Journal of Statistical Planning and Inference 140 (2010) 2089–2098

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A profile likelihood method for normal mixture with unequal variance

Weixin Yao

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

ARTICLE INFO

Article history:
Received 4 September 2009
Received in revised form
2 February 2010
Accepted 3 February 2010
Available online 10 February 2010

Keywords:
EM algorithm
Maximum likelihood
Mixture models
Profile likelihood
Unbounded likelihood

ABSTRACT

It is well known that the normal mixture with unequal variance has unbounded likelihood and thus the corresponding global maximum likelihood estimator (MLE) is undefined. One of the commonly used solutions is to put a constraint on the parameter space so that the likelihood is bounded and then one can run the EM algorithm on this constrained parameter space to find the constrained global MLE. However, choosing the constraint parameter is a difficult issue and in many cases different choices may give different constrained global MLE. In this article, we propose a profile log likelihood method and a graphical way to find the maximum interior mode. Based on our proposed method, we can also see how the constraint parameter, used in the constrained EM algorithm, affects the constrained global MLE. Using two simulation examples and a real data application, we demonstrate the success of our new method in solving the unboundness of the mixture likelihood and locating the maximum interior mode.

1. Introduction

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

$$f(x;\boldsymbol{\theta})=\pi_1\phi(x;\mu_1,\sigma_1^2)+\pi_2\phi(x;\mu_2,\sigma_2^2)+\cdots+\pi_m\phi(x;\mu_m,\sigma_m^2),$$

where $\boldsymbol{\theta}=(\pi_1,\ldots,\pi_{m-1},\mu_1,\ldots,\mu_m,\sigma_1,\ldots,\sigma_m)$, $\phi(\cdot;\mu,\sigma^2)$ is the normal density with mean μ and σ^2, and π_j is the proportion of j th component with $\sum_{j=1}^{m}\pi_j=1$. The log-likelihood for $\mathbf{x}$ is

$$\log L(\boldsymbol{\theta};\mathbf{x})=\sum_{i=1}^{n}\log\{\pi_1\phi(x_i;\mu_1,\sigma_1^2)+\pi_2\phi(x_i;\mu_2,\sigma_2^2)+\cdots+\pi_m\phi(x_i;\mu_m,\sigma_m^2)\}. \quad (1)$$

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

It is well known that $\log L(\boldsymbol{\theta};\mathbf{x})$ in (1) is unbounded without any restriction on the component variance, and so the global maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, by maximizing (1), does not exist. For example, if we set $\mu_1=x_1$ and let $\sigma_1^2\to 0$, the likelihood value goes to infinity. However, for mixtures of normal distributions, at least in the univariate case, there is a sequence of roots corresponding to local maxima in the interior of the parameter space that are consistent and asymptotically normal and efficient (Kiefer, 1978; Peters and Walker, 1978). Note that if there are multiple local maxima in the interior of the parameter space, there is also a problem of identifying the consistent sequence, which is a very difficult problem itself. In this article, we do not focus on this issue. Instead, when the likelihood is unbounded, we define the MLE as the maximum interior/local mode. Hathaway (1985) provided some theoretical support of using the maximum interior/local mode.

E-mail address: wxyao@ksu.edu

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.02.004

One of the commonly used methods to avoid the unboundness of the log likelihood and to find the maximum interior mode is to run the EM algorithm (Dempster et al., 1977) over a constrained parameter space

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

where $C \in (0, 1]$, and Ω denotes the unconstrained parameter space. See Hathaway (1985, 1986) and Bezdek et al. (1985) for more detail. However, a big challenge for this method is to choose the appropriate cut point C. If C is too large, it is possible that the consistent local maxima does not belong to the constrained parameter space Ω_C and thus the found estimate will be misleading. Even the consistent local maxima is in Ω_C, it is still possible that Ω_C misses some interior modes worthy of consideration. On the other hand, if C is too small, it is possible that some boundary point, satisfying $\sigma_h/\sigma_j = C$ for some h and j, maximizes the log likelihood over the constrained parameter space Ω_C. In this situation, the found estimate is on the boundary of Ω_C and thus depends on the choice of C.

Another commonly used method is to use maximum penalized likelihood estimator that adds penalty term to the unequal variance. See Chen et al. (2008) and Chen and Tan (2009).

In this article, we propose a profile log-likelihood method and a graphical way to solve the unboundness issue of likelihood and find the maximum interior mode for the normal mixture with unequal variance. Unlike the constrained EM algorithm (Hathaway, 1985, 1986), our proposed method does not need to specify a cut point C. In addition, based on our proposed method, we can clearly check whether there are some other minor interior modes and see how the choice of C in (2) affects the constrained global MLE. Using the simulation study and a real data application, we demonstrate the effectiveness of our proposed method and show how the selection of cut point C affects the constrained MLE (Hathaway, 1985, 1986).

The rest of the paper is organized as follows. Section 2 proposes a profile log likelihood method to solve the unboundness issue of the likelihood function for the normal mixture with unequal variance. In Section 3, we use two simulation examples and a real data application to demonstrate how our proposed method works. We summarize our proposed method and give the discussion in Section 4.

2. New method

In this section, we will first introduce our profile log likelihood method for two component normal mixtures and provide a simple EM algorithm. We will then extend the profile log likelihood method to normal mixtures of more than two components.

2.1. Mixtures of two components

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ from the two-component normal mixture, the log-likelihood for $\mathbf{x}$ is

$$\log L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \log\{\pi_1 \phi(x_i; \mu_1, \sigma_1^2) + \pi_2 \phi(x_i; \mu_2, \sigma_2^2)\}, \tag{3}$$

where $\boldsymbol{\theta} = (\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$ and

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

Note that without any restriction, the above log-likelihood is unbounded and the global MLE is undefined. In this section, we propose a profile likelihood method to avoid the unboundness issue and to find the maximum interior mode of $\log L(\boldsymbol{\theta}; \mathbf{x})$.

Let $\sigma_1 = k\sigma_2 \equiv k\sigma$, where $k \in (0, 1]$. Then the log-likelihood of (3), for each fixed k, is

$$\log L(\boldsymbol{\eta}; \mathbf{x}, k) = \sum_{i=1}^{n} \log\{\pi_1 \phi(x_i; \mu_1, k^2\sigma^2) + \pi_2 \phi(x_i; \mu_2, \sigma^2)\}. \tag{4}$$

where $\boldsymbol{\eta} = (\pi_1, \mu_1, \mu_2, \sigma)$. *Note that for each fixed k, the log-likelihood of* (4) *is bounded.* Hence the global MLE for (4) is well defined. In order to estimate k, we define the profile log-likelihood for k as

$$p(k) = \max_{\boldsymbol{\eta}} \log L(\boldsymbol{\eta}; \mathbf{x}, k), \tag{5}$$

where $\log L(\boldsymbol{\eta}; \mathbf{x}, k)$ is defined in (4).

Let

$$\Omega_C = \{\boldsymbol{\theta} \in \Omega : \min(\sigma_1, \sigma_2)/\max(\sigma_1, \sigma_2) \geq C > 0\}, \tag{6}$$

where Ω is the unconstrained parameter space for $\boldsymbol{\theta}$.

Theorem 2.1. *We have the following properties about the profile likelihood* $p(k)$ *defined in* (5).

(a) *The profile likelihood* $p(k)$ *is unbounded and goes to infinity when k goes to zero.*
(b) *The* $\hat{\boldsymbol{\theta}} = (\hat{\pi}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ *maximizes the log likelihood* $\log L(\boldsymbol{\theta}; \mathbf{x})$ *of* (3) *constrained in* Ω_C, *where* $\hat{\sigma}_1 \leq \hat{\sigma}_2$, *if and only if* $\hat{k} = \hat{\sigma}_1/\hat{\sigma}_2$ *maximizes the profile log-likelihood* $p(k)$ *of* (5) *in* K_C, *where* $K_C = \{k \in (0, 1] : k \geq C\}$.

(c) *Suppose* $\tilde{k}$ *is a local mode for the profile log-likelihood* $p(k)$ *with the corresponding* $\tilde{\boldsymbol{\eta}} = (\tilde{\pi}_1, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma})$. *Let* $\tilde{\boldsymbol{\theta}} = (\tilde{\pi}_1, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{k}\tilde{\sigma}, \tilde{\sigma})$. *Then* $\tilde{\boldsymbol{\theta}}$ *is a local mode for the log likelihood* $\log L(\boldsymbol{\theta}; \mathbf{x})$ *of* (3).

The proof of Theorem 2.1 is given in the Appendix. From (a), one can know that $p(k)$ is also unbounded. Therefore, we cannot estimate k by maximizing $p(k)$ directly. Based on (b), one can know that finding the maximum interior mode of $\log L(\boldsymbol{\theta}; \mathbf{x})$ of (3) is equivalent to finding the maximum interior mode of $p(k)$. Noting that k is a one-dimensional parameter, hence our profile likelihood method transfers the problem of locating the maximum interior mode for a high-dimensional function $\log L(\boldsymbol{\theta}; \mathbf{x})$ into locating the maximum interior model for a one-dimensional function $p(k)$.

For one dimension function $p(k)$, one can easily use the plot of $p(k)$ versus k to locate the maximum interior mode of $p(k)$ without choosing a cut point C in advance, which is one of the major advantages of our proposed method and will be illustrated in more detail in Section 3. Let $\hat{k}$ be the maximum interior mode of (5). Then fixing k at $\hat{k}$, we can find the MLE of (4), denoted by $\hat{\boldsymbol{\eta}}(\hat{k})$, and the corresponding $\hat{\boldsymbol{\theta}}(\hat{k})$. The $\hat{\boldsymbol{\theta}}(\hat{k})$ is our proposed maximum interior mode of (3).

Based on the plot of $p(k)$ versus k, one can also clearly see how the cut point C in (6) affects the constrained MLE (Hathaway, 1985, 1986). We will demonstrate this using examples in Section 3.

Note that the profile log-likelihood $p(k)$ does not have an explicit form. Therefore, we can only numerically evaluate $p(k)$ for a set of grid points of k. The following is the EM algorithm to find $p(k)$ for any fixed k.

Algorithm 1. Starting with the initial parameter values $\{\hat{\pi}_1^{(0)}, \hat{\mu}_1^{(0)}, \hat{\mu}_2^{(0)}, \hat{\sigma}_1^{(0)} = k\hat{\sigma}_2^{(0)}\}$, iterate the following two steps until convergence.

E Step: Compute the classification probabilities:

$$\hat{p}_{ij}^{(t+1)} = \frac{\hat{\pi}_j^{(t)} \phi(x_i; \hat{\mu}_j^{(t)}, \hat{\sigma}_j^{2(t)})}{\sum_{l=1}^{2} \hat{\pi}_l^{(t)} \phi(x_i; \hat{\mu}_l^{(t)}, \hat{\sigma}_l^{2(t)})}, \quad i = 1, \ldots, n, \ j = 1, 2$$

M step: Update the component parameters:

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{p}_{ij}^{(t+1)} x_i}{\sum_{i=1}^{n} \hat{p}_{ij}^{(t+1)}}, \quad \hat{\pi}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{p}_{ij}^{(t+1)}}{n}, \ j = 1, 2,$$

$$\hat{\sigma}_1^{2(t+1)} = \frac{\sum_{i=1}^{n} [\hat{p}_{i1}^{(t+1)} (x_i - \hat{\mu}_1^{(t+1)})^2 + k^2 \hat{p}_{i2}^{(t+1)} (x_i - \hat{\mu}_2^{(t+1)})^2]}{n}, \hat{\sigma}_2^{(t+1)} = \hat{\sigma}_1^{(t+1)}/k.$$

Similar to the general EM-algorithm, this algorithm is only guaranteed to converge to a local mode. In order to find the maximal mode (global MLE) for each fixed k, we may run the algorithm from several initial values and choose the converged mode which has the largest log-likelihood (note that the maximal mode is well defined since the log likelihood (4) is bounded for each fixed k).

2.2. Mixtures of more than two components

When there are more than two components, i.e. $m > 2$, let $k = \sigma_{(1)}/\sigma_{(m)}$, where $\sigma_{(1)} \le \sigma_{(2)} \le \cdots \le \sigma_{(m)}$ are ordered sequence of $(\sigma_1, \ldots, \sigma_m)$. Let

$$\Theta_k = \{\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{m-1}, \mu_1, \ldots, \mu_m, \sigma_1, \ldots, \sigma_m) | \sigma_{(1)} = k\sigma_{(m)}\}.$$

Then one can define the profile log likelihood as

$$p(k) = \max_{\boldsymbol{\theta} \in \Theta_k} \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}, k), \quad k \in (0, 1]. \tag{7}$$

It can be easily seen that the above defined profile log likelihood $p(k)$ also has the properties given in Theorem 2.1. In addition, similar to the way proposed in Section 2.1, one can also use $p(k)$ in (7) to find the maximum interior mode and check how the constraint parameter affects the constrained MLE for the constrained EM algorithm.

Due to the complicated nature of the constrained optimization, finding $p(k)$ is not trivial for each fixed k. In $(t+1)$th step of EM algorithm, E step finds the classification probabilities

$$\hat{p}_{ij}^{(t+1)} = \frac{\hat{\pi}_j^{(t)} \phi(x_i; \hat{\mu}_j^{(t)}, \hat{\sigma}_j^{2(t)})}{\sum_{l=1}^{m} \hat{\pi}_l^{(t)} \phi(x_i; \hat{\mu}_l^{(t)}, \hat{\sigma}_l^{2(t)})}, \quad i = 1, \ldots, n, \ j = 1, \ldots, m.$$

In M step, the component means and the mixing proportions are updated by

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{p}_{ij}^{(t+1)} x_i}{\sum_{i=1}^{n} \hat{p}_{ij}^{(t+1)}}, \quad \hat{\pi}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{p}_{ij}^{(t+1)}}{n}, \ j = 1, \ldots, m.$$

Let $n_j=\sum_{i=1}^n \hat{p}_{ij}^{(t+1)}$ and $S_j^2=\sum_{i=1}^n \hat{p}_{ij}^{(t+1)}(x_i-\mu_j^{(t+1)})^2$. For simplicity of notation, we omit the dependence of n_j and S_j on $t+1$. For a fixed $k\in(0,1]$, based on the EM algorithm theory, $\hat{\boldsymbol{\sigma}}^{(t+1)}=(\hat{\sigma}_1^{(t+1)},\ldots,\hat{\sigma}_m^{(t+1)})$ are updated by minimizing

$$\sum_{j=1}^m\left(n_j\log\sigma_j+\frac{S_j^2}{2\sigma_j^2}\right), \tag{8}$$

subject to $\sigma_{(1)}=k\sigma_{(m)}$.

Note that due to the label switching issue of mixture models (Yao and Lindsay, 2009), the component index does not have real meaning. Without loss of generality, we will assume that the component index satisfies $S_1^2/n_1\le S_2^2/n_2\le\cdots\le S_m^2/n_m$. (If the component index does not satisfy the above constraint, we can always permute the component index such that the above constraint holds.)

Note that when $k=1$, the component variance are all equal and thus the computation of p(1) is straightforward. In the following, we will mainly consider the situation when $0<k<1$.

Proposition 2.1. *Let $\hat{\boldsymbol{\sigma}}^{(t+1)}=(\hat{\sigma}_1^{(t+1)},\ldots,\hat{\sigma}_m^{(t+1)})$ be the maximizer of (8), subject to $\sigma_{(1)}=k\sigma_{(m)}$, where $k\in(0,1)$. Let $(\hat{\sigma}_{(1)}^{(t+1)},\ldots,\hat{\sigma}_{(m)}^{(t+1)})$ be the corresponding ordered sequence. Then, we have the following results about $\hat{\boldsymbol{\sigma}}^{(t+1)}$.*

(a) *If $S_1^2/n_1\le k^2S_m^2/n_m$, there exists $1\le i<j\le m$ such that $\hat{\sigma}_1^{(t+1)}=\hat{\sigma}_2^{(t+1)}=\cdots=\hat{\sigma}_i^{(t+1)}\le S_{i+1}/\sqrt{n_{i+1}}$, $\hat{\sigma}_j^{(t+1)}=\hat{\sigma}_{j+1}^{(t+1)}=\cdots=\hat{\sigma}_m^{(t+1)}\ge S_{j-1}/\sqrt{n_{j-1}}$, and $\hat{\sigma}_l^{(t+1)}=S_l/\sqrt{n_l}$, $l=i+1,\ldots,j-1$.*
(b) *If $S_1^2/n_1>k^2S_m^2/n_m$, there exists $1\le i<j\le m$ such that $\hat{\sigma}_i^{(t+1)}=\hat{\sigma}_{(1)}^{(t+1)}\le S_1/\sqrt{n_1}$, $\hat{\sigma}_j^{(t+1)}=\hat{\sigma}_{(m)}^{(t+1)}\ge S_m/\sqrt{n}_m$, and $\hat{\sigma}_l^{(t+1)}=S_l/\sqrt{n}_l$, $l\ne i$ and $l\ne j$.*

The proof of Proposition 2.1 is given in the Appendix. From the Proposition 2.1, we can see that the constrained maximizer of (8) depends on whether $S_1^2/n_1<k^2S_m^2/n_m$ holds. When $S_1^2/n_1\le k^2S_m^2/n_m$, $\hat{\sigma}_l^{(t+1)}=\hat{\sigma}_{(l)}^{(t+1)}$, $l=1,\ldots,m$. (Note that we have assumed $S_1^2/n_1\le S_2^2/n_2\le\cdots\le S_m^2/n_m$.) However, when $S_1^2/n_1>k^2S_m^2/n_m$, $\hat{\sigma}_{(1)}^{(t+1)}$ is not necessarily equal to $\hat{\sigma}_1^{(t+1)}$ and $\hat{\sigma}_{(m)}^{(t+1)}$ is not necessarily equal to $\hat{\sigma}_m^{(t+1)}$.

Proposition 2.2. (a) *For any $1\le i<j\le m$,under the constraint that $\sigma_1=\sigma_2=\cdots=\sigma_i=\sigma$ and $\sigma_j=\sigma_{j+1}=\cdots=\sigma_m=\sigma/k$, the objective function (8), as a function of σ by fixing $\{\sigma_{i+1},\ldots,\sigma_{j-1}\}$, is minimized at*

$$\hat{\sigma}_{(i,j)}^2=\frac{\sum_{l=1}^i S_l^2+k^2\sum_{l=j}^m S_l^2}{\sum_{l=1}^i n_l+\sum_{l=j}^m n_l}. \tag{9}$$

In addition, (8) is monotone decreasing when $\sigma<\hat{\sigma}_{(i,j)}$ and monotone increasing when $\sigma>\hat{\sigma}_{(i,j)}$.

(b) *For any $1\le i<j\le m$, under the constraint that $\sigma_i=\sigma=k\sigma_j$, the objective function (8), as a function of σ by fixing $\{\sigma_l, l\ne i$ and $l\ne j\}$, is minimized at*

$$\breve{\sigma}_{(i,j)}^2=\frac{S_i^2+k^2S_j^2}{n_i+n_j}. \tag{10}$$

In addition, (8) is monotone decreasing when $\sigma<\breve{\sigma}_{(i,j)}$ and monotone increasing when $\sigma>\breve{\sigma}_{(i,j)}$.

The proof of Proposition 2.2 is given in the Appendix. Based on the Propositions 2.1 and 2.2, we propose to use the following two steps to find $\hat{\boldsymbol{\sigma}}^{(t+1)}$ that minimizes (8) subject to $\sigma_{(1)}=k\sigma_{(m)}$.

Step 1: If $S_1^2/n_1\le k^2S_m^2/n_m$, for all pairs $1\le i<j\le m$, let $\tilde{\boldsymbol{\sigma}}_{(i,j)}$ be the minimizer of (8) under the constraint $\tilde{\sigma}_1=\tilde{\sigma}_2=\cdots=\tilde{\sigma}_i$, $\tilde{\sigma}_j=\tilde{\sigma}_{j+1}=\cdots=\tilde{\sigma}_m=\tilde{\sigma}_1/k$, $\tilde{\sigma}_1^2\le S_{i+1}^2/n_{i+1}$, and $\tilde{\sigma}_m^2\ge S_{j-1}^2/n_{j-1}$, when $\{\tilde{\sigma}_l^2=S_l^2/n_l, l=i+1,\ldots,j-1\}$ are fixed, where

$$\tilde{\sigma}_1^2=\begin{cases}\hat{\sigma}_{(i,j)}^2, & k^2S_{j-1}^2/n_{j-1}\le\hat{\sigma}_{(i,j)}^2\le S_{i+1}^2/n_{i+1},\\ S_{i+1}^2/n_{i+1}, & \hat{\sigma}_{(i,j)}^2>S_{i+1}^2/n_{i+1},\\ k^2S_{j-1}^2/n_{j-1}, & \hat{\sigma}_{(i,j)}^2<k^2S_{j-1}^2/n_{j-1},\end{cases}$$

where $\hat{\sigma}_{(i,j)}$ is defined in (9).

If $S_1^2/n_1>k^2S_m^2/n_m$, for all pairs $1\le i<j\le m$, let $\tilde{\boldsymbol{\sigma}}_{(i,j)}$ be the minimizer of (8) under the constraint $\tilde{\sigma}_i=k\tilde{\sigma}_j$ and $\tilde{\sigma}_i^2\le S_1^2/n$ and $\tilde{\sigma}_j^2\ge S_m^2/n_m$, when $\{\tilde{\sigma}_l^2=S_l^2/n_l, l\ne i$ and $l\ne j\}$ are fixed, where

$$\tilde{\sigma}_i^2=\begin{cases}\breve{\sigma}_{(i,j)}^2, & k^2S_m^2/n_m\le\breve{\sigma}_{(i,j)}^2\le S_1^2/n_1,\\ S_1^2/n_1, & \breve{\sigma}_{(i,j)}^2>S_1^2/n_1,\\ k^2S_m^2/n_m, & \breve{\sigma}_{(i,j)}^2<k^2S_m^2/n_m,\end{cases}$$

where $\breve{\sigma}_{(i,j)}$ is defined in (10).

Step 2: Let $(\tilde{i},\tilde{j})$ be the index of (i,j) such that $\tilde{\boldsymbol{\sigma}}_{(\tilde{i},\tilde{j})}$ minimizes (8) among $\tilde{\boldsymbol{\sigma}}_{(i,j)}$s, $1\le i<j\le m$. Then $\hat{\boldsymbol{\sigma}}^{(t+1)}=\tilde{\boldsymbol{\sigma}}_{(\tilde{i},\tilde{j})}$ minimizes (8) subject to $\sigma_{(1)}=k\sigma_{(m)}$.

By careful analysis of the properties of $\hat{\boldsymbol{\sigma}}^{(t+1)}$, one might be able to further shorten the computations of Step 1 by skipping the calculation of $\tilde{\boldsymbol{\sigma}}_{(i,j)}$'s for some (i,j). See the remarks after the proof of Proposition 2.1 in the Appendix for more detail.

3. Example

In this section, we will use two simulation examples and a real data application to show how our proposed method works. For simplicity of reporting, we mainly consider the case when $m=2$. When $m>2$, the results are similar. Algorithm

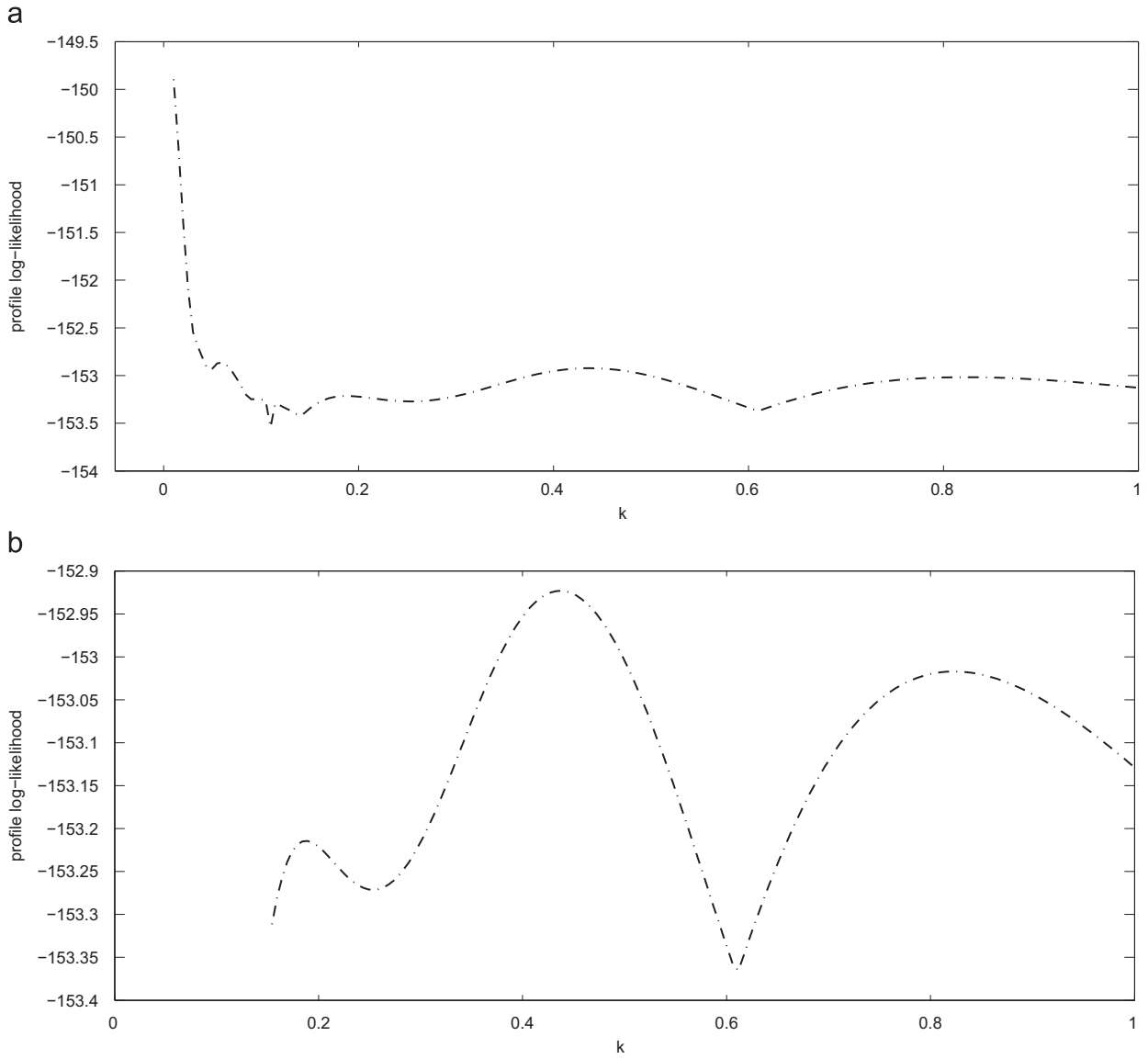

Fig. 1. Profile log-likelihood plot for Example 1: (a) for all k values from 10^{-4} to 1; (b) for k values from 0.15 to 1.

Table 1
Local maximizers for Example 1.

Local maximizer	$\log L$	π_1	μ_1	μ_2	σ_1	σ_2
k=0.1891	−153.2144	0.0934	−0.1700	0.8280	0.2175	1.1503
k=0.4378	−152.9230	0.2199	−0.0567	0.9578	0.5092	1.1629
k=0.8209	−153.0170	0.2796	2.0455	0.2260	0.6791	0.8273

1 is used to find the profile log-likelihood $p(k)$ in (5) over 200 equally spaced grid points of k from 10^{-4} to 1. Note that when k is close to zero, the smaller component variance, say σ_1^2, is also close to zero. Therefore, when k is small, the initial value for μ_1 should be one of the observations, otherwise, it is possible that there will be no observations assigned to the first component. For Algorithm 1, we used 30 initial values for each k. The initial values for mixing proportions π_1 and π_2 are both $\frac{1}{2}$. The initial values for the larger component variance σ_2^2 is half of the sample variance. The first 15 initial values for the component means are randomly sampled from the observations $(x_1,\ldots,x_n)$. For each of the sampled component means, say (x_i,x_j) for some $i\neq j$, we also used its permuted values (x_j,x_i) as the initial component means in order to avoid misspecifying the labels between component means and component variance. When k is not close to zero, one might try some other methods to choose the initial values. See McLachlan and Peel (2000, Section 2.12) and Karlis and Xekalaki (2003).

3.1. Simulation studies

Example 1. One hundred observations are generated from $0.3\mathrm{N}(0,0.5^2)+0.7\mathrm{N}(1,1)$. Fig. 1 is the profile log-likelihood plot of $p(k)$ versus k. From the plot, we can see that $p(k)$ goes to infinity when k goes to zero. To better look at the structure of the profile log-likelihood plot for the interior parameter space, in Fig. 1(b), we also provide the plot excluding the area where k

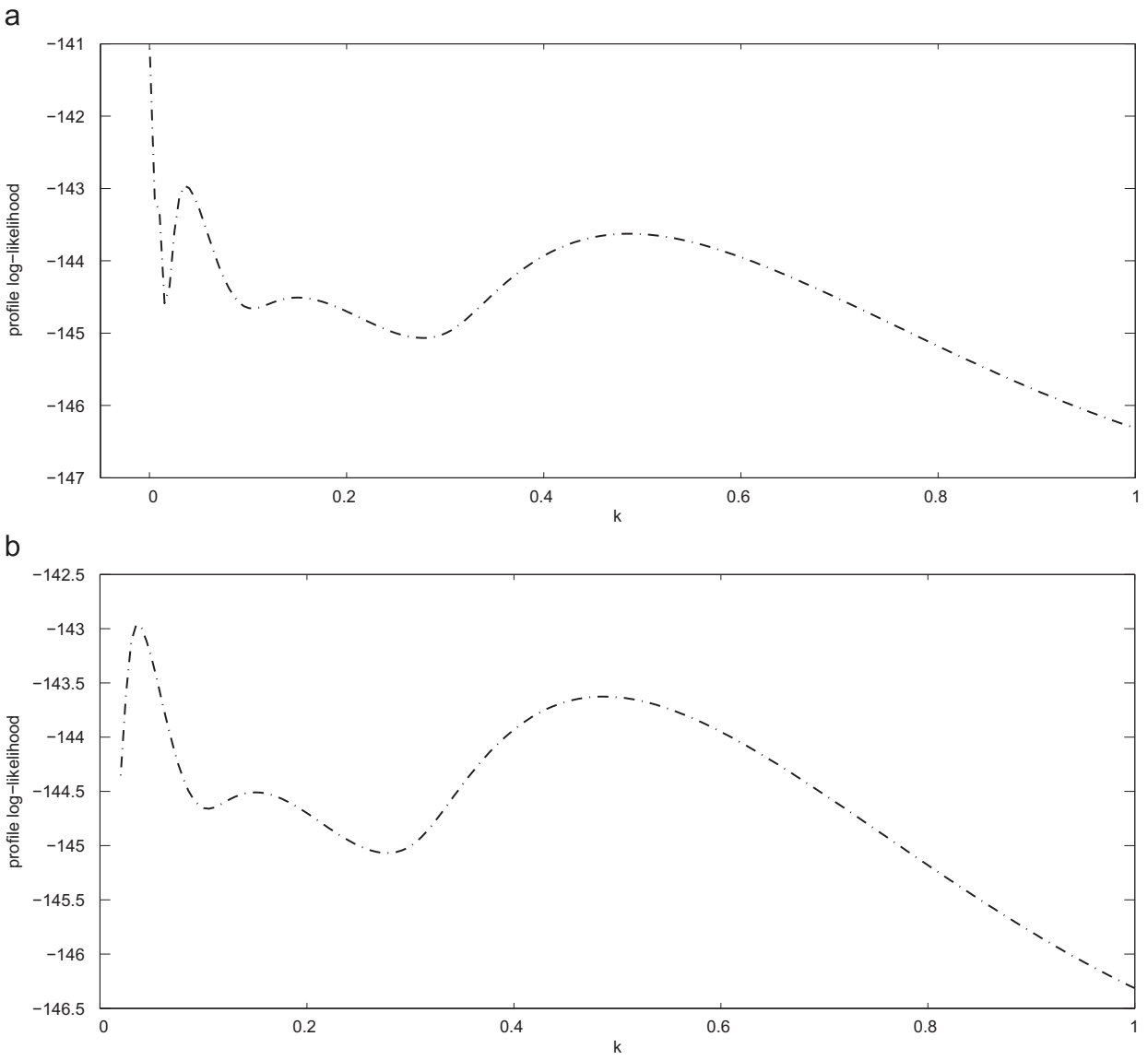

Fig. 2. Profile log-likelihood plot for Example 2: (a) for all k values from 10^{-4} to 1; (b) for k values from 0.03 to 1.

is very close to zero and the corresponding log-likelihood is relatively very large. From Fig. 1(b), one can see that there are three interior modes. The information about these three modes are reported in Table 1 (they can be easily located based on the estimated profile log-likelihood $p(k)$). By comparing the values of $\log L$, one can know that the maximum interior mode is at k=0.4378.

Based on the profile log-likelihood $p(k)$ and Fig. 1, one can also see that when $k < 0.07$ the profile log likelihood is greater than -152.9230 (the profile log likelihood value of the maximum interior mode). The value 0.07 can be found based on the estimated $p(k)$. Therefore, when the constrained EM algorithm (Hathaway, 1985, 1986) is used to find the MLE, if $C < 0.07$ in Ω_C of (6), the constrained MLE is on the boundary of the parameter space Ω_C. In fact, in this case, the constrained MLE even depends on the cut point C, which is not reasonable. If $0.07 < C < 0.4378$, the constrained EM algorithm can find the maximum interior mode and give the same result as our profile likelihood method. However, if C is too large, it is possible for the constrained EM algorithm to miss some interior modes. For example, if $0.1891 < C < 0.4378$, the constrained EM algorithm will miss the first interior mode (k=0.1891). Although the missed one is not the maximum interior mode, in many cases the interior mode can also provide useful information, especially for clustering application (McLachlan and Peel, 2000, Section 8.3.2).

Example 2. One hundred observations are generated from $0.3\mathrm{N}(0,0.5^2)+0.7\mathrm{N}(1.5,1)$. Fig. 2 is the profile log-likelihood plot. From the plot, we can see that there are about three interior modes. The corresponding information is reported in Table 2. The main controversy is on the first mode with k=0.0361, denoted by $\hat{\theta}_1$. Although $\hat{\theta}_1$ has the largest log-likelihood among all three modes, it is hard to say whether it is a real interior mode or a spurious mode that is very close to the boundary of the parameter space. If one thinks that the mode $\hat{\theta}_1$ with k=0.0361 is reasonable, then one might use it since it has the largest likelihood among all three modes. If one thinks that $\hat{\theta}_1$ is not of practical interest since one of the component

Table 2
Local maximizers for Example 2.

Local maximizer	$\log L$	π_1	μ_1	μ_2	σ_1	σ_2
k=0.0361	−142.9583	0.0685	−0.3670	1.0979	0.0376	1.0394
k=0.1516	−144.5090	0.1045	−0.4303	1.1642	0.1521	1.0036
k=0.4879	−143.6260	0.3515	0.0387	1.5172	0.4577	0.9380

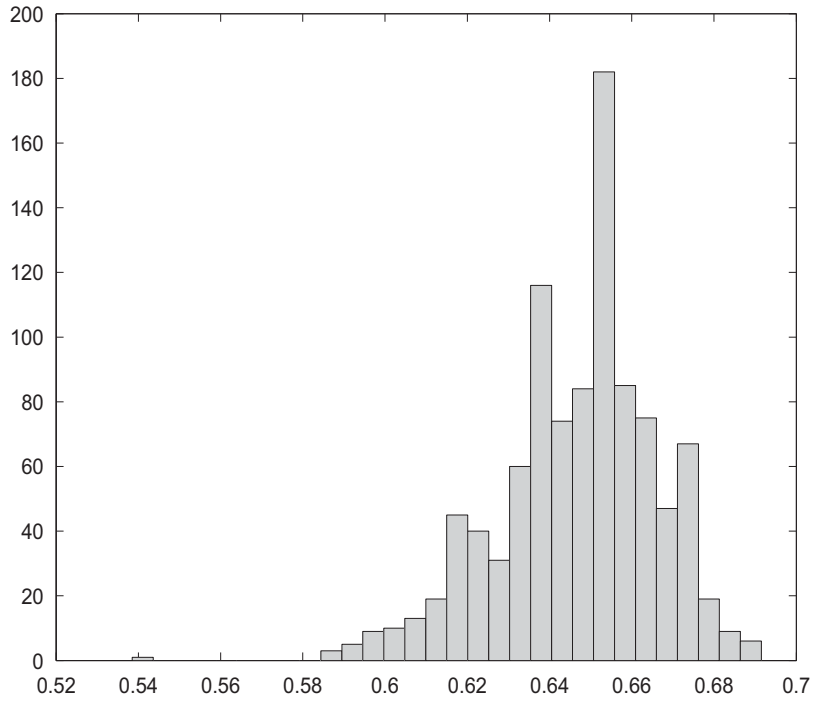

Fig. 3. Histogram of crab data. The number of bins used is 30.

proportions is only about 0.07 and the corresponding variance is also very small, then one might choose the mode with k=0.4879, which has the second largest likelihood in Table 2. In addition, from Fig. 2, one can also see that the area around the mode with k=0.4879 is much larger than the area around the mode $\hat{\theta}_1$ with k=0.0361. Therefore, when using the general EM algorithm, one might expect that most of the initial values will converge to the mode with k=0.4879.

Based on Fig. 2 and the estimated $p(k)$, one can also get that when $C > 0.0361$ in Ω_C of (6), the constrained EM algorithm (Hathaway, 1985, 1986) will miss the first mode. When $C < 0.06$, the constrained EM algorithm can always find the estimate with larger log likelihood than the mode with k=0.4879. In this case, the constrained global MLE also depends on the cut point C. If $C < 0.01$, the constrained global MLE occurs at the boundary of Ω_C and has larger log-likelihood than the first mode of k=0.0361.

3.2. Real data application

The crab data: We consider the famous crab data set analyzed by Pearson (1894). The histogram of the data is shown in Fig. 3. The data set consists of the measurements on the ratio of forehead to body length of 1000 crabs sampled from the bay of Naples. Following Pearson (1894), we use a two-component normal mixture model to analyze this data set.

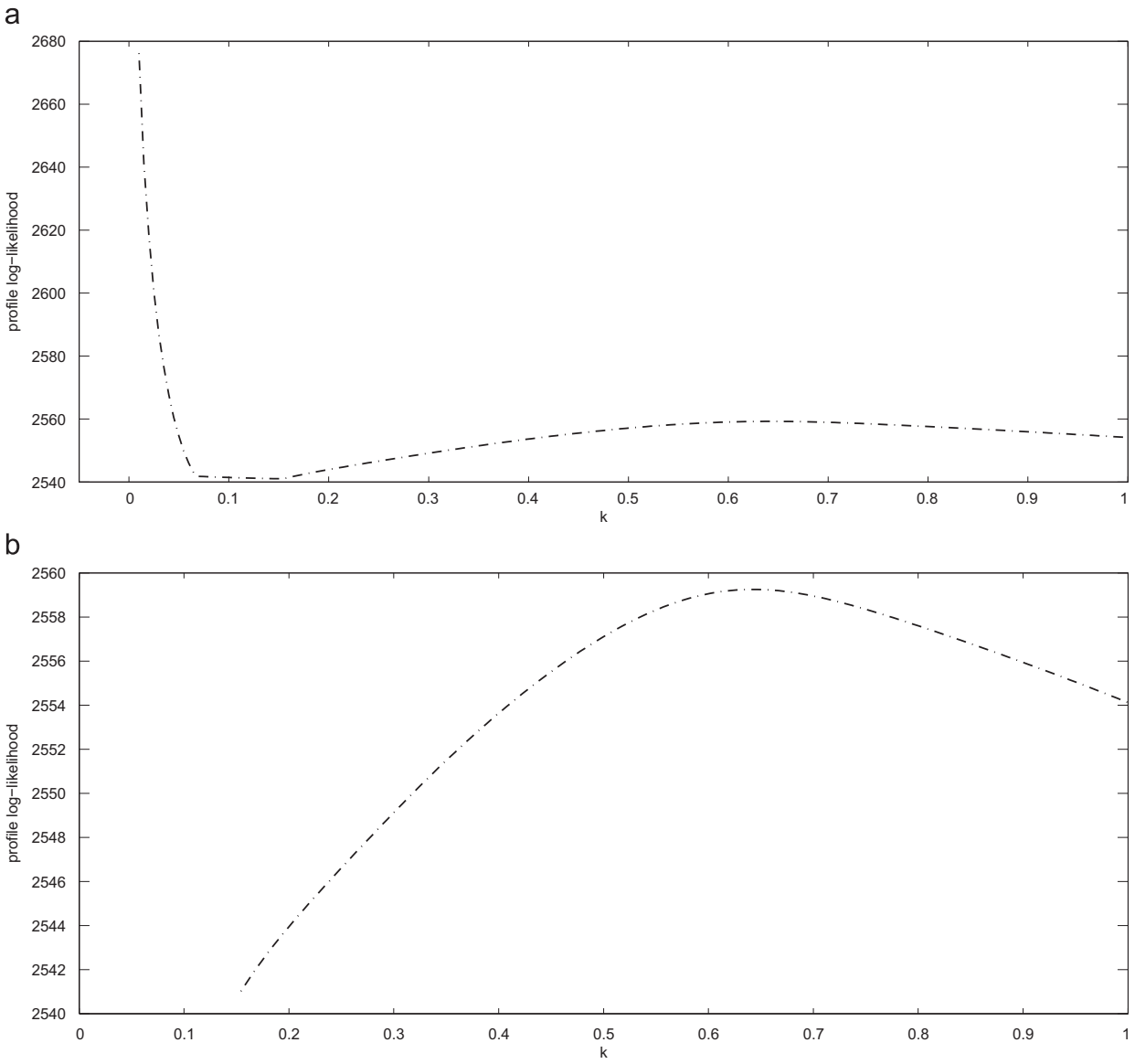

Fig. 4. Profile log-likelihood plot for crab data: (a) for all k values from 10^{-2} to 1; (b) for k values from 0.15 to 1.

Fig. 4 is our proposed profile log-likelihood plot. For this example, when k is from 10^{-4} to 10^{-2}, the corresponding log-likelihood is too large, which will affect the display of the plot. Therefore, we only provide the profile log-likelihood plot for k values from 10^{-2} to 1. From the plot, we can see that there are only one interior mode (with k=0.6418). When k=0.6418, the corresponding MLE of $(\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$ is (0.5360, 0.6563, 0.6355, 0.0126, 0.0196).

If the constrained EM algorithm is used, based on Fig. 4 and the estimated $p(k)$, when the cut point $C < 0.05$ in Ω_C of (6) the constrained global MLE occurs on the boundary of Ω_C and thus depends on the value C. When $C > 0.05$, the constrained MLE is the same as the maximum interior mode found by our proposed profile log-likelihood method.

4. Discussion

In this paper, we proposed a profile log likelihood method to solve the unboundness issue of the likelihood function for the normal mixture with unequal variance. Unlike the usual constrained EM algorithm (Hathaway, 1985, 1986), our proposed method does not need to specify a cutting point C in advance. Based on the profile log-likelihood plot and the estimated $p(k)$, one can easily identify the maximum interior mode. In addition, based on our proposed method, one can also clearly see how the cutting point C in (6) affects the constrained global MLE for the constrained EM algorithm (Hathaway, 1985, 1986). The Matlab programs for calculating the profile likelihood is available to download at "http://www-personal.ksu.edu/∼wxyao/".

For multivariate normal mixture with unequal covariance matrix, $\mathbf{\Sigma}_i$ $(i = 1, \ldots, m)$, the likelihood function is also unbounded. Similar to the univariate case, one can also put some constraint on the covariance matrix. For example, let k be the minimum of all the eigenvalues of $\mathbf{\Sigma}_h\mathbf{\Sigma}_j^{-1} (1 \le h \ne j \le m)$ or let k be the minimum of $|\mathbf{\Sigma}_h|/|\mathbf{\Sigma}_j| (1 \le h \ne j \le m)$ (Hathaway, 1985; Ingrassia, 2004). Then one can define the profile log likelihood for k similar to (7) and use it to find the maximum interior mode. The main difficulty lies on how to maximize the mixture likelihood under the above constraints. These require further research.

Acknowledgments

The authors are grateful to Bruce G. Lindsay and the two referees for their insightful comments, which greatly improved this article.

Appendix A. Proofs

Proof of Theorem 1. (a) Let $\mu_1 = x_1$. Then $\log L(\boldsymbol{\eta}; \mathbf{x}, k)$ in (4) goes into infinity when k goes to zero. Then the result follows.

(b) Given any $k \in K_C$, let $\boldsymbol{\eta}(k)$ be the corresponding parameter maximizing $\log L(\boldsymbol{\eta}; \mathbf{x}, k)$ and $\boldsymbol{\theta}(k)$ be the parameter value corresponding to $\boldsymbol{\eta}(k)$. Noting that $\boldsymbol{\theta}(k) \in \Omega_C$ and $\hat{\boldsymbol{\theta}}$ maximizes $\log L(\boldsymbol{\theta}; \mathbf{x})$ in Ω_C, hence

$$p(k) = \log L(\boldsymbol{\eta}(k); \mathbf{x}, k) = \log L(\boldsymbol{\theta}(k); \mathbf{x}) \le \log L(\hat{\boldsymbol{\theta}}; \mathbf{x}).$$

Since $\hat{k} = \hat{\sigma}_1/\hat{\sigma}_2$, one can easily know that $\boldsymbol{\theta}(\hat{k}) = \hat{\boldsymbol{\theta}}$ and $p(\hat{k}) = L(\hat{\boldsymbol{\theta}}; \mathbf{x})$. Hence $p(\hat{k}) \ge p(k)$. Therefore, $\hat{k}$ maximizes $p(k)$ in K_C. The reverse argument can be proved similarly.

(c) Suppose $\tilde{\boldsymbol{\theta}}$ is not a local mode for the log likelihood of $\log L(\boldsymbol{\theta}; \mathbf{x})$ of (3). Then for any given small $\varepsilon > 0$, then exists a $\overline{\boldsymbol{\theta}}$ satisfying $\|\overline{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| \le \varepsilon$ and $\log L(\overline{\boldsymbol{\theta}}; \mathbf{x}) > \log L(\tilde{\boldsymbol{\theta}}; \mathbf{x})$, where $\|\cdot\|$ is the Euclidian norm. Let $\overline{\boldsymbol{\theta}} = (\overline{\pi}_1, \overline{\mu}_1, \overline{\mu}_2, \overline{\sigma}_1, \overline{\sigma}_2)$ and $\overline{k} = \overline{\sigma}_1/\overline{\sigma}_2$, where $\overline{\sigma}_1 \le \overline{\sigma}_2$. Then $p(\overline{k}) = \log L(\overline{\boldsymbol{\eta}}; \mathbf{x}, \overline{k}) = \log L(\overline{\boldsymbol{\theta}}; \mathbf{x})$, where $\overline{\boldsymbol{\eta}} = (\overline{\pi}_1, \overline{\mu}_1, \overline{\mu}_2, \overline{\sigma}_2)$. Noting that $p(\tilde{k}) = \log L(\tilde{\boldsymbol{\theta}}; \mathbf{x}) < p(\overline{k})$, hence $\tilde{k} \ne \overline{k}$. Since $\|\overline{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| \le \varepsilon$, where $\tilde{\boldsymbol{\theta}} = (\tilde{\pi}_1, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{k}\tilde{\sigma}, \tilde{\sigma})$, hence $|\overline{\sigma}_1 - \tilde{k}\tilde{\sigma}| \le \varepsilon$ and $|\overline{\sigma}_2 - \tilde{\sigma}| \le \varepsilon$. Therefore

$$\frac{\tilde{k}\tilde{\sigma} - \varepsilon}{\tilde{\sigma} + \varepsilon} \le \overline{k} = \frac{\overline{\sigma}_1}{\overline{\sigma}_2} \le \frac{\tilde{k}\tilde{\sigma} + \varepsilon}{\tilde{\sigma} - \varepsilon}.$$

Let $\varepsilon \to 0$, then $\overline{k} \to \tilde{k}$. Since $p(\tilde{k}) < p(\overline{k})$ for all $\overline{k}$, $\tilde{k}$ cannot be a local mode, which contradicts the assumption. Hence $\tilde{\boldsymbol{\theta}}$ is a local mode for the log likelihood of $\log L(\boldsymbol{\theta}; \mathbf{x})$ of (3). □

Before we prove Proposition 2.1, we first provide a useful lemma.

Lemma A.1. *Let* $\hat{\boldsymbol{\sigma}}^{(t+1)} = (\hat{\sigma}_1^{(t+1)}, \ldots, \hat{\sigma}_m^{(t+1)})$ *be the minimizer of* (8), *subject to* $\sigma_{(1)} = k\sigma_{(m)}$, *where* $k \in (0, 1)$. *Let* $\hat{\sigma}_{(1)}^{(t+1)} \le \hat{\sigma}_{(2)}^{(t+1)} \cdots \le \hat{\sigma}_{(m)}^{(t+1)}$ *be the corresponding ordered minimizer. Then* $\hat{\sigma}_{(1)}^{(t+1)} \le S_1/\sqrt{n_1}$ *and* $\hat{\sigma}_{(m)}^{(t+1)} \ge S_m/\sqrt{n_m}$ *or* $\hat{\sigma}_{(1)}^{(t+1)} \ge S_1/\sqrt{n_1}$ *and* $\hat{\sigma}_{(m)}^{(t+1)} \le S_m/\sqrt{n_m}$.

Proof. For simplicity of proof, we will assume that $S_1/n_1 < S_1/n_2 < \cdots < S_m/n_m$. Let

$$Q(\boldsymbol{\sigma}) = \sum_{j=1}^{m} \left(n_j \log\sigma_j + \frac{S_j^2}{2\sigma_j^2} \right).$$

Note that

$$\frac{\partial Q(\boldsymbol{\sigma})}{\partial \sigma_j^2} = \frac{n_j}{2\sigma_j^4}(\sigma_j^2 - S_j^2/n_j).$$

Hence $Q(\boldsymbol{\sigma})$ is minimized when $\sigma_j^2 = S_j^2/n_j$. In addition, $Q(\boldsymbol{\sigma})$ is monotone increasing when $\sigma_j^2 > S_j^2/n_j$ and monotone decreasing when $\sigma_j^2 < S_j^2/n_j$.

If $\hat{\sigma}_{(1)}^{(t+1)} < S_1/\sqrt{n_1}$ and $\hat{\sigma}_{(m)}^{(t+1)} < S_m/\sqrt{n_m}$, one can easily see that $\hat{\sigma}_m^{(t+1)} = \hat{\sigma}_{(m)}^{(t+1)} = \sigma_{(1)}^{(t+1)}/k$, if considering $Q(\boldsymbol{\sigma})$ as a function σ_m by fixing other arguments. Suppose $\hat{\sigma}_{(1)}^{(t+1)} = \hat{\sigma}_i^{(t+1)}$ and $S_j/\sqrt{n_j} \le \hat{\sigma}_{(m)}^{(t+1)} < S_{j+1}/\sqrt{n_{j+1}}$. It can be seen that $\hat{\sigma}_l^{(t+1)} = S_l/\sqrt{n_l}, l \ne i$ and $l \le j$, and $\hat{\sigma}_l^{(t+1)} = \hat{\sigma}_m^{(t+1)}, j \ne i$ and $l > j$. However, under the above assumptions, when $\hat{\sigma}_{(1)}^{(t+1)}$ moves closer to $S_1/\sqrt{n_1}$ and $\hat{\sigma}_m^{(t+1)} = \hat{\sigma}_{(1)}^{(t+1)}/k$ moves closer to $S_m/\sqrt{n_m}$, the $Q(\boldsymbol{\sigma})$ will decrease. Therefore, the contradiction occurs.

Similarly, we can prove the contradiction if we assume $\hat{\sigma}_{(1)}^{(t+1)} > S_1/\sqrt{n_1}$ and $\hat{\sigma}_{(m)}^{(t+1)} > S_m/\sqrt{n_m}$. Therefore, the result follows. □

Proof of Proposition 2.1. (a) Based on Lemma A.1, since $S_1^2/n_1 \le k^2 S_m^2/n_m$, $\hat{\sigma}_{(1)}^{(t+1)} \ge S_1/\sqrt{n_1}$ and $\hat{\sigma}_{(m)}^{(t+1)} \le S_m/\sqrt{n_m}$. Suppose $S_i/\sqrt{n_i} \le \hat{\sigma}_{(1)}^{(t+1)} < S_{i+1}/\sqrt{n_{i+1}}$ and $S_{j-1}/\sqrt{n_{j-1}} < \hat{\sigma}_{(m)}^{(t+1)} \le S_j/\sqrt{n_j}$.

Based on the properties of $Q(\boldsymbol{\sigma})$ as a function of σ_j, one can easily see that $\hat{\sigma}_1^{(t+1)} = \hat{\sigma}_2^{(t+1)} = \cdots = \hat{\sigma}_i^{(t+1)} = \hat{\sigma}_{(1)}^{(t+1)} < S_{i+1}/\sqrt{n_{i+1}}$, $\hat{\sigma}_j^{(t+1)} = \hat{\sigma}_{j+1}^{(t+1)} = \cdots = \hat{\sigma}_m^{(t+1)} = \hat{\sigma}_{(m)}^{(t+1)} > S_{j-1}/\sqrt{n_{j-1}}$, and $\hat{\sigma}_l^{(t+1)} = S_l/\sqrt{n_l}, l = i+1, \ldots, j-1$.

(b) Based on Lemma A.1, since $S_1^2/n_1 > k^2 S_m^2/n_m$, $\hat{\sigma}_{(1)}^{(t+1)} \le S_1/\sqrt{n_1}$ and $\hat{\sigma}_{(m)}^{(t+1)} \ge S_m/\sqrt{n_m}$. Suppose $\hat{\sigma}_{(1)}^{(t+1)} = \hat{\sigma}_i^{(t+1)}$ and $\hat{\sigma}_{(2)}^{(t+1)} = \hat{\sigma}_j^{(t+1)}$. It can be easily seen that $\hat{\sigma}_l^{(t+1)} = S_l/n_l, l \ne i, l \ne j$ and $i < j$. In addition, if $\hat{\sigma}_{(1)}^{(t+1)} = S_1/\sqrt{n_1}$, then $\hat{\sigma}_{(1)}^{(t+1)} = \hat{\sigma}_1^{(t+1)}$. Suppose $\hat{\sigma}_{(m)}^{(t+1)} = \hat{\sigma}_j^{(t+1)} = k\hat{\sigma}_1^{(t+1)}$. If considering $Q(\boldsymbol{\sigma})$ as a function of σ_1, we can easily prove that the minimizer is not $S_1/\sqrt{n_1}$. The contradiction occurs. Hence, $\hat{\sigma}_{(1)}^{(t+1)} < S_1/\sqrt{n_1}$. Similarly, we can also prove $\hat{\sigma}_{(m)}^{(t+1)} > S_m/\sqrt{n_m}$. □

Remarks. 1. From the above proof, we can see that we have proved the stronger results than Proposition 2.1, i.e. the strict inequality holds for $\hat{\boldsymbol{\sigma}}^{(t+1)}$. Hence, in Step 1 of Section 2.2, we only need to consider $\tilde{\boldsymbol{\sigma}}_{(i,j)}$'s when the strict inequality constraint holds. For example, if $S_1^2/n_1 > k^2 S_m^2/n_m$, we only need to consider $\tilde{\boldsymbol{\sigma}}_{(i,j)}$'s when $k^2 S_m^2/n_m < \breve{\sigma}_{(i,j)} < S_1^2/n_1$, where $\breve{\sigma}_{(i,j)}$ is defined in (10).

2. In addition, when $S_1^2/n_1 \le k^2 S_m^2/n_m$, it can be seen that $Q(\tilde{\boldsymbol{\sigma}}_{(i,j)}) < Q(\tilde{\boldsymbol{\sigma}}_{(i',j')})$ when $i' > i, j' < j$, and the strict inequality constraint holds for $\tilde{\boldsymbol{\sigma}}_{(i,j)}$ and $\tilde{\boldsymbol{\sigma}}_{(i',j')}$, since $\tilde{\boldsymbol{\sigma}}_{(i,j)}$ minimizes $Q(\boldsymbol{\sigma})$ over larger parameter space than $\tilde{\boldsymbol{\sigma}}_{(i',j')}$. Let $n(i)$ be the largest j values for fixed i such that the inequality constraint holds for $\tilde{\boldsymbol{\sigma}}_{(i,j)}$ and $\tilde{n}(i) = \max\{n(1), \ldots, n(i-1)\}$. Then, we only need to consider i when $n(i) > \tilde{n}(i)$, i.e. for i, we only need to consider $j = \tilde{n}(i)+1, \ldots, m$. If $\tilde{n}(i) = m$ for some i, then we can stop and need not calculate $\tilde{\boldsymbol{\sigma}}_{(lj)}$ for $l = i+1, \ldots, m-1$. □

Proof of Proposition 2.2. (a) Under the constraint that $\sigma_1 = \sigma_2 = \cdots = \sigma_i = \sigma$ and $\sigma_j = \sigma_{j+1} = \cdots = \sigma_m = \sigma/k$,

$$\frac{Q(\boldsymbol{\sigma})}{\partial \sigma^2} = \frac{\sum_{l=1}^{i} n_l + \sum_{l=j}^{m} n_l}{2\sigma^4}\left(\sigma^2 - \frac{\sum_{l=1}^{i} S_l^2 + k^2 \sum_{l=j}^{m} S_l^2}{\sum_{l=1}^{i} n_l + \sum_{l=j}^{m} n_l}\right).$$

Therefore the result follows.

(b) The proof is similar to the proof of (a). □

References

Bezdek, J.C., Hathaway, R.M., Huggins, V.J., 1985. Parametric estimation for normal mixtures. Pattern Recognition 3, 79–84.

Böhning, D., 1999. Computer-Assisted Analysis of Mixtures and Applications. Chapman & Hall, CRC, Boca Raton, FL.

Chen, J., Tan, X., Zhang, R., 2008. Inference for normal mixture in mean and variance. Statistica Sincia 18, 443–465.

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

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

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer, Berlin.

Hathaway, R.J., 1985. A constrained formulation of maximum-likelihood estimation for nomral mixture distributions. Annals of Statistics 13, 795–800.

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

Ingrassia, S., 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. Statistical Methods & Applications 13, 151–166.

Karlis, D., Xekalaki, E., 2003. Choosing initial values for the EM algorithm for finite mixtures. Computational Statistics & Data Analysis 41, 577–590.

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

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

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

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

Peters, B.C., Walker, H.F., 1978. An iterative procedure for obtaining maximum likelihood estimators of the parameters for a mixture of normal distributions. SIAM Journal on Applied Mathematics 35, 362–378.

Yao, W., Lindsay, B.G., 2009. Bayesian mixture labelling by highest posterior density. Journal of the American Statistical Association 104, 758–767.