorithm—the parameter estimates of the MC algorithm—prevents the local maximum problem, as well as the problem of an unbounded likelihood, because the parameter estimates lie in the middle of the parameter space. The DS-MLE approach shows the poorest performance for Model I, although the median estimates are still relatively close to the true parameter values, but it works very well for Model II. The DS algorithm seems to work quite well, although its parameter estimates show high mean square errors. Moreover, as the tuning parameter h increases, that is, as the parameter estimates are farther from the boundary of the parameter space, the bias, and therefore the mean squared error, increases significantly.

Comparing the sample standard deviations with the standard errors estimated using the asymptotic method described in Section 7.1, we see in the case of Model I that the sandwich formula performs quite comparably to the empirical standard deviation, except in certain cases. In these cases, our MC method occasionally produces a strong outlier, inflating the sample standard deviation. On the other hand, the doubly smoothed method appears to produce asymptotic standard errors that are consistently too small. Under Model II, in which there is a significant overlap between the mixture components, the asymptotic formula tends to underestimate the sample standard deviation more consistently and more dramatically, for all of the methods we tested, than for Model I.

8 Acidity index of lakes

In this section, we apply our proposal to a data set on an acidity index measured in a sample of 155 lakes in the Northeastern United States (Crawford et al., 1992). Previously, this data set has been modeled by a mixture of Gaussian distributions on the log scale, with a number of components from two to five (e.g., Richardson and Green, 1997; McLachlan and Peel, 2000). The authors reported the most plausible solution among all local maxima,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

removing infinite spikes and spurious maxima. Here, we choose the two-component normal mixture model to simplify our main point, as in Seo and Lindsay (2010). We compare the performance of our proposal with the approaches proposed by Ingrassia and Rocci (2007), Seo and Lindsay (2010), the algorithm proposed in Section 6, and the unconstrained EM algorithm initialized using the best solution of our proposal. Table 1 shows the local maximizers of the usual log-likelihood, based on 100 randomly generated partitions. Under our proposal, the full set of parameters is obtained by combining the marginal likelihood with the conditional likelihood; that is, given the invariant parameter estimates, we obtain the mean for the *reference* component, set equal to zero. For the best solution under the above-mentioned approaches, the asymptotic standard errors are estimated. All methods produce parameter estimates that appear quite accurate in terms of precision; that is, their standard error estimates are fairly small. However, as we saw in Section 7.2, there may be reason to doubt the standard errors in the case of the doubly smoothed estimates.

The constrained EM algorithm works well, reducing the number of local maxima. This is also true for the DS-MLE algorithm presented in Section 6, and the unconstrained EM algorithms initialized with the best solutions of our proposal. All of these methods reach the same local maximum. For the approach of Seo and Lindsay (2010), as expected, a lower h (0.01) produces a higher number of local maxima. Indeed, the doubly smoothed method with small h is similar to the unconstrained case, in which the likelihood is known to have spikes to infinity.

9 Concluding Remarks

This study provides a solution for the unbounded likelihood issue in a normal mixture problem. Currently, the boundedness has only been proved for the univariate two-component case, though with further theoretical development, perhaps these ideas hold

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

Table 1: Local maximizers from the constrained EM algorithm, doubly smoothed log-likelihood, and our proposal. Asymptotic standard errors for the best solutions are reported in parentheses.

Constrained EM algorithm (Ingrassia and Rocci, 2007)					
μ_1	μ_2	σ_1^2	σ_2^2	p	Log-Lik
4.2506	5.8916	0.0678	0.7180	0.4793	-187.2345
4.3301	6.2491	0.1388	0.2701	0.5962	-184.6447
(0.0415)	(0.0720)	(0.0340)	(0.0457)	(0.0408)	
Doubly smoothed likelihood (Seo and Lindsay, 2010) with $h = 0.01$					
μ_1	μ_2	σ_1^2	σ_2^2	p	Log-Lik
5.1431	5.0555	1.0457	1.1025	0.4506	-226.5162
5.9807	4.2635	0.0834	0.6119	0.4990	-191.1817
5.9817	4.2637	0.0837	0.6106	0.4987	-191.1815
4.2629	5.9620	0.0780	0.6422	0.4972	-191.1669
4.2589	5.9363	0.0740	0.6727	0.4887	-191.1640
4.3315	6.2542	0.1391	0.2670	0.5941	-187.9188
(0.0416)	(0.0719)	(0.0340)	(0.0450)	(0.0408)	
Doubly smoothed likelihood (Seo and Lindsay, 2010) with $h = 0.1$					
μ_1	μ_2	σ_1^2	σ_2^2	p	Log-Lik
5.1060	5.1051	1.1768	1.1775	0.5352	-232.5730
5.1607	5.0589	1.1417	1.2135	0.5412	-232.5552
5.1786	5.0433	1.1313	1.2209	0.5385	-232.5408
6.0485	4.2506	0.1963	0.5638	0.4700	-212.6224
4.3475	6.2858	0.2513	0.3316	0.6008	-210.7553
(0.0436)	(0.0700)	(0.2791)	(0.1402)	(0.0416)	
Our MC method with $B = 500$					
μ_1	μ_2	σ_1^2	σ_2^2	p	Log-Lik
4.9213	5.0896	1.3838	0.3147	0.7100	-238.9258
4.6522	5.4254	0.8937	1.2140	0.4096	-224.7499
4.5808	5.5108	0.8444	1.1755	0.4286	-224.1096
4.3570	6.3050	0.1734	0.2146	0.6233	-185.2983
(0.0438)	(0.0681)	(0.0447)	(0.0332)	(0.0400)	
Full EM initialized with the best MC solution					
μ_1	μ_2	σ_1^2	σ_2^2	p	Log-Lik
4.3302	6.2493	0.1389	0.2699	0.5962	-184.6447
(0.0416)	(0.0720)	(0.0341)	(0.0456)	(0.0408)	
Alternative DS MLE of Section 6 with $B = 10$					
μ_1	μ_2	σ_1^2	σ_2^2	p	Log-Lik
4.3302	6.2494	0.1390	0.2698	0.5963	-184.6447
(0.0416)	(0.0720)	(0.0341)	(0.0456)	(0.0408)	
Alternative DS MLE of Section 6 with $B = 100$					
μ_1	μ_2	σ_1^2	σ_2^2	p	Log-Lik
4.3302	6.2492	0.1389	0.2700	0.5962	-184.6447
(0.0416)	(0.0720)	(0.0341)	(0.0457)	(0.0408)	

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

promise for more general cases. For instance, when the multivariate invariant likelihood may be written as the product of the univariate invariant likelihoods, an extension to the multivariate cases is straightforward. In order to decompose the multivariate density as a product of univariate densities, we restrict the covariance matrices $\mathbf{\Sigma}_g$ to be diagonal. This particular model belongs to a wider class, called parsimonious Gaussian mixture models, introduced by Celeux and Govaert (1995).

It may be possible to investigate a large group of transformations, the so-called affine transformations, generated by $\mathbf{x} \to \mathbf{a} + \mathbf{B}\mathbf{x} = \mathbf{y}$, where $\mathbf{B}$ is an arbitrary nonsingular matrix. The corresponding transformation of the parameters gives

$$\boldsymbol{\tau} \to (p, \mathbf{a} + \mathbf{B}\boldsymbol{\mu}_1, a + \mathbf{B}\boldsymbol{\mu}_2, \mathbf{B}\mathbf{\Sigma}_1\mathbf{B}^T, \mathbf{B}\mathbf{\Sigma}_2\mathbf{B}^T).$$

It follows that, in this case, a set of maximally invariant statistics is

$$(\mathbf{S}^{-1/2}(\mathbf{x}_1 - \mathbf{x}_n), ..., \mathbf{S}^{-1/2}(\mathbf{x}_{n-1} - \mathbf{x}_n)).$$

This larger group reduces the active parameters still further. If we take a fixed $\boldsymbol{\tau}_0$, and apply this family of transformations to it, we get a collection of parameter values called the orbit of $\boldsymbol{\tau}_0$, Orb($\boldsymbol{\tau}_0$). The orbits are the contours of the maximal invariant parameter. The maximal invariant statistics have the same invariant-statistics-based likelihood for every $\boldsymbol{\tau} \in$ Orb($\boldsymbol{\tau}_0$). For example, the transformation $\mathbf{x} \to \mathbf{\Sigma}_1^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_1)$ transforms observations from the first component to the standard normal, $N(\mathbf{0}, \mathbf{I})$, while the second component becomes $N(\mathbf{\Sigma}_1^{-1/2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2})$. We can fit the invariant likelihood based on this parametrization, recognizing that we must use additional equations to solve directly for $\boldsymbol{\mu}_1$ and $\mathbf{\Sigma}_1$. It might be possible to show that the invariant likelihood is bounded for $d > 1$ by constrained solutions fulfilling $\det(\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}) \geq 1$; imposing this constraint

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

also solves the labeling nonidentifiability issue.

Finally, note that the algorithms proposed here can be adapted easily to the multivariate case, despite the problem of the unbounded likelihood.

Supplementary Material

The online Supplementary Material includes the proof of Lemma 1 (S1) and additional results from the simulation studies (S2).

Acknowledgements. During the preparation of this manuscript, Bruce G. Lindsay passed away due to an illness. We lost a brilliant statistician, a warm and humble human being, and a great friend. We miss him dearly. The first author is particularly grateful to Professor Lindsay, a wise and excellent mentor whose contribution to her early-stage research has been invaluable. The author will treasure it forever.

References

Broniatowski, M., G. Celeux, and J. Diebolt (1983). Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data analysis and informatics 3*, 359–373.

Celeux, G., D. Chauveau, and J. Diebolt (1995). On Stochastic Versions of the EM Algorithm. Research Report RR-2514, INRIA.

Celeux, G. and J. Diebolt (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational statistics quarterly 2*(1), 73–82.

Celeux, G. and J. Diebolt (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastic Reports 41*(1-2), 119–134.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition 28*(5), 781 – 793.

Chen, J. and X. Tan (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis 100*(7), 1367–1383.

Chen, J., X. Tan, and R. Zhang (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica 18*(2), 443.

Ciuperca, G., A. Ridolfi, and J. Idier (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics 30*(1), 45–59.

Cox, D. and D. Hinkley (1979). *Theoretical Statistics*. Taylor & Francis.

Crawford, S. L., M. H. DeGroot, J. B. Kadane, and M. J. Small (1992). Modeling lake-chemistry distributions: Approximate bayesian methods for estimating a finite-mixture model. *Technometrics 34*(4), 441–453.

Geyer, C. J. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B 56*(1), pp. 261–274.

Geyer, C. J. and E. A. Thompson (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B 54*(3), pp. 657–699.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics 31*(4), 1208–1211.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics 13*, 795–800.

Hathaway, R. J. (1986). A constrained em algorithm for univariate normal mixtures. *Journal of Statistical Computation and Simulation 23*(3), 211–230.

Hummel, R. M., D. R. Hunter, and M. S. H. Handcock (2012). Improving simulation-based algorithms for fitting ERGMs. *Journal of Computational and Graphical Statistics 21*(4), 920–939.

Hunter, D. R., P. Kuruppumullage Don, and B. G. Lindsay (2018). An expansive view of EM algorithms. In

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0483

Handbook of Mixture Analysis, pp. 41–54. CRC Press.

Ingrassia, S. and R. Rocci (2007). Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Computational Statistics & Data Analysis 51*(11), 5339–5351.

Ingrassia, S. and R. Rocci (2011). Degeneracy of the em algorithm for the mle of multivariate gaussian mixtures and dynamic constraints. *Computational statistics & data analysis 55*(4), 1715–1725.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics 27*(4), 887–906.

Kiefer, N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica 46*(2), 427–434.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models* (1 ed.). Wiley Series in Probability and Statistics. Wiley-Interscience.

Owen, A. and Y. Zhou (2000). Safe and effective importance sampling. *Journal of the American Statistical Association 95*(449), 135–143.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A 185*, 71–110.

Peters, Jr, B. C. and H. F. Walker (1978). An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal on Applied Mathematics 35*(2), 362–378.

Policello II, G. E. (1981). Conditional maximum likelihood estimation in gaussian mixtures. In *Statistical Distributions in Scientific Work*, pp. 111–125. Springer.

Richardson, S. and P. J. Green (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology) 59*(4), 731–792.

Seo, B. and B. G. Lindsay (2010). A computational strategy for doubly smoothed mle exemplified in the normal mixture model. *Computational Statistics & Data Analysis 54*(8), 1930–1941.

Sung, Y. J. and C. J. Geyer (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics 35*(3), 990–1011.

Tanaka, K. and A. Takemura (2006). Strong consistency of the maximum likelihood estimator for finite mixtures of location: Scale distributions when the scale parameters are exponentially small. *Bernoulli 12*(6), 1003–1017.

Tanner, M. A. (1991). *Tools for Statistical Inference*, Volume 3. Springer.

Wei, G. C. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*(411), 699–704.

Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference 140*(7), 2089–2098.

E-mail: mxr459@psu.edu

Department of Statistics, Pennsylvania State University

E-mail: dhunter@psu.edu