Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Empirical likelihood estimation in multivariate mixture models with repeated measurements

Yuejiao Fu, Yukun Liu, Hsiao-Hsuan Wang & Xiaogang Wang

To cite this article: Yuejiao Fu, Yukun Liu, Hsiao-Hsuan Wang & Xiaogang Wang (2020) Empirical likelihood estimation in multivariate mixture models with repeated measurements, Statistical Theory and Related Fields, 4:2, 152-160, DOI: 10.1080/24754269.2019.1630544

To link to this article: https://doi.org/10.1080/24754269.2019.1630544

Published online: 19 Jun 2019.

Submit your article to this journal

Article views: 95

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
2020, VOL. 4, NO. 2, 152–160
https://doi.org/10.1080/24754269.2019.1630544

Taylor & Francis
Taylor & Francis Group

Check for updates

Empirical likelihood estimation in multivariate mixture models with repeated measurements

Yuejiao Fu [a], Yukun Liu [b], Hsiao-Hsuan Wang [a] and Xiaogang Wang [a,c]

[a]Department of Mathematics and Statistics, York University, Toronto, Canada; [b]School of Statistics, East China Normal University, Shanghai, China; [c]Institute of Data Science, Tsinghua University, Beijing, China

ABSTRACT
Multivariate mixtures are encountered in situations where the data are repeated or clustered measurements in the presence of heterogeneity among the observations with unknown proportions. In such situations, the main interest may be not only in estimating the component parameters, but also in obtaining reliable estimates of the mixing proportions. In this paper, we propose an empirical likelihood approach combined with a novel dimension reduction procedure for estimating parameters of a two-component multivariate mixture model. The performance of the new method is compared to fully parametric as well as almost nonparametric methods used in the literature.

ARTICLE HISTORY
Received 12 November 2018
Revised 12 May 2019
Accepted 7 June 2019

KEYWORDS
Empirical likelihood; estimating equation; repeated measurements; multivariate mixture model

1. Introduction

Mixture models provide a flexible way of modelling complex data obtained from a population with observed or unobserved heterogeneity. Mixture models have been applied in astronomy, biology, fishery, human genetics, and other scientific areas of research. See Titterington, Smith, and Makov (1985), Lindsay (1995), McLachlan and Peel (2000), and references therein.

We consider a special multivariate mixture model where repeated measurements are available for each subject. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent and identically distributed (i.i.d.) d-variate random vectors from a finite mixture model with m components. If the elements of the vector $\boldsymbol{X}_i$ are independent conditional on belonging to a subpopulation, then the mixture density is given by

$$h(\boldsymbol{x}) = \sum_{j=1}^{m} \pi_j \prod_{r=1}^{d} f_{jr}(x_r), \tag{1}$$

where π_j's are mixing proportions such that $\sum_{j=1}^{m} \pi_j = 1$, $\pi_j > 0$ for all j, and $f(\cdot)$, with or without subscripts, denotes a univariate density function.

The above data structure is quite common especially in social sciences where measurements are taken repeatedly for various reasons. For example, the goal of research on preschool children's inclusion task responses is to study different solution strategies with which young children solve a given cognitive task. The solution strategy is often called the latent variable since it is hidden and unobservable. A group of preschool children can be considered as a sample from a mixture model where the components correspond to the various solution strategies; see Thomas and Horton (1997). In a simplified setting, one could assume that there are two main solution strategies which lead to a mixture model with two components.

Many researchers studied the nonparametric identifiability of the above multivariate mixture model. Hall and Zhou (2003) showed that the model (1) is always nonparametrically unidentifiable when $d = 2$ and $m = 2$. Under some mild regularity conditions, Hall and Zhou (2003) proved that the two-component mixture model is nonparametrically identifiable for $d \geq 3$. Kasahara and Shimotsu (2014) discussed the identifiability of the number of components in multivariate mixture models in which each component distribution has independent marginals. Hettmansperger and Thomas (2000) considered the situation where the elements of the vector $\boldsymbol{X}_i$ are, not only conditionally independent, but also identically distributed. Under such an assumption, the mixture density (1) can be rewritten as

$$h(\boldsymbol{x}) = \sum_{j=1}^{m} \pi_j \prod_{r=1}^{d} f_j(x_r). \tag{2}$$

They proposed an almost nonparametric approach to estimate the mixing proportions. Their key idea is to categorise data into 0 or 1 by setting an optimal cut point and then apply the EM algorithm to estimate the mixing proportion in the resulting binomial mixture

CONTACT Yukun Liu ykliu@sfs.ecnu.edu.cn

© East China Normal University 2019

models. Cruz-Medina, Hettmansperger, and Thomas (2004) extended the work of Hettmansperger and Thomas (2000) by transforming the observed vector into a count vector which leads to a multinomial mixture model.

To avoid possible loss of efficiency in categorising continuous data into count data, we propose a nonparametric approach to estimate the mixing proportions using empirical likelihood (EL). The EL, which was first introduced by Owen (1988), is a nonparametric method of inference based on a data-driven likelihood ratio function. This nonparametric and likelihood-based approach has become one of the most effective statistical methods. See Owen (2001) for a comprehensive review. As shown in Qin and Lawless (1994), the EL is a prominent efficient tool in estimating parameters by incorporating estimating equations into constrained maximisation of the empirical likelihood function.

We first develop the proposed methodology for the 3-dimensional mixture models, and later on extend it to higher dimensions. For the multivariate mixture model, we propose linking the various moment estimating equations through the EL to provide a more efficient estimation. In the d-dimensional mixture model, there are $2^d - 1$ moment estimating equations. When d is large, it is impracticable to search for the optimal solution. We propose a simple and intuitive bootstrap-like modification of the method. First we obtain K sets of three indices chosen randomly and without replacement from $1, 2, \ldots, d$, and then multiply the K nonparametric likelihoods pertinent to the chosen indices to obtain the profile empirical likelihood ratio function.

Our simulation results show that, when the parametric model is correctly specified, our EL estimators perform similarly to the parametric estimators. However, when the parametric model is misspecified, the EL estimators perform uniformly better than the parametric estimators and the almost nonparametric estimators.

The paper is organised as follows. The proposed empirical likelihood approach for multivariate mixture model and its theoretical properties are presented in Section 2. The extension to d-dimensional ($d > 3$) mixtures is also presented. Simulation studies and real data analysis are provided in Section 3. Discussions are given in Section 4.

2. Methodology

We first discuss the methodology for the three-variate mixture model, and then extend to multivariate mixtures with higher dimensions.

2.1. Three-variate mixture model

Let $\boldsymbol{X} = (X_1, X_2, X_3)^{\mathrm{T}}$ be a 3-dimensional random vector with distribution function $H(\boldsymbol{x})$ and joint probability density function

$$h(\boldsymbol{x}) = \pi \prod_{i=1}^{3} f_1(x_i) + (1-\pi) \prod_{i=1}^{3} f_2(x_i), \qquad (3)$$

where $0 \le \pi \le 1$, and the component density functions f_1 and f_2 are different but unspecified. This model is a special case of model (2) with $m = 2$ and $d = 3$.

The parameters of interest are the expectations of the random variables and the mixing proportion π. Suppose μ_0 and μ_1 are the expected values of the two components:

$$\mu_0 = \int x f_1(x)\,\mathrm{d}x, \quad \mu_1 = \int x f_2(x)\,\mathrm{d}x,$$

and that they satisfy $\mu_0 < \mu_1$. We then have the following moment estimating equations

$$\begin{aligned} \mathbb{E}(X_1X_2X_3) &= \pi\mu_0^3 + (1-\pi)\mu_1^3, \\ \mathbb{E}(X_1X_2) = \mathbb{E}(X_1X_3) &= \mathbb{E}(X_2X_3) \\ &= \pi\mu_0^2 + (1-\pi)\mu_1^2, \\ \mathbb{E}(X_1) = \mathbb{E}(X_2) &= \mathbb{E}(X_3) = \pi\mu_0 + (1-\pi)\mu_1. \end{aligned}$$

There are seven estimating equations in total with three unknown parameters (π, μ_0, μ_1).

Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$, $i = 1, \ldots, n$, be i.i.d. observations from the multivariate mixture model (3), and $p_i = dH(\boldsymbol{x}_i)$. According to Owen (1988), the EL function based on the observed data is

$$\prod_{i=1}^{n} dH(\boldsymbol{x}_i) = \prod_{i=1}^{n} p_i. \qquad (4)$$

Let $\boldsymbol{\theta} = (\pi, \mu_0, \mu_1)^{\mathrm{T}}$. For the distribution $H(\boldsymbol{x})$ under study, feasible p_i's satisfy

$$\sum_{i=1}^{n} p_i = 1, \quad p_i \ge 0, \quad \text{and} \quad \sum_{i=1}^{n} p_i \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{0}, \qquad (5)$$

where

$$\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) = (g_1(\boldsymbol{x}_i, \boldsymbol{\theta}), \boldsymbol{g}_2^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{\theta}), \boldsymbol{g}_3^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{\theta}))^{\mathrm{T}} \qquad (6)$$

with $g_1(\boldsymbol{x}_i, \boldsymbol{\theta}) = x_{i1}x_{i2}x_{i3} - \pi\mu_0^3 - (1-\pi)\mu_1^3$,

$$\boldsymbol{g}_2(\boldsymbol{x}_i, \boldsymbol{\theta}) = \begin{pmatrix} x_{i1}x_{i2} - \pi\mu_0^2 - (1-\pi)\mu_1^2 \\ x_{i1}x_{i3} - \pi\mu_0^2 - (1-\pi)\mu_1^2 \\ x_{i2}x_{i3} - \pi\mu_0^2 - (1-\pi)\mu_1^2 \end{pmatrix} \quad \text{and}$$

$$\boldsymbol{g}_3(\boldsymbol{x}_i, \boldsymbol{\theta}) = \begin{pmatrix} x_{i1} - \pi\mu_0 - (1-\pi)\mu_1 \\ x_{i2} - \pi\mu_0 - (1-\pi)\mu_1 \\ x_{i3} - \pi\mu_0 - (1-\pi)\mu_1 \end{pmatrix}.$$

Inference on $\boldsymbol{\theta}$ is usually made through their profile likelihood, which is obtained by maximising (4) with respect to p_i's subject to the constraints in (5). Up to

a constant not depending on $\boldsymbol{\theta}$, the resulting empirical log-likelihood is

$$\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})\},$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier determined by

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})} = \mathbf{0}.$$

We can show that in a $O(n^{-1/3})$ neighbourhood of the true values of $\boldsymbol{\theta}$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\theta})$ is determined uniquely by an implicit function of $\boldsymbol{\theta}$. We denote the maximum empirical likelihood estimators as $\hat{\boldsymbol{\theta}} = (\hat{\pi}, \hat{\mu}_0, \hat{\mu}_1)^{\mathrm{T}}$. Their asymptotic properties are given in the following theorem by Qin and Lawless (1994). When $\boldsymbol{\theta}$ takes its true value $\boldsymbol{\theta}_0$, we write $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta}_0)$ to be $\boldsymbol{g}(\boldsymbol{x})$ for short.

Theorem 2.1: *Under the regularity conditions specified in* Qin and Lawless (1994). *As n goes to infinity,* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{V}_1)$, *where*

$$\boldsymbol{V}_1 = \left[\mathbb{E}\left\{ \frac{\partial \boldsymbol{g}(\boldsymbol{X})}{\partial \boldsymbol{\theta}} \right\}^{\mathrm{T}} \{\mathbb{E}\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{X})\}^{-1} \mathbb{E}\left\{ \frac{\partial \boldsymbol{g}(\boldsymbol{X})}{\partial \boldsymbol{\theta}} \right\} \right]^{-1}.$$

With $(\mathbf{1}(X_1 \le t), \mathbf{1}(X_2 \le t), \mathbf{1}(X_3 \le t))$ in place of $\boldsymbol{X}$, we can estimate the underlying distribution functions $F_1(t)$ and $F_2(t)$. The asymptotic normality of the resulting empirical likelihood estimators can be established in a similar way to Theorem 2.1.

2.2. Multivariate mixtures with higher dimensions

We now extend the methodology discussed in the previous section to the case with $d > 3$. Suppose the d-variate data $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{id})^{\mathrm{T}}$, $i = 1, \ldots, n$, arise from the mixture model with the following mixture density

$$h(\boldsymbol{w}_i) = \pi \prod_{j=1}^{d} f_1(w_{ij}) + (1 - \pi) \prod_{j=1}^{d} f_2(w_{ij}).$$

In principle, we can adopt the same approach as in the case $d = 3$ in order to make inferences about $\boldsymbol{\theta}$. When d is large, however, the number of estimating equations we must deal with is

$$\binom{d}{d} + \binom{d}{d-1} + \cdots + \binom{d}{1} = 2^d - 1,$$

which can be extremely large. Consequently, it is impractical to find the optimal solution to embrace that many estimating equations in the empirical likelihood setup.

We now propose a simple and intuitive solution to the high-dimensional problem. Let $M_d = \binom{d}{3}$, and Ω_i $(i = 1, 2, \ldots, M_d)$ be all the possible samples of size 3 from $\{1, 2, \ldots, d\}$ drawn by simple random sampling without replacement. We randomly select K sets from $\{\Omega_1, \ldots, \Omega_{M_d}\}$ by simple random sampling without replacement. Let $\Omega_k^* = \{s_{k1}, s_{k2}, s_{k3}\}$ $(k = 1, 2, \ldots, K)$ be the resulting K index sets, and $\boldsymbol{u}_{ki} = (x_{ki}, y_{ki}, z_{ki})^{\mathrm{T}}$ denote $(w_{i,s_{k1}}, w_{i,s_{k2}}, w_{i,s_{k3}})^{\mathrm{T}}$. We assume $s_{k1} < s_{k2} < s_{k3}$ for each k, and treat the data with different Ω_k^* as independent samples. The profile empirical likelihood ratio function of $\boldsymbol{\theta}$ based on the selected index sets is

$$R(\boldsymbol{\theta}) = \max \left\{ \prod_{k=1}^{K} \prod_{i=1}^{n} (np_{ki}) \,\middle|\, \sum_{i=1}^{n} p_{ki} = 1,\ p_{ki} \ge 0, \right.$$
$$\left. \sum_{i=1}^{n} p_{ki} \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta}) = \mathbf{0},\ k = 1, \ldots, K \right\},$$

where the function $\boldsymbol{g}$ is defined in (6).

Applying the method of constrained optimisation, we have

$$G = \sum_{k=1}^{K} \sum_{i=1}^{n} \log(np_{ki}) - n \sum_{k=1}^{K} \sum_{i=1}^{n} p_{ki} \boldsymbol{\lambda}_k^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})$$
$$+ \sum_{k=1}^{K} \gamma_k \left(\sum_{i=1}^{n} p_{ki} - 1 \right),$$

where $\boldsymbol{\lambda}_k$ and γ_k are the Lagrange multipliers. Setting the first derivative of G with respect to p_{ki} to zero, we have

$$\frac{\partial G}{\partial p_{ki}} = \frac{1}{p_{ki}} - n\boldsymbol{\lambda}_k^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta}) + \gamma_k = 0.$$

Multiplying both sides of the above equation by p_{ki} and summing over i give

$$\sum_{i=1}^{n} p_{ki} \frac{\partial G}{\partial p_{ki}} = n + \gamma_k = 0,$$

which leads to $\gamma_k = -n$. Therefore, the maximum of $\prod_{k=1}^{K} \prod_{i=1}^{n} (np_{ki})$ is attained at

$$\hat{p}_{ki} = \frac{1}{n} \frac{1}{1 + \boldsymbol{\lambda}_k^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})}, \quad k = 1, \ldots, K,$$

where the Lagrange multipliers $\boldsymbol{\lambda}_k = \boldsymbol{\lambda}_k(\boldsymbol{\theta})$'s are the solutions to

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}_k^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})} = 0.$$

Putting $\hat{p}_{ki}$ back and taking logarithm, we have the profile empirical log-likelihood ratio function of $\boldsymbol{\theta}$,

$$\ell(\boldsymbol{\theta}) = \log\{R(\boldsymbol{\theta})\} = -\sum_{k=1}^{K} \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}_k^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})\}.$$

We show that with probability tending to one, there must be a local maximum point in a very small neighbourhood of the true parameter value of $\boldsymbol{\theta}$. Let $\Omega^* = \{\Omega_1^*, \ldots, \Omega_K^*\}$.

Lemma 2.1: *Let* $\boldsymbol{\theta}_0 = (\pi_*, \mu_{0*}, \mu_{1*})$ *be the true value of* $\boldsymbol{\theta}$. *Suppose* $\int |x|^9 \, \mathrm{d}F_0(x) + \int |x|^9 \, \mathrm{d}F_1(x) < \infty$, $\pi_* \in (0,1)$ *and* $\mu_{0*} \neq \mu_{1*}$, *and that* F_0 *and* F_1 *are non-degenerate distributions. Conditioning on* Ω^*, *as* $n \to \infty$, $\ell(\boldsymbol{\theta})$ *attains its maximum value at some point* $\hat{\boldsymbol{\theta}}$ *with probability 1 in the interior of the ball* $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq n^{-1/3}$. *Let* $\hat{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\lambda}}_1^{\mathrm{T}}, \ldots, \hat{\boldsymbol{\lambda}}_K^{\mathrm{T}})^{\mathrm{T}}$ *with* $\hat{\boldsymbol{\lambda}}_i = \boldsymbol{\lambda}(\hat{\boldsymbol{\theta}})$. *Consequently,* $\hat{\boldsymbol{\theta}}$ *and* $\hat{\boldsymbol{\lambda}}$ *satisfy*

$$\boldsymbol{Q}_{kn}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}}_k) = 0 \quad \textit{for } k = 1, \ldots, K, \quad \textit{and}$$
$$\boldsymbol{Q}_{0n}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}}) = 0,$$

where

$$\boldsymbol{Q}_{kn}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}_k^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})},$$

$$\boldsymbol{Q}_{0n}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \frac{1}{1 + \boldsymbol{\lambda}_k^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})} \left(\frac{\partial \boldsymbol{g}(\boldsymbol{u}_{ki}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{\mathrm{T}} \boldsymbol{\lambda}_k.$$

Lemma 2.1 implies that the proposed EL estimator $\hat{\boldsymbol{\theta}}$ is consistent. Based on Lemma 2.1, we further establish the asymptotic normality of $\hat{\boldsymbol{\theta}}$ in the following theorem. This result is an extension of Theorem 1 in Qin and Lawless (1994). It embraces the correlation structure of the selected elements within the random vectors.

Theorem 2.2: *Assume the conditions of Lemma* 2.1. *Let* $S_{11} = \mathbb{E}\{\boldsymbol{g}(\boldsymbol{X})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{X})\}$, $S_{12} = S_{21}^{\mathrm{T}} = -\mathbb{E}\{\partial \boldsymbol{g}(\boldsymbol{X})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$, *and*

$$\Sigma_{\mathrm{off}} = \frac{1}{K(K-1)} \sum_{1 \leq k \neq j \leq K} \mathbb{E}\{\boldsymbol{g}(\boldsymbol{u}_{k1})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{j1}) | \Omega^*\}.$$

Conditioning on Ω^*, *as n goes to infinity,* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *converges in distribution to* $N(0, \boldsymbol{V}_2)$, *where*

$$\boldsymbol{V}_2 = \frac{1}{K}(S_{21}S_{11}^{-1}S_{12})^{-1} + \frac{K-1}{K}(S_{21}S_{11}^{-1}S_{12})^{-1}(S_{21}S_{11}^{-1}) \Sigma_{\mathrm{off}}(S_{11}^{-1}S_{12})(S_{21}S_{11}^{-1}S_{12})^{-1}.$$

If there are no common elements in Ω_k^* and Ω_j^*, then $\mathbb{E}\{\boldsymbol{g}(\boldsymbol{u}_{k1})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{j1}) \mid \Omega^*\} = 0$. Further, if d is quite large, and there are no common elements in any pair of Ω_k^* and Ω_j^* $(k \neq j)$, then $\Sigma_{\mathrm{off}} = 0$, and $\boldsymbol{V}_2 = (S_{21}S_{11}^{-1}S_{12})^{-1}/K$. At the other extreme, if $\Omega_k^* = \Omega_1^*$ for $k = 2, \ldots, K$, then $\Sigma_{\mathrm{off}} = S_{11}$, and $\boldsymbol{V}_2 = (S_{21}S_{11}^{-1}S_{12})^{-1}$. Therefore, the second term in $\boldsymbol{V}_2$ stands for the efficiency loss due to the fact that some data are used more than once.

3. Simulation studies and data analysis

3.1. Simulation studies

We have carried out simulations to evaluate the finite-sample performance of the proposed empirical likelihood estimators (EL). For comparison, we have also considered two of its competitors: the maximum likelihood estimators (ML) under the multivariate normal mixture model, and the almost nonparametric estimators based on multinomial mixtures (Cruz-Medina et al. (2004); MN for short). Both the ML and MN estimators can be calculated by the EM algorithm.

We generate data from the mixture model (3). Different specifications of component distributions f_1 and f_2 are listed below:

(a) (Normal mixtures) f_1 and f_2 are the density functions of $N(\mu_1, 1)$ and $N(\mu_2, 1)$, respectively. Here $\mu_1 = 0$ and $\mu_2 = 1$ or 2.
(b) (Non-central t mixtures) f_1 and f_2 are the density functions of $t(r, a(r)\mu_1)$ and $t(r, a(r)\mu_2)$, respectively. Here $t(r, a(r)\mu)$ denotes a t-distribution with r degrees of freedom, non-centrality parameter $a(r)\mu$, and mean μ, where $a(r) = (2/r)(\Gamma(r/2)/\Gamma((r-1)/2))$. Here $r = 4$, $\mu_1 = 0$, and $\mu_2 = 1.5$ or 2.
(c) (Chi-square mixtures) f_1 and f_2 are the density functions of $\chi^2_{\mu_1}$ and $\chi^2_{\mu_2}$. Here $\mu_1 = 5$ and $\mu_2 = 10$ or 20.

For each setting, we generate 1000 samples with sample size $n = 400$, $d = 3$ or 6, and $\pi = 0.2$, 0.5, or 0.8. When $d = 6$, we set $K = 8$ in the proposed EL method. We calculate the biases and standard deviations of the estimators under comparison, and summarise the results in Tables 1–3.

Let us first examine Table 1, where the multivariate normal mixture model is correctly specified. As expected, the ML estimators have the smallest standard deviations in all cases and the smallest absolute biases in most cases. The proposed EL estimators perform very similarly to the ML estimators and both of them are uniformly better than the MN estimators. As μ_2 goes further away from $\mu_1 = 0$, all estimators have decreasing standard deviations. This may be because the two component distributions in the mixture model also get further away from each other. When π increases from 0.2 to 0.8, the performances of all the three estimators for μ_1 are getting better, while those for μ_2 are getting worse. This is probably because as π increases, the multivariate normal mixture contains increasing information about μ_1 but decreasing information about μ_2. All the three estimators for π have better performance when π lies in the middle than on the boundaries of its parameter space.

Table 1. Biases (%) and standard deviations (%) (in parentheses) of different estimators based on 1,000 simulations with $n = 400$. Data were generated from the multivariate mixture model with f_1 and f_2 being $N(\mu_1, 1)$ and $N(\mu_2, 1)$, respectively. Here $\mu_1 = 0$, $\mu_2 = 1$ or 2 and $d = 3$ or 6.

	$d=3$			$d=6$		
Method	π	μ_1	μ_2	π	μ_1	μ_2
			$\pi = 0.2, \mu_2 = 1$			
EL	3.34(14.96)	−4.5(33.33)	2.33(11.54)	0.97(6.52)	0.18(14.11)	0.49(4.95)
MN	16.78(19.18)	13.84(40.78)	10.38(16.05)	2.46(9.62)	−20.72(25.32)	5.83(5.96)
ML	2.19(11.32)	−1.33(23.21)	1.45(8.56)	0.24(3.89)	−0.07(8.73)	0.08(3.28)
			$\pi = 0.2, \mu_2 = 2$			
EL	0.29(3.25)	0.61(13.31)	0.27(4.96)	−0.05(2.39)	−0.10(7.34)	−0.06(2.98)
MN	1.74(5.43)	−32.19(28.69)	10.57(7.03)	0.08(2.53)	−46.28(13.04)	10.53(3.53)
ML	0.10(2.46)	0.16(9.20)	0.07(3.75)	0.01(2.09)	0.11(5.00)	0.14(2.42)
			$\pi = 0.5, \mu_2 = 1$			
EL	0.72(13.16)	−1.09(15.10)	2.46(15.25)	−0.2(6.37)	−0.39(6.81)	−0.14(6.89)
MN	0.94(13.75)	−6.76(23.29)	8.86(23.67)	−0.3(5.50)	−13.79(10.83)	13.04(11.17)
ML	−0.2(10.45)	−1.37(11.76)	0.66(11.46)	−0.29(4.35)	−0.28(4.70)	−0.25(4.81)
			$\pi = 0.5, \mu_2 = 2$			
EL	0.06(3.68)	0.09(6.54)	0.22(6.49)	−0.16(2.87)	0.01(4.06)	−0.28(3.98)
MN	0.00(4.15)	−16.14(16.52)	16.42(16.16)	−0.15(2.78)	−19.97(7.75)	19.64(6.88)
ML	0.00(2.94)	−0.04(4.94)	0.11(4.96)	−0.17(2.63)	−0.10(2.94)	−0.16(3.08)
			$\pi = 0.8, \mu_2 = 1$			
EL	−3.34(15.06)	−2.09(11.78)	4.23(34.49)	−0.97(7.08)	−0.38(5.06)	0.17(14.21)
MN	−17.77(20.32)	−10.54(16.36)	−14.67(41.56)	−2.17(8.89)	−5.83(5.81)	21.41(22.80)
ML	−3.71(13.25)	−2.26(9.88)	−0.95(25.60)	−0.53(3.90)	−0.23(3.28)	−0.37(8.70)
			$\pi = 0.8, \mu_2 = 2$			
EL	−0.37(3.36)	−0.24(5.07)	−0.07(13.19)	0.01(2.36)	−0.12(3.09)	−0.29(7.01)
MN	−1.95(6.22)	−10.16(7.70)	32.08(32.32)	−0.24(4.00)	−10.3(4.17)	45.26(19.93)
ML	−0.25(2.37)	−0.14(3.81)	0.06(8.70)	0.07(2.09)	−0.07(2.42)	0.12(4.61)

Table 2. Biases (%) and standard deviations (%) (in parentheses) of different estimators based on 1,000 simulations with $n = 400$. Data were generated from the multivariate mixture model with f_1 and f_2 being $t(4, \mu_1)$ and $t(4, \mu_2/\{\sqrt{2}\Gamma(3/2)/\Gamma(2)\})$, respectively. Here $\mu_1 = 0$, $\mu_2 = 1.5$ or 2 and $d = 3$ or 6.

	$d=3$			$d=6$		
Method	π	μ_1	μ_2	π	μ_1	μ_2
			$\pi = 0.2, \mu_2 = 1.5$			
EL	3.45(15.40)	−7.18(50.85)	3.84(21.38)	−0.49(5.90)	−3.91(20.41)	−1.44(7.11)
MN	9.79(15.66)	−14.9(82.18)	51.68(46.55)	1.08(7.48)	−71.86(67.91)	59.47(43.38)
ML	61.85(12.85)	101.00(22.48)	66.01(59.61)	56.13(21.56)	95.17(35.27)	38.41(37.84)
			$\pi = 0.2, \mu_2 = 2$			
EL	1.13(8.87)	−2.98(37.27)	0.92(13.61)	−0.46(3.92)	−2.47(16.23)	−1.53(6.36)
MN	3.42(8.78)	−51.51(87.04)	59.89(60.83)	0.67(5.98)	−92.36(66.26)	79.47(59.61)
ML	58.3(19.12)	128.91(42.01)	77.41(164.19)	17.33(29.78)	41.3(64.06)	21.8(68.62)
			$\pi = 0.5, \mu_2 = 1.5$			
EL	−0.56(12.44)	−3.27(21.17)	1.19(22.59)	−0.93(5.99)	−1.50(9.17)	−1.98(9.76)
MN	−0.42(11.04)	−31.21(48.91)	61.71(65.51)	0.01(4.61)	−42.77(35.48)	85.2(55.73)
ML	30.44(11.70)	53.29(16.19)	21.03(52.65)	15.15(16.04)	29.22(24.27)	8.13(30.20)
			$\pi = 0.5, \mu_2 = 2$			
EL	−0.61(7.80)	−1.46(15.47)	−0.73(19.03)	−0.54(3.86)	−0.72(7.20)	−2.54(8.42)
MN	−0.48(6.31)	−38.24(44.4)	82.58(64.56)	0.13(3.78)	−51.07(40.48)	109.59(71.1)
ML	19.19(17.25)	46.87(32.11)	29.21(95.27)	−0.01(8.46)	5.37(16.65)	−3.63(36.38)
			$\pi = 0.8, \mu_2 = 1.5$			
EL	−3.74(14.02)	−3.58(15.79)	3.55(50.90)	−1.00(5.70)	−0.74(6.35)	−2.41(19.01)
MN	−12.42(16.63)	−22.01(35.80)	41.44(96.03)	−1.57(8.17)	−18.67(29.21)	116.35(94.33)
ML	−1.71(12.38)	13.9(25.24)	−54.54(46.13)	−4.63(7.52)	7.89(7.06)	−49.65(26.15)
			$\pi = 0.8, \mu_2 = 2$			
EL	−1(6.46)	−1.06(9.54)	−1.50(31.48)	−0.49(3.40)	−0.32(4.88)	−4.20(14.72)
MN	−4.73(10.12)	−23.08(33.81)	100.56(104.09)	−0.69(5.67)	−23.85(28.05)	169.97(103.51)
ML	−6.44(9.50)	8.54(10.14)	−66.54(40.47)	−6.77(5.18)	−0.15(30.64)	−49.79(24.33)

However, when data are generated from non-normal mixtures, the ML estimators lose their optimality. From Tables 2–3, we can see that compared with the MN estimators, they have smaller absolute biases in some cases, but larger standard deviations in other cases. The proposed EL estimators perform reasonably well as they have uniformly smaller biases and standard deviations than the other two competitors.

If the mixing proportion is of primary interest, we see that when the multivariate normal mixture is correctly specified, the ML estimator again performs the best and the EL estimator has almost the

Table 3. Biases (%) and standard deviations (%) (in parentheses) of different estimators based on 1,000 simulations with $n = 400$. Data were generated from the multivariate mixture model with f_1 and f_2 being $\chi^2_{\mu_1}$ and $\chi^2_{\mu_2}$, respectively. Here $\mu_1 = 5$, $\mu_2 = 10$ or 20 and $d = 3$ or 6.

	$d=3$			$d=6$		
Method	π	μ_1	μ_2	π	μ_1	μ_2
			$\pi = 0.2, \mu_2 = 10$			
EL	1.72(9.81)	−9.58(106.79)	5.19(37.99)	−0.22(4.61)	−6.59(45.73)	−1.33(18.89)
MN	5.15(11.12)	41.52(125.27)	57.42(40.56)	0.53(4.9)	−14.54(52.22)	63.59(23.5)
ML	3.79(12.44)	−3.66(134.41)	13.29(52.78)	−2.27(4.25)	−22.54(57.72)	−10.11(21.07)
			$\pi = 0.2, \mu_2 = 20$			
EL	0.03(3.06)	1.94(92.73)	−1.27(27.15)	−0.17(2.10)	−0.51(25.73)	−1.22(17.57)
MN	0.27(3.41)	−1.38(91.57)	63.70(35.50)	−0.03(3.37)	−10.19(70.36)	75.51(29.02)
ML	−1.28(3.28)	−25.91(56.89)	−17.70(47.12)	−0.48(2.79)	−5.06(34.02)	−5.14(34.02)
			$\pi = 0.5, \mu_2 = 10$			
EL	−0.24(7.99)	−2.89(38.92)	1.2(48.92)	−0.51(4.19)	−2.26(18.18)	−3.29(24.06)
MN	−0.54(7.31)	7.22(60.73)	71.58(71.78)	0.02(3.33)	0.78(27.14)	96.32(37.29)
ML	−10.06(8.06)	−56.55(59.13)	−48.5(43.21)	−8.14(9.39)	−36.58(59.29)	−45.44(46.36)
			$\pi = 0.5, \mu_2 = 20$			
EL	0.08(2.82)	1.01(51.90)	−1.55(34.49)	−0.04(2.58)	−1.07(11.77)	−2.41(22.25)
MN	0.04(3.66)	2.25(68.84)	96.59(83.74)	−0.01(2.49)	−4.98(28.04)	125.48(47.09)
ML	−1.70(4.21)	−16.59(31.39)	−36.1(59.26)	−1.12(7.03)	−8.23(44.42)	−18.47(101.29)
			$\pi = 0.8, \mu_2 = 10$			
EL	−0.73(6.76)	−1.61(22.24)	2.94(85.03)	−0.38(3.52)	−0.95(11.57)	−7.90(41.22)
MN	−5.67(11.64)	9.08(34.32)	71.59(142.80)	−0.30(3.78)	20.31(14.44)	164.10(73.63)
ML	−20.50(8.55)	−67.61(37.71)	−153.48(48.88)	−15.45(14.64)	−43.67(59.79)	−138.41(67.07)
			$\pi = 0.8, \mu_2 = 20$			
EL	−0.14(2.16)	0.03(12.65)	−2.88(54.59)	−0.06(2.09)	0.00(8.44)	−7.86(37.19)
MN	−0.83(5.43)	1.64(24.09)	245.75(176.53)	−0.22(3.77)	2.40(15.46)	315.56(133.24)
ML	−2.40(2.19)	−13.19(11.43)	−106.07(66.16)	−2.47(13.03)	−10.13(53.34)	−48.96(194.17)

same reasonable performance. Both of them perform better than the MN estimator. When the model is misspecified, the EL estimator has the best performance followed by the MN estimator. These two estimators usually win the ML estimator by a large amount. For example, in Table 2, when $\pi = 0.5$, $\mu_2 = 1.5$, and $d = 3$, all three estimators for π have similar standard deviations, however, the ML estimator has a much larger absolute bias (0.3044) compared with the EL estimator (0.0056), and the MN estimator (0.0042).

When the data dimension d increases from 3 to 6, the standard deviations of both the EL and MN estimators are getting smaller but they have different performances in bias. The absolute biases of the EL estimators are always getting smaller, while those of the ML and MN estimators are not the case. For example, in Table 3, when $\pi = 0.2$ and $\mu_2 = 10$, the absolute bias of the MN estimator for μ_2 increases from 0.5742 to 0.6359 and that of the ML estimator for μ_1 increases from 0.0366 to 0.2254. By contrast, that of the EL estimators for both (μ_1, μ_2) decreases from $(0.0958, 0.0519)$ to $(0.0659, 0.0133)$.

Overall, the EL method exhibits more robust performance than the MN and ML methods for different model specifications. When the normal mixture is correctly specified, the proposed EL estimators have comparable performance as the ML estimators. When the normal mixture is misspecified, the EL estimators perform uniformly better than the other two competitors.

3.2. Data analysis

Reaction time (RT) task is one of the most common experimental methods in psychology to study individual differences. In this section, we apply our proposed empirical likelihood method to a RT data set which was analysed by Cruz-Medina et al. (2004). In this experiment, 197 nine-year-old children were tested on mental rotation task in which a target figure was presented on the left and another one on the right. Children thus had to determine whether the second figure was identical to the first or simply a mirror image instead. The RT was recorded in milliseconds. There were 6 trials, and we considered these trials as $d = 6$ repeated measurements. The time delays between trials were randomly chosen so that children would unable to anticipate the length of delays. The subsequent trials were then expected or assumed to be independent. We display only the histogram of the first measurement of the data in Figure 1; those for the rest are similar. Cruz-Medina et al. (2004) suggested using a two-component mixture to fit the heterogeneous RT distribution.

Since recorded in milliseconds, the RT values range from around 700 to 7000. For convenience, we re-scale them in seconds; the resulting numbers are no greater than 10. Although the mixing proportion π is of primary interest, we calculate the EL, MN and ML estimators for all the three parameters π, μ_1 and μ_2. The results are tabulated in Table 4. Based on these point estimates, we also provide 95% Wald interval estimates

Table 4. Point and interval estimates of the EL, MN and ML methods for π, μ_1 and μ_2. EL_0: EL with $K = \binom{6}{3} = 20$; EL_1, EL_2, EL_3: EL with $K = 8$; MN_1: MN with cut points $c_1, \ldots, c_{10}$ being the deciles of the empirical distribution, which was suggested by Cruz-Medina et al. (2004) for general use; MN_2: MN with cut points $(c_1, \ldots, c_{10}) = (0.5, 1, 1.2, 1.4, 1.6, 2, 2.5, 3, 4, 5)$, which was used by Cruz-Medina et al. (2004) when they analysed this dataset.

Parameter	EL_0	EL_1	EL_2	EL_3	MN_1	MN_2	ML
				Point estimates			
π	0.70	0.72	0.68	0.70	0.52	0.59	0.60
μ_1	1.68	1.70	1.66	1.67	1.58	1.64	1.64
μ_2	2.90	2.95	2.87	2.91	2.79	2.67	2.64
				95% Interval estimates			
π	[0.58, 0.82]	[0.59, 0.84]	[0.55, 0.82]	[0.57, 0.83]	[0.34, 0.70]	[0.39, 0.78]	[0.42, 0.77]
μ_1	[1.57, 1.79]	[1.59, 1.82]	[1.53, 1.79]	[1.56, 1.79]	[1.39, 1.76]	[1.47, 1.81]	[1.49, 1.78]
μ_2	[2.58, 3.22]	[2.61, 3.28]	[2.53, 3.22]	[2.57, 3.25]	[2.46, 3.13]	[2.35, 2.98]	[2.35, 2.94]

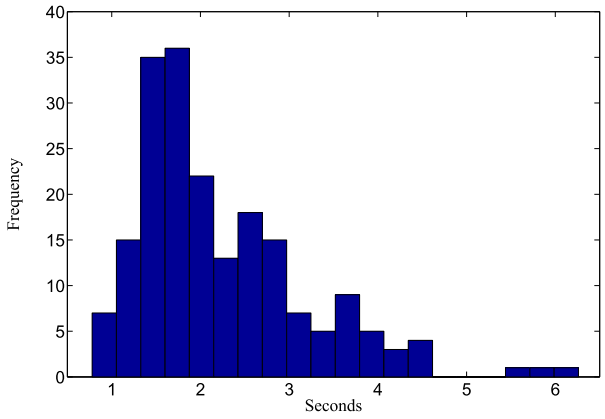

Figure 1. Histogram of the first measurement of the RT data.

for all the three parameters with variances estimated by 200 bootstrap repetitions.

As mentioned in Section 2.2, the EL estimator depends on the K randomly selected sets Ω_k^* $(k = 1, 2, \ldots, K)$. Therefore, we shall obtain different EL estimates in general when applying the EL method more than one time if $K < \binom{d}{3}$. We apply the EL method with $K = 8$ three times, and denote the results by EL_1, EL_2 and EL_3, respectively. In this example, $d = 6$. When $K = \binom{6}{3} = 20$, the results are denoted by EL_0. We see that the EL estimates with $K = 8$ are very close to those with $K = 20$. This confirms that the proposed random selection strategy works very well. The EL proportion estimates are all around 0.7, and the EL estimates for μ_1 and μ_2 are around 1.6 and 2.9, respectively.

When applying the MN method, we need to determine the cut points c_i's. For general use, Cruz-Medina et al. (2004) suggested using 10 cut points and choosing $c_1, \ldots, c_{10}$ to be the deciles of the empirical distribution of the data. The resulting MN method, denoted by MN_1, is also the MN method compared in our simulation study. When analysing the RT data, Cruz-Medina et al. (2004) used $(c_1, \ldots, c_{10}) = (0.5, 1, 1.2, 1.4, 1.6, 2, 2.5, 3, 4, 5)$. We denote the resulting MN method by MN_2. It seems that the MN results depend to some extent on the choice of cutting points, because the MN_1 proportion estimate 0.52 is quite different from that of MN_2 0.59. In the meantime, the MN_2 point and interval estimates are both nearly equal to those of the ML method.

According to our simulation studies, the EL method exhibits more robust performance than the MN and ML methods. This indicates that the EL analysis results are more trustworthy than those of the other two methods.

4. Discussions

In this paper, we proposed an empirical likelihood-based estimation method for the parameters of a multivariate two-component mixture model. We discussed three-variate mixtures in detail and extended the methodology to high-dimensional mixtures by giving a permutation-like method which reduces the high-dimensional problem to a three-dimensional situation. The performance and efficiency of the method are demonstrated through a real data example as well as simulation studies. The simulation results show that the proposed method is quite efficient in comparison to both completely parametric and almost nonparametric methods in the literature. Furthermore, the proposed method can accommodate parameter estimation in high-dimensional mixtures by requiring estimation only in three dimensions.

The extension of our approach to mixtures with more than two components is valuable and interesting. Similar to the two-component mixture situation, one can use a set of moment conditions implied by the mixture model to identify and estimate mixing proportions and other component parameters. When the number of components grows, the number of unknown parameters increases. The improvement in the performance of the proposed approach in terms of better identification and higher efficiency may crucially depend on the choice of the set of moment conditions. We will consider it in future research.

Acknowledgements

The authors would like to thank the editor, the AE, and the referee for their insightful comments and suggestions. The authors would like to thank Dr Jing Qin for valuable discussions and many helpful comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants (RGPIN-2018-05846, RGPIN-2018-05981), the National Natural Science Foundation of China (Grant Numbers 11771144, 11501354 and 11501208), and the Chinese 111 Project (B14019).

Notes on contributors

Yuejiao Fu is an Associate Professor of Statistics in the Department of Mathematics and Statistics at York University, Canada. She received her PhD in Statistics in 2004 from the University of Waterloo. Her research interests include mixture models, empirical likelihood, and statistical genetics.

Yukun Liu is a Professor in the School of Statistics, Faculty of Economic and Management, East China Normal University, China. He received his PhD in Statistics in 2009 from Nankai University, China. His research interests include nonparametric and semiparametric statistics based on empirical likelihood and their applications in case-control data, capture-recapture data, selection biased data, and finite mixture models.

Hsiao-Hsuan Wang received her PhD in Statistics in 2010 from York University, Canada. She is now a director in Model Quantification, Enterprise Risk Management, CIBC, Canada.

Xiaogang Wang is a Professor in Statistics in the Department of Mathematics and Statistics of York University. He is also holding an adjunct position as a senior research fellow at the Institute of Data Science of Tsinghua University in Beijing. He received his PhD in Statistics from the University of British Columbia in 2001. His current research is on statistical analysis of complex data in health and life sciences.

ORCID

Yuejiao Fu http://orcid.org/0000-0001-8606-570X

References

Cruz-Medina, I. R., Hettmansperger, T. P., & Thomas, H. (2004). Semiparametric mixture models and repeated measures: The multinomial cut point model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*, 463–474.

Hall, P., & Zhou, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics*, *31*, 201–224.

Hettmansperger, T. P., & Thomas, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society. Series B*, *62*, 811–825.

Kasahara, H., & Shimotsu, K. (2014). Nonparametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society. Series B*, *76*(1), 97–111.

Lindsay, B. G. (1995). *Mixture models: Theory, geometry and applications*. Hayward: Institute for Mathematical Statistics.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*, 237–249.

Owen, A. B. (2001). *Empirical likelihood*. New York: Chapman & Hall/CRC.

Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*, 300–325.

Thomas, H., & Horton, J. J. (1997). Competency criteria and the class inclusion task: Modeling judgments and justifications. *Developmental Psychology*, *33*, 1060–1073.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

Appendix

Since both Lemma 2.1 and Theorem 2.2 are established conditionally on the K selected sets Ω_k^* $(k = 1, 2, \ldots, K)$, for convenience we regard the K selected sets as fixed sets throughout the proofs. Note that $\boldsymbol{u}_{ki}$'s are i.i.d. random vectors for fixed k and varying i, while they are not independent for fixed i and varying k.

Proof of Lemma 2.1***:*** We consider $\boldsymbol{\theta} \in \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = n^{-1/3}\}$, which can be rewritten as $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + n^{-1/3}\boldsymbol{v}$ with $\|\boldsymbol{v}\| = 1$. From Qin and Lawless (1994), we can show that $\|\boldsymbol{\lambda}_k\| = O(n^{-1/3})$ and

$$\boldsymbol{\lambda}_k(\boldsymbol{\theta}) = \left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\right\}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\right\} + o(n^{-1/3}) \quad \text{(a.s.)}$$

uniformly about $\boldsymbol{\theta} \in \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le n^{-1/3}\}$, for each $k = 1, \ldots, K$. By Taylor's expansion, we have

$$\begin{aligned}
-\ell(\boldsymbol{\theta}) &= \sum_{k=1}^{K}\sum_{i=1}^{n}\log\{1 + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\} \\
&= \frac{n}{2}\sum_{k=1}^{K}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\right]^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\right]^{-1} \\
&\quad \times \left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\right] + o(n^{1/3}) \quad \text{(a.s.)} \\
&= \frac{n}{2}\sum_{k=1}^{K}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta}_0) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\boldsymbol{v}n^{-1/3}\right]^{\mathrm{T}} \\
&\quad \times \left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\right]^{-1} \\
&\quad \times \left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta}_0) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\boldsymbol{v}n^{-1/3}\right] \\
&\quad + o(n^{1/3}) \quad \text{(a.s.)} \\
&= \frac{nK}{2}\left[O(n^{-1/2}(\log\log n)^{1/2}) + \mathbb{E}\left(\frac{\partial\boldsymbol{g}(\boldsymbol{u},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)\boldsymbol{v}n^{-1/3}\right]^{\mathrm{T}} \\
&\quad \times \left[\mathbb{E}(\boldsymbol{g}(\boldsymbol{u},\boldsymbol{\theta}_0)\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u},\boldsymbol{\theta}_0))\right]^{-1} \\
&\quad \times \left[O(n^{-1/2}(\log\log n)^{1/2}) + \mathbb{E}\left(\frac{\partial\boldsymbol{g}(\boldsymbol{u},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)\boldsymbol{v}n^{-1/3}\right] \\
&\quad + o(n^{1/3}) \quad \text{(a.s.)} \\
&\ge (c/2)n^{1/3}, \quad \text{(a.s.)},
\end{aligned}$$

where c is the smallest eigenvalue of

$$\mathbb{E}\left(\frac{\partial\boldsymbol{g}(\boldsymbol{u},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\left[\mathbb{E}(\boldsymbol{g}(\boldsymbol{u},\boldsymbol{\theta}_0)\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u},\boldsymbol{\theta}_0))\right]^{-1}\mathbb{E}\left(\frac{\partial\boldsymbol{g}(\boldsymbol{u},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right).$$

Similarly,

$$\begin{aligned}-\ell(\boldsymbol{\theta}_0) &= \frac{n}{2}\sum_{k=1}^{K}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta}_0)\right]^{\mathrm{T}} \\ &\quad\times\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta}_0)\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{ki},\boldsymbol{\theta}_0)\right]^{-1} \\ &\quad\times\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta}_0)\right]+o(1)\quad\text{(a.s.)} \\ &= O(\log\log n).\quad\text{(a.s.)}\end{aligned}$$

Since $\ell(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ when $\boldsymbol{\theta}$ belongs to the ball $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le n^{-1/3}$, as n is large, $\ell(\boldsymbol{\theta})$ must have a maximum point $\widehat{\boldsymbol{\theta}}$ in the interior of this ball such that

$$\begin{aligned}&\left.\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= -\sum_{k=1}^{K}\sum_{i=1}^{n}\left.\frac{(\partial\boldsymbol{\lambda}_k^{\mathrm{T}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})+(\partial\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})/\partial\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\lambda}_k(\boldsymbol{\theta})}{1+\boldsymbol{\lambda}_k^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= -\sum_{k=1}^{K}\sum_{i=1}^{n}\left.\frac{1}{1+\boldsymbol{\lambda}_k^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})}\left(\frac{\partial\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\boldsymbol{\lambda}_k(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \mathbf{0}.\end{aligned}$$

■

Proof of Theorem 2.2**:** Taking derivatives about $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}^{\mathrm{T}}$, we have

$$\begin{aligned}\frac{\partial\boldsymbol{Q}_{kn}(\boldsymbol{\theta},\mathbf{0})}{\partial\boldsymbol{\theta}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},\quad \frac{\partial\boldsymbol{Q}_{kn}(\boldsymbol{\theta},\mathbf{0})}{\partial\boldsymbol{\lambda}_j^{\mathrm{T}}} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{ji},\boldsymbol{\theta})\delta_{kj}, \\ \frac{\partial\boldsymbol{Q}_{0n}(\boldsymbol{\theta},\mathbf{0})}{\partial\boldsymbol{\theta}} &= 0,\quad \frac{\partial\boldsymbol{Q}_{0n}(\boldsymbol{\theta},\mathbf{0})}{\partial\boldsymbol{\lambda}_k^{\mathrm{T}}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\boldsymbol{g}(\boldsymbol{u}_{ki},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}},\end{aligned}$$

for $k,j=1,\ldots,K$, and δ_{kj} is the Kronecker delta. Expanding $\boldsymbol{Q}_{kn}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\lambda}})$ and $\boldsymbol{Q}_{0n}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\lambda}})$ at $(\boldsymbol{\theta}_0,\mathbf{0})$, we have

$$\begin{aligned}\mathbf{0} = \boldsymbol{Q}_{kn}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\lambda}}_k) &= \boldsymbol{Q}_{kn}(\boldsymbol{\theta}_0,\mathbf{0})+\frac{\partial\boldsymbol{Q}_{kn}(\boldsymbol{\theta}_0,\mathbf{0})}{\partial\boldsymbol{\lambda}_k^{\mathrm{T}}}(\hat{\boldsymbol{\lambda}}_k-\mathbf{0}) \\ &\quad+\frac{\partial\boldsymbol{Q}_{kn}(\boldsymbol{\theta}_0,\mathbf{0})}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+o_p(\delta_n), \\ \mathbf{0} = \boldsymbol{Q}_{0n}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\lambda}}) &= \boldsymbol{Q}_{0n}(\boldsymbol{\theta}_0,\mathbf{0})+\sum_{k=1}^{K}\frac{\partial\boldsymbol{Q}_{0n}(\boldsymbol{\theta}_0,\mathbf{0})}{\partial\boldsymbol{\lambda}_k^{\mathrm{T}}}(\hat{\boldsymbol{\lambda}}_k-\mathbf{0}) \\ &\quad+\frac{\partial\boldsymbol{Q}_{0n}(\boldsymbol{\theta}_0,\mathbf{0})}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+o_p(\delta_n),\end{aligned}$$

where $\delta_n=\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|+\sum_{k=1}^{K}\|\hat{\boldsymbol{\lambda}}_k\|$.

It follows from the above equations that

$$\begin{pmatrix}\hat{\boldsymbol{\lambda}}\\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\end{pmatrix}=\boldsymbol{S}_n^{-1}\begin{pmatrix}\boldsymbol{D}_n\\ \mathbf{0}\end{pmatrix}+o_p(\delta_n).\qquad\text{(A1)}$$

Here

$$\boldsymbol{D}_n=\begin{pmatrix}\boldsymbol{Q}_{1n}(\boldsymbol{\theta}_0,\mathbf{0})\\ \vdots\\ \boldsymbol{Q}_{Kn}(\boldsymbol{\theta}_0,\mathbf{0})\end{pmatrix},\quad \boldsymbol{S}_n=\begin{pmatrix}\boldsymbol{S}_{11n} & \boldsymbol{S}_{12n}\\ \boldsymbol{S}_{21n} & \boldsymbol{S}_{22n}\end{pmatrix},$$

where

$$\begin{aligned}\boldsymbol{S}_{11n} &= \left(-\frac{\partial\boldsymbol{Q}_{kn}(\boldsymbol{\theta}_0,\mathbf{0})}{\partial\boldsymbol{\lambda}_j^{\mathrm{T}}}\right)_{1\le j,\,k\le K} \\ &= \operatorname{diag}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{1i},\boldsymbol{\theta}_0)\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{1i},\boldsymbol{\theta}_0),\ldots,\right. \\ &\qquad\left.\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{u}_{Ki},\boldsymbol{\theta}_0)\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{Ki},\boldsymbol{\theta}_0)\right),\end{aligned}$$

$$\begin{aligned}\boldsymbol{S}_{12n} &= -\left(\frac{\partial\boldsymbol{Q}_{1n}(\boldsymbol{\theta}_0,\mathbf{0})}{\partial\boldsymbol{\theta}},\ldots,\frac{\partial\boldsymbol{Q}_{Kn}(\boldsymbol{\theta}_0,\mathbf{0})}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}} \\ &= -\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{g}(\boldsymbol{u}_{1i},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}},\ldots,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{g}(\boldsymbol{u}_{Ki},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right),\end{aligned}$$

$\boldsymbol{S}_{21n}=\boldsymbol{S}_{12n}^{\mathrm{T}}$ and $\boldsymbol{S}_{22n}=-\partial\boldsymbol{Q}_{0n}(\boldsymbol{\theta}_0,\mathbf{0})/\partial\boldsymbol{\theta}=0$.

Define $\boldsymbol{S}_{11}=I_K\otimes S_{11}$ and $\boldsymbol{S}_{12}=\mathbf{1}_K\otimes S_{12}$, where $\otimes$ is the Kronecker product operator. Under the conditions of Theorem 2.2, as $n\to\infty$, it can be verified that

$$\boldsymbol{S}_{11n}=\boldsymbol{S}_{11}+o_p(1),\quad \boldsymbol{S}_{12n}=\boldsymbol{S}_{12}+o_p(1),$$

and therefore $\boldsymbol{S}_n=\boldsymbol{S}+o_p(1)$, where

$$\boldsymbol{S}=\begin{pmatrix}\boldsymbol{S}_{11} & \boldsymbol{S}_{12}\\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22}\end{pmatrix}=\begin{pmatrix}I_K\otimes S_{11} & \mathbf{1}_K\otimes S_{12}\\ \mathbf{1}_K^{\mathrm{T}}\otimes S_{12}^{\mathrm{T}} & \mathbf{0}\end{pmatrix}.$$

In addition, $\sqrt{n}\boldsymbol{D}_n$ converges in distribution to $N(\mathbf{0},\boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma}=\left(\mathbb{E}\{\boldsymbol{g}(\boldsymbol{u}_{k1})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{j1})|\Omega^*\}\right)_{1\le k,j\le K}.$$

Therefore, $\delta_n=O_p(n^{-1/2})$. Since the inverse of $\boldsymbol{S}$ is

$$\boldsymbol{S}^{-1}=\begin{pmatrix}\boldsymbol{S}_{11}^{-1}+\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22.1}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1} & -\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22.1}^{-1}\\ -\boldsymbol{S}_{22.1}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1} & \boldsymbol{S}_{22.1}^{-1}\end{pmatrix},$$

where $\boldsymbol{S}_{22.1}=-\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}$, we further have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\boldsymbol{S}_{22.1}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\cdot\sqrt{n}\boldsymbol{D}_n,$$

which converges in distribution to $N(0,\boldsymbol{V}_2)$ with

$$\boldsymbol{V}_2=\boldsymbol{S}_{22.1}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22.1}^{-1}.\qquad\text{(A2)}$$

With some algebra, it can be seen that $\boldsymbol{S}_{22.1}=-KS_{21}S_{11}^{-1}S_{12}$ and $\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}=\mathbf{1}_K^{\mathrm{T}}\otimes(S_{21}S_{11}^{-1})$, which implies

$$\begin{aligned}\boldsymbol{V}_2 &= K^{-2}(S_{21}S_{11}^{-1}S_{12})^{-1}\mathbf{1}_K^{\mathrm{T}}\otimes(S_{21}S_{11}^{-1}) \\ &\quad\boldsymbol{\Sigma}\mathbf{1}_K\otimes(S_{11}^{-1}S_{12})(S_{21}S_{11}^{-1}S_{12})^{-1} \\ &= K^{-2}(S_{21}S_{11}^{-1}S_{12})^{-1}(S_{21}S_{11}^{-1}) \\ &\quad\sum_{k,j=1}^{K}\mathbb{E}\{\boldsymbol{g}(\boldsymbol{u}_{k1})\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{u}_{j1})|\Omega^*\}(S_{11}^{-1}S_{12})(S_{21}S_{11}^{-1}S_{12})^{-1} \\ &= K^{-1}(S_{21}S_{11}^{-1}S_{12})^{-1}+\frac{K-1}{K}(S_{21}S_{11}^{-1}S_{12})^{-1}(S_{21}S_{11}^{-1}) \\ &\quad\Sigma_{\mathrm{off}}(S_{11}^{-1}S_{12})(S_{21}S_{11}^{-1}S_{12})^{-1}.\end{aligned}$$

This finishes the proof of Theorem 2.2. ■