Statistics and Its Interface Volume 7 (2014) 211–217

Inference for multivariate mixtures of two unknown symmetric components

Wenxiu Ge, Xiaobo Guo, Xueqin Wang*, and Heping Zhang

Modeling heterogeneity for multivariate data is an important research topic. In this paper, we give a sufficient condition to establish the identifiability for semiparametric multivariate mixture models with unknown location-shifted symmetric components, and propose a novel minimum distance method to estimate the location and proportion parameters. Strong consistency and asymptotic normality of our estimators under some regularity conditions are established. Simulation studies show that the proposed method is robust to misspecified component distributions. The Old Faithful data is also used as a real benchmark to assess the performance of the proposed method.

Keywords and phrases: E-distance, Identifiability, Multivariate symmetry, Semiparametric mixtures, V-process.

1. INTRODUCTION

Parametric mixture models are common approaches to dealing with the unobserved heterogeneity in the datasets, and have been widely employed in diverse areas such as biometrics, genetics, medicine, economics and finance [6, 14, 15, 18, 23]. However, as [9] stated, the specification of components' distributions is necessary yet a great trouble, and seldom well-established theories exist to guide the selection of components' distributions based on the observed data.

In recent years, nonparametric (semiparametric) mixture models have emerged as efficient approaches to characterizing the heterogenous datasets [3, 4, 7–10, 13]. Unlike the classical parametric mixture model, nonparametric mixture models are generally not identifiable without additional information of components' distributions. To avoid the identifiability issue, the assumption of *conditional independence* is usually imposed on the component distributions. Specifically, it assumes that the coordinates of the observed d-dimensional vector are independent conditional on the component from which the observed data are drawn; that is

$$G(\boldsymbol{x}) = \sum_{j=1}^{m} \lambda_j \prod_{l=1}^{d} F_{jl}(x_l), \quad \forall\, \boldsymbol{x} \in \mathbb{R}^d, \tag{1}$$

*Corresponding author.

where $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$, $G(\cdot)$ denotes cumulative distribution function (CDF) of the d-variate random vector, and F_{jl} is the univariate CDF of the lth variate of the jth component. Furthermore, the mixing proportions $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ satisfy $\sum_{j=1}^{m} \lambda_j = 1, \lambda_j \geq 0$. [8] showed that when $m = 2$, the model identifiability only holds for $d \geq 3$. [1] gave a general result for more than two components' case and showed that the parameters in model (1) are identifiable for $d \geq 3$ if the component density functions $f_{1l}, \ldots, f_{ml}$ are linearly independent except possibly on a set of Lebesgue measure zero.

In this article, we employ a slightly different approach by taking into account the joint distribution of each component in the mixture model rather than imposing the conditionally independent restriction on the d-variate component distribution, and show that the mixture model is identifiable if each component distribution is symmetric with respect to some fixed locations. Similar ideas can be traced back to [3] and [10] who studied independently the following univariate, symmetric, location-shifted, semiparametric, mixture model,

$$G(x) = \sum_{j=1}^{m} \lambda_j F(x - \mu_j), \quad \forall\, x \in \mathbb{R}, \tag{2}$$

where $F(\cdot)$ denotes an unspecified CDF of a component which is symmetric about zero, namely, $F(x) = 1 - F(-x)$ for all continuity points x of F. However, [3] and [10] focused on only the univariate mixture component, i.e., $d = 1$. To our best knowledge, the case of $d \geq 2$ has not been well studied.

The identifiability of our introduced semiparametric multivariate mixture model relies on the assumption of symmetric component distribution. To utilize the symmetry, we develop a novel E-distance [21] based method to estimate the parameters $\boldsymbol{\theta} = (\lambda, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ defined in Section 2 in the mixture model. This method exploits the symmetry of the component distributions as well as the dependence information of multivariate data. We shall demonstrate its higher efficiency through simulation studies and real data analysis.

The rest of this paper is organized as follows. In Section 2, we introduce the model and address its identifiability. In Section 3, we present the distance-based method for estimating the parameters, as well as the asymptotic properties. We assess the performance of the proposed estimators by numerical studies and a real data analysis in Section 4. Detailed proofs of the theoretical results are deferred to Appendices A, B and C.

2. MODEL AND IDENTIFIABILITY

Considering the following m-term mixture distribution,

$$G(\boldsymbol{x};\boldsymbol{\lambda},\boldsymbol{\mu},F)=\sum_{j=1}^{m}\lambda_j F(\boldsymbol{x}-\boldsymbol{\mu}_j),\quad \forall\,\boldsymbol{x}\in\mathbb{R}^d, \tag{3}$$

where λ_j denotes the mixture proportion of the jth component which satisfies $\sum_{j=1}^{m}\lambda_j=1,\lambda_j\geq 0$, $F(\cdot)$ is a symmetric multivariate distribution function with origin $\mathbf{0}$, and $\boldsymbol{\mu}_j=(\mu_{j1},\ldots,\mu_{jd})^\top$ is the location vector for the jth component $(j=1,\ldots,m)$. In this paper, we assume that m is fixed and $m=2$ unless stated otherwise. Let $\mathcal{F}$ denote the set of all CDFs $F(\cdot)$ which are symmetric about the origin $\mathbf{0}$. Then, the parameter space of finite mixture model (3) is $\boldsymbol{\Theta}\times\mathcal{F}$, where $\boldsymbol{\Theta}=\{\boldsymbol{\theta}=(\lambda,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2):0<\lambda<1/2,(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)\in\mathbb{R}^{2d}\setminus\Delta\}$, $\Delta=\{(\boldsymbol{x},\boldsymbol{x}),\boldsymbol{x}\in\mathbb{R}^d\}$ and the constraint $0<\lambda<1/2$ is used to avoid the "label switching" problem [17].

For clarity, we rewrite the mixture distribution (3) by using its characteristic function. Let $\phi_G(\cdot)$ be the characteristic function of mixture distribution (3), then we have

$$\begin{aligned}\phi_G(\boldsymbol{t})&=E\big\{\exp\big(i\boldsymbol{t}^\top\boldsymbol{X}\big)\big\}\\&=\int_{R^d}\exp\big(i\boldsymbol{t}^\top\boldsymbol{x}\big)dG(\boldsymbol{x})\\&=\big\{\lambda\exp\big(i\boldsymbol{t}^\top\boldsymbol{\mu}_1\big)+(1-\lambda)\exp\big(i\boldsymbol{t}^\top\boldsymbol{\mu}_2\big)\big\}\\&\quad\times\int_{R^d}\exp\big(i\boldsymbol{t}^\top\boldsymbol{x}\big)dF(\boldsymbol{x})\\&\triangleq\phi(\boldsymbol{t};\boldsymbol{\theta})\times\phi_F(\boldsymbol{t}),\quad\forall\,\boldsymbol{t}\in\mathbb{R}^d,\end{aligned}\tag{4}$$

where $\phi(\boldsymbol{t};\boldsymbol{\theta})=\lambda\exp(i\boldsymbol{t}^\top\boldsymbol{\mu}_1)+(1-\lambda)\exp(i\boldsymbol{t}^\top\boldsymbol{\mu}_2)$, and $\phi_F(\boldsymbol{t})=\int_{R^d}\exp(i\boldsymbol{t}^\top\boldsymbol{x})dF(\boldsymbol{x})$ is the characteristic function of symmetric multivariate distribution $F(\cdot)$.

The identifiability of the model (3) means that no two different parameter vectors $(\boldsymbol{\theta},F)$ and $(\boldsymbol{\theta}',F')$ in $\boldsymbol{\Theta}\times\mathcal{F}$ satisfy

$$\phi(\boldsymbol{t};\boldsymbol{\theta})\times\phi_F(\boldsymbol{t})=\phi(\boldsymbol{t};\boldsymbol{\theta}')\times\phi_{F'}(\boldsymbol{t}),\quad\forall\boldsymbol{t}\in\mathbb{R}^d. \tag{5}$$

The following theorem shows that the proposed two-component mixture model (3) is identifiable.

Theorem 2.1. *Assume that there are two parameter vectors $(\boldsymbol{\theta},F)$ and $(\boldsymbol{\theta}',F')$ in $\boldsymbol{\Theta}\times\mathcal{F}$ satisfying (5), then $(\boldsymbol{\theta},F)=(\boldsymbol{\theta}',F')$.*

The proof of this theorem is given in Appendix A.

3. ESTIMATION PROCEDURE

3.1 E-distance estimator

In this subsection, we introduce the E-distance based method to estimate the parameters $\boldsymbol{\theta}=(\lambda,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)$. First, we have:

Theorem 3.1. *If the mixture model (3) is identifiable, then there is a unique $\boldsymbol{\theta}=(\lambda,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)\in\boldsymbol{\Theta}$ such that $\phi(\boldsymbol{t};\lambda,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)\times\phi(\boldsymbol{t};\lambda,-\boldsymbol{\mu}_1,-\boldsymbol{\mu}_2)$ is a real value function.*

The proof of Theorem 3.1 is similar to that of Theorem 1 in [10]. Theorem 3.1 implies that the multivariate symmetric distribution corresponding to $\phi(\boldsymbol{t};\lambda,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)\times\phi(\boldsymbol{t};\lambda,-\boldsymbol{\mu}_1,-\boldsymbol{\mu}_2)$ is unique. Next, we construct a random vector $\boldsymbol{W}$ which is the key to estimate the parameters. Suppose the random vector $\boldsymbol{U}$ has the mixture distribution G_0 (3) with the true parametric vector $(\boldsymbol{\theta}_0,F_0)$, and let $\boldsymbol{V}$ be a random vector with the support $(-\boldsymbol{\mu}_1,-\boldsymbol{\mu}_2)$ and the weights $(\lambda,\ 1-\lambda)$. The random vector $\boldsymbol{W}=\boldsymbol{U}+\boldsymbol{V}$ is centrally symmetric about the origin $\mathbf{0}$ when $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, since the characteristic function of $\boldsymbol{W}$,

$$\begin{aligned}&\phi_{G_0}(\boldsymbol{t})\times\phi(\boldsymbol{t};\lambda,-\boldsymbol{\mu}_1,-\boldsymbol{\mu}_2)\\&=\phi_{F_0}(\boldsymbol{t})\times\big\{\phi(\boldsymbol{t};\boldsymbol{\theta}_0)\times\phi(\boldsymbol{t};\lambda,-\boldsymbol{\mu}_1,-\boldsymbol{\mu}_2)\big\},\end{aligned}$$

is a real function if $\boldsymbol{\theta}=\boldsymbol{\theta}_0$.

To exploit the symmetry of $\boldsymbol{W}$, we use the concept of E-distance between two random vectors developed in [21]. Suppose that $\boldsymbol{X}'$ and $\boldsymbol{Y}'$ are i.i.d. random vectors in $\mathbb{R}^d$ corresponding to $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. $\|\cdot\|$ is the Euclidean norm and $\stackrel{\mathrm{d}}{=}$ means that two random variables (or vectors) are identically distributed. We define $\mathcal{D}(\boldsymbol{X},\boldsymbol{Y})\triangleq 2E\|\boldsymbol{X}-\boldsymbol{Y}\|-E\|\boldsymbol{X}-\boldsymbol{X}'\|-E\|\boldsymbol{Y}-\boldsymbol{Y}'\|$, then $\mathcal{D}(\boldsymbol{X},\boldsymbol{Y})$ is a distance. The proof can be found in [21]. The following lemma summaries some properties of the E-distance helping us estimate the parameters:

Lemma 3.2. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two independent random vectors in $\mathbb{R}^d$, $E\|\boldsymbol{X}\|<\infty$ and $E\|\boldsymbol{Y}\|<\infty$. Then,*
(i) $\mathcal{D}(\boldsymbol{X},\boldsymbol{Y})\geq 0$;
(ii) $\mathcal{D}(\boldsymbol{X},\boldsymbol{Y})=0$ if and only if $\boldsymbol{X}\stackrel{\mathrm{d}}{=}\boldsymbol{Y}$.

Now, let $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ be i.i.d. random vectors and have the same distribution as $\boldsymbol{W}$, we define

$$\begin{aligned}\mathcal{D}(\boldsymbol{\theta})&\triangleq\mathcal{D}(\boldsymbol{W},-\boldsymbol{W})\\&=E\{\|\boldsymbol{W}_1+\boldsymbol{W}_2\|-\|\boldsymbol{W}_1-\boldsymbol{W}_2\|\}\\&=E\big\{h_{\boldsymbol{\theta}}(\boldsymbol{X}_1,\boldsymbol{X}_2)\big\},\end{aligned}\tag{6}$$

where

$$\begin{aligned}&h_{\boldsymbol{\theta}}(\boldsymbol{X}_1,\boldsymbol{X}_2)\\&=\lambda^2(\|\boldsymbol{X}_1+\boldsymbol{X}_2-2\boldsymbol{\mu}_1\|-\|\boldsymbol{X}_1-\boldsymbol{X}_2\|)\\&\quad+2\lambda(1-\lambda)\big(\|\boldsymbol{X}_1+\boldsymbol{X}_2-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\|\\&\qquad-\|\boldsymbol{X}_1-\boldsymbol{X}_2+(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\|\big)\\&\quad+(1-\lambda)^2(\|\boldsymbol{X}_1+\boldsymbol{X}_2-2\boldsymbol{\mu}_2\|-\|\boldsymbol{X}_1-\boldsymbol{X}_2\|),\end{aligned}\tag{7}$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are i.i.d random vectors with the same mixture distributions (3). It follows from Theorem 3.1 and Lemma 3.2 that $\mathcal{D}(\boldsymbol{\theta}) \geq 0$ and $\mathcal{D}(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Hence, we can write

$$(8) \qquad \boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{D}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\big\{h_{\boldsymbol{\theta}}(\boldsymbol{X}_1, \boldsymbol{X}_2)\big\}.$$

The corresponding V-process of $\mathcal{D}(\boldsymbol{\theta})$ can be calculated as

$$(9) \qquad \mathcal{D}_n(\boldsymbol{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Thus, $\boldsymbol{\theta}_0$ can be estimated by

$$(10) \qquad \hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

We call $\hat{\boldsymbol{\theta}}_n$ the E-distance estimator (EDE).

Note that the estimator (18) in [10] is a special setting of our EDE for $d = 1$.

3.2 Asymptotic properties

Let f be the component density function of mixture distribution (3). If f is continuous and bounded in an open neighborhood of $\boldsymbol{\theta}_0$, then the minimizer $\hat{\boldsymbol{\theta}}_n$ of $\mathcal{D}_n(\boldsymbol{\theta})$ is strongly consistent and asymptotically normal under some regularity conditions.

The following theorem presents the consistency:

Theorem 3.3. *Suppose that the covariance matrix of component in mixture model (3) is positive definite. Then $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ almost surely as $n \to \infty$.*

The proof of this theorem is given in Appendix B.

Next, we study the asymptotic normality of $\hat{\boldsymbol{\theta}}_n$. Note that $\mathcal{D}_n(\boldsymbol{\theta})$ in Equation (9) is a V-process as it is indexed by the parameter vector $\boldsymbol{\theta}$. For convenience, denote the functions

$$V^{(k)}(h) = Eh(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$$

and

$$V_n^{(k)}(h) = \frac{1}{n^k}\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n} h(\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_k}),$$

where h is some vector-valued function of k variables and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are random samples from some CDF $G_{\boldsymbol{\theta}_0}(\cdot)$. Then, we have $\mathcal{D}(\boldsymbol{\theta}) = V^{(2)}(h_{\boldsymbol{\theta}})$ and $\mathcal{D}_n(\boldsymbol{\theta}) = V_n^{(2)}(h_{\boldsymbol{\theta}})$. Note that $\mathcal{D}_n(\boldsymbol{\theta})$ is a V-statistic for fixed $\boldsymbol{\theta} \in \Theta$. According to the Hoeffding decomposition for V-statistics, we have

$$(11) \qquad \mathcal{D}_n(\boldsymbol{\theta}) - \mathcal{D}(\boldsymbol{\theta}) = 2V_n^{(1)}(\pi_1 h_{\boldsymbol{\theta}}) + V_n^{(2)}(\pi_2 h_{\boldsymbol{\theta}}),$$

where $\pi_k h_{\boldsymbol{\theta}}, k = 1, 2$ is the kth Hoeffding projection defined by

$$\pi_1 h_{\boldsymbol{\theta}}(\boldsymbol{x}_1) = (\delta_{\boldsymbol{x}_1} - G_{\boldsymbol{\theta}_0})G_{\boldsymbol{\theta}_0}h_{\boldsymbol{\theta}}$$

and

$$\pi_2 h_{\boldsymbol{\theta}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\delta_{\boldsymbol{x}_1} - G_{\boldsymbol{\theta}_0})(\delta_{\boldsymbol{x}_2} - G_{\boldsymbol{\theta}_0})h_{\boldsymbol{\theta}}$$

where $Gh = \int h dG$ and $\delta_{\boldsymbol{x}}$ denotes a point mass at $\boldsymbol{x}$.

Meanwhile, the V-statistic and U-statistic are closely related in behavior under appropriate moment conditions ([19], page 206). Therefore, we can take full advantage of the U-process theory to obtain the asymptotic normality.

Theorem 3.4. *In addition to the conditions of Theorem 3.3, we assume that the following conditions hold:*

C1. For all $\epsilon > 0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty} Pr\Big\{\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta} |nV_n^{(2)}(\pi_2 h_{\boldsymbol{\theta}} - \pi_2 h_{\boldsymbol{\theta}_0})| > \epsilon\Big\} = 0;$$

C2. For the functions defined by $r_{\boldsymbol{\theta}_0}(\boldsymbol{x}) = 0$ and

$$r_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\pi_1 h_{\boldsymbol{\theta}}(\boldsymbol{x}) - \pi_1 h_{\boldsymbol{\theta}_0}(\boldsymbol{x}) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \eta(\boldsymbol{x})}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|}, \ \boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash\{\boldsymbol{\theta}_0\}$$

and for any $\epsilon > 0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty} Pr\Big\{\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta} |n^{1/2}V_n^{(1)}(r_{\boldsymbol{\theta}})| > \epsilon\Big\} = 0.$$

Then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to N(\mathbf{0}, 4A^{-1}BA^{-1}), \quad \text{in distribution},$$

where $A = -\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\mathcal{D}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, $B = Cov\{\eta(\boldsymbol{X})\}$ and $\eta(\boldsymbol{x}) = E\frac{\partial}{\partial\boldsymbol{\theta}}\pi_1 h_{\boldsymbol{\theta}}(\boldsymbol{x})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$.

The proof of this theorem is given in Appendix C.

Note that the asymptotic covariance in Theorem 3.4 has a simple expression but becomes quite involved in computation. We adopt the bootstrap approach to evaluate the standard error of $\boldsymbol{\theta}$ in application.

4. SIMULATION AND APPLICATION

4.1 Numerical study

In this section, we will assess the finite sample behavior of our proposed method by simulation studies. We consider the following two scenarios:

1. $\lambda N(\boldsymbol{\mu}_1, \Sigma) + (1-\lambda)N(\boldsymbol{\mu}_2, \Sigma)$,
2. $\lambda t_3(\boldsymbol{\mu}_1, \Sigma) + (1-\lambda)t_3(\boldsymbol{\mu}_2, \Sigma)$.

In scenario 1, the value of mixing proportion λ is 0.3, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are set to be $(0,1)^\top$ and $(2,3)^\top$, respectively, and Σ is chosen to be the correlation matrix with the correlation coefficient ρ with values 0.05, 0.50, 0.95, respectively. In scenario 2, we follow the same parameter setting except that $\boldsymbol{\mu}_2 = (4,5)^\top$ and the covariance matrix equals to 0.5 times the correlation matrix. In the simulation, the sample size is $n = 200$ and we replicate 500 times.

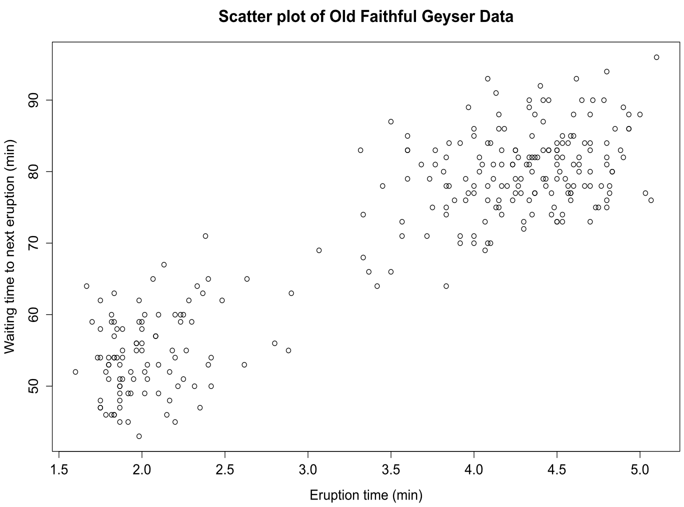

Figure 1. The scatter plot of eruption time and waiting time in Old Faithful Geyser data.

Here, we compare the performance of our proposed E-distance method with existing maximum likelihood estimates of the homoscedastic normal mixture using the EM algorithm (Norm-EM). In this paper, we use the function *NMixEM* in the R package *mixAK* [11] to obtain the Norm-EM estimates as well as the start values used in E-distance method. Two criteria, the mean of the estimates and the mean squared error (MSE), are used to measure the performance. The results are summarized in Tables 1 and 2.

From Table 1, we can observe that the difference between E-distance method and Norm-EM is ignorable if the true model is the multivariate normal mixture model. However, when the data are generated from the multivariate t-mixture model, we can observe clearly from Table 2 that the MSEs of E-distance method are much smaller than the MSEs of Norm-EM. These results indicate that the E-distance method has a comparable performance with the parametric method if the underlying model is true. However, when the underlying distribution is misspecified, the proposed semiparametric E-distance method is more robust than the parametric method.

4.2 Case study

In this subsection, we will assess our proposed method by using the benchmark data–Old Faithful Geyser in Yellowstone National Park, USA. We are especially interested in the joint distribution of two measurements: eruption time and waiting time. We should note that [10] analyzed the waiting time between eruptions as a univariate random variable.

Figure 1 presents the scatter plot of the eruption time and waiting time, revealing two subpopulations. Hence, a two-component mixture model is a reasonable choice. We applied the E-distance method and the Norm-EM to analyze the bivariate data, and Table 3 contains the results based on 200 bootstrap samples.

From Table 3, we can observe that both methods yield nearly the same sample average values and standard errors, indicating that the two-normal mixture model fits the data well. We note that the E-distance method does not require a specification of the component distribution, hence it is more flexible as also apparent in the simulation studies.

5. DISCUSSION

To the best of our knowledge, this is the first attempt to establish the identifiability of a two-component mixture model with multivariate symmetric component distributions. The idea and property of "symmetry" are valuable and have been widely adopted in theory and practice, see [5, 20, 24] and so on. We prove that the multivariate two-component mixture model is identifiable when the component distributions are symmetric about some location parameters and do not need to be conditionally independent. It is noteworthy that the symmetric component distribution is only a sufficient condition for the identifiability of a semi-parametric multivariate mixture model. It warrants further effort to establish the necessary and sufficient condition.

Interestingly, we employ the E-distance estimation method to estimate the parameters. Our method takes advantage of the symmetry in the component distributions, and hence increases the estimation efficiency. Simulations and real data analysis suggest great promise of our proposed E-distance method as opposed to an existing method.

It should be noted that we only focused on the two-component case, i.e., $m = 2$. For $m \geq 3$, the identifiability of a multivariate location-changed mixture model is a challenging topic. Imposing constraints on the proportion coefficient λ_j or location parameters $\boldsymbol{\mu}_j (j = 1, \ldots, m)$ in (3) may not address the "label switching" problem.

Through preliminary simulation studies, we find that the proposed E-distance method performs similarly to the parametric Norm-EM method when the model is correctly specified, and is more robust when the model is misspecified. This observation is useful in practice. If our proposed E-distance method performs similarly to the parametric Norm-EM method, chances are that the model assumptions are reasonably valid and so is the statistical inference. We illustrate this point with the re-analysis of the Old Faithful Geyser data.

APPENDIX A. IDENTIFIABILITY PROOF

Proof of Theorem 2.1. Suppose that there are independent random vectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ such that $\boldsymbol{Y}$ has the probability $(\lambda, 1 - \lambda)$ in the two-point support $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and $\boldsymbol{Z}$ is symmetric about the origin $\mathbf{0}$. Let $\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{Z}$, then, according to Equation (4), we have $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}}(\boldsymbol{t})\phi_{\boldsymbol{Z}}(\boldsymbol{t})$.

Table 1. Sample average of estimated values (EST) and mean squared errors (MSE) of the two estimating methods (E-distance and Norm-EM) for the two-component bivariate normal mixture model with $\boldsymbol{\mu}_1 = (0,1)^\top$, $\boldsymbol{\mu}_2 = (2,3)^\top$ and $\lambda = 0.3$

ρ	Measures	Methods	Parameters				
			μ_{11}	μ_{12}	μ_{21}	μ_{22}	λ
0.05	EST	E-distance	0.0375	1.0158	2.0114	3.0196	0.3110
		Norm-EM	0.0020	0.9933	2.0007	3.0027	0.3019
	MSE	E-distance	0.0549	0.0429	0.0164	0.0159	0.0029
		Norm-EM	0.0374	0.0322	0.0108	0.0116	0.0022
0.50	EST	E-distance	0.0453	1.0331	2.0436	3.0448	0.3241
		Norm-EM	0.0031	0.9930	2.0091	3.0068	0.3070
	MSE	E-distance	0.0761	0.0785	0.0343	0.0355	0.0067
		Norm-EM	0.0739	0.0760	0.0268	0.0235	0.0059
0.95	EST	E-distance	0.0253	1.0197	2.0608	3.0612	0.3255
		Norm-EM	0.0101	1.0091	2.0426	3.0412	0.3201
	MSE	E-distance	0.1007	0.1036	0.0455	0.0468	0.0090
		Norm-EM	0.1148	0.1147	0.0334	0.0327	0.0096

Table 2. Sample average of estimated values (EST) and mean squared errors (MSE) of the two estimating methods (E-distance and Norm-EM) for the two-component bivariate Student's-t mixture model with degrees of freedom $\nu = 3$, $\boldsymbol{\mu}_1 = (0,1)^\top$, $\boldsymbol{\mu}_2 = (4,5)^\top$ and $\lambda = 0.3$

ρ	Measures	Methods	Parameters				
			μ_{11}	μ_{12}	μ_{21}	μ_{22}	λ
0.05	EST	E-distance	−0.0027	1.0076	4.0030	4.9984	0.2993
		Norm-EM	−0.0226	0.9839	4.0258	5.0211	0.3031
	MSE	E-distance	0.0188	0.0184	0.0057	0.0051	0.0011
		Norm-EM	0.1842	0.1901	0.0187	0.0169	0.0013
0.50	EST	E-distance	0.0042	1.0103	4.0046	5.0064	0.3009
		Norm-EM	−0.0490	0.9595	4.0339	5.0339	0.3054
	MSE	E-distance	0.0245	0.0249	0.0084	0.0085	0.0014
		Norm-EM	0.7873	0.9386	0.0231	0.0240	0.0017
0.95	EST	E-distance	0.0085	1.0046	3.9925	4.9893	0.2991
		Norm-EM	−0.0849	0.9082	4.0272	5.0252	0.3059
	MSE	E-distance	0.0831	0.0791	0.0145	0.0143	0.0013
		Norm-EM	1.5099	1.4337	0.0352	0.0351	0.0020

Table 3. Sample average of the estimated values (EST) and corresponding standard errors (SE) for the two estimate method (E-distance and Norm-EM) based on the bootstrapped samples of the Old Faithful Geyser data

Measures	Methods	Parameters				
		μ_{11}	μ_{12}	μ_{21}	μ_{22}	λ
EST	E-distance	2.0553	54.5892	4.3020	80.0316	0.3561
	Norm-EM	2.0462	54.6279	4.2969	80.1044	0.3575
SE	E-distance	0.0494	0.6984	0.0386	0.4475	0.0287
	Norm-EM	0.0295	0.6586	0.0305	0.4305	0.0284

Assume that there are alternative independent random vectors $(\boldsymbol{Y}', \boldsymbol{Z}')$ like $(\boldsymbol{Y}, \boldsymbol{Z})$ satisfying $\boldsymbol{X} = \boldsymbol{Y}' + \boldsymbol{Z}'$, that is

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t})\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}'}(\boldsymbol{t})\phi_{\boldsymbol{Z}'}(\boldsymbol{t}). \tag{12}$$

Multiplying each side of Equation (12) by the complex conjugate of $\phi_{\boldsymbol{Y}'}(\boldsymbol{t})$, then we have

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t})\phi_{-\boldsymbol{Y}'}(\boldsymbol{t})\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \phi_{\boldsymbol{Z}'}(\boldsymbol{t}). \tag{13}$$

Since Z and Z' are symmetric about the origin, their characteristic functions $\Phi_Z(\boldsymbol{t})$ and $\Phi'_Z(\boldsymbol{t})$ are real continuous functions at $\boldsymbol{t} = \boldsymbol{0}$. It follows from Equation (13) that the imaginary part of $\phi_{\boldsymbol{Y}}(\boldsymbol{t})\phi_{-\boldsymbol{Y}'}(\boldsymbol{t})$ is equal to 0 in an open ball of $\boldsymbol{t} = \boldsymbol{0}$. Furthermore, it must be identically zero on R^d by the analytic property of the characteristic function.

Suppose that one d-variate random vector $\boldsymbol{Y}$ is symmetric about d-dimensional vector $\boldsymbol{\mu}$. Then, the random vector $\boldsymbol{Y}$ can be written as $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a d-dimensional random vector which is symmetric about the origin. For any unit vector $\boldsymbol{u} \in R^d$, we have $\boldsymbol{u}^\top\boldsymbol{Y} = \boldsymbol{u}^\top\boldsymbol{\mu} + \boldsymbol{u}^\top\boldsymbol{\epsilon}$. Then, $w = \boldsymbol{u}^\top\boldsymbol{\epsilon}$ is a univariate random variable and it is symmetric about zero. According to Theorem 2 in [10], for any linear independent vector $\boldsymbol{u}$, we have $\boldsymbol{\mu} = \boldsymbol{\mu}'$ if $\boldsymbol{u}^\top\boldsymbol{\mu} = \boldsymbol{u}^\top\boldsymbol{\mu}'$. Furthermore, we have $\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}'}(\boldsymbol{t})$. According to Equation (12), $\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \phi_{\boldsymbol{Z}'}(\boldsymbol{t})$ if $\phi_{\boldsymbol{Y}}(\boldsymbol{t}) \neq 0$. Because $\phi_{\boldsymbol{Y}}(t)$ is an analytic function, it is not identically zero. Thus, $\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \phi_{\boldsymbol{Z}'}(\boldsymbol{t})$ holds except for a discrete set. Therefore, Equation (12) implies both $\boldsymbol{Y} \overset{\mathrm{d}}{=} \boldsymbol{Y}'$ and $\boldsymbol{Z} \overset{\mathrm{d}}{=} \boldsymbol{Z}'$; that is $(\boldsymbol{\theta}, F) = (\boldsymbol{\theta}', F')$. □

APPENDIX B. CONSISTENCY PROOF

Proof of Theorem 3.3. First, according to Theorem 3.1 and Lemma 3.2, $\boldsymbol{\theta}_0$ is a unique minimizer of the criterion function $\mathcal{D}(\boldsymbol{\theta})$. Next, note that $\mathcal{D}(\boldsymbol{\theta})$ is continuous about $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$. Denote $\delta = \inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\geq\epsilon} \mathcal{D}(\boldsymbol{\theta}) - \mathcal{D}(\boldsymbol{\theta}_0)$, then we have $\delta > 0$. Hence,

$$\begin{aligned}
&P\Big\{\sup_{n\geq N} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > \epsilon\Big\} \\
&\leq P\Big[\sup_{n\geq N}\big\{\mathcal{D}(\hat{\boldsymbol{\theta}}_n) - \mathcal{D}(\boldsymbol{\theta}_0)\big\} \geq \delta\Big] \\
&\leq P\Big[\sup_{n\geq N}\big\{\mathcal{D}(\hat{\boldsymbol{\theta}}_n) - \mathcal{D}_n(\hat{\boldsymbol{\theta}}_n)\big\} + \sup_{n\geq N}\big\{\mathcal{D}_n(\boldsymbol{\theta}_0) - \mathcal{D}(\boldsymbol{\theta}_0)\big\} \geq \delta\Big] \\
&\leq 2P\Big\{\sup_{n\geq N}\sup_{\boldsymbol{\theta}} |\mathcal{D}_n(\boldsymbol{\theta}) - \mathcal{D}(\boldsymbol{\theta})| \geq \delta/2\Big\}.
\end{aligned}$$

If $\sup_{\boldsymbol{\theta}} |\mathcal{D}_n(\boldsymbol{\theta}) - \mathcal{D}(\boldsymbol{\theta})| \to 0$, a.s., then the theorem holds. In fact, denote $\mathcal{H} = \{h_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, which is a class function on $\mathbb{R}^{2d}$. Note that $0 < \lambda < 1/2$ for $h_{\boldsymbol{\theta}}$ in (7). Next using the Minkowski inequality, we have

$$\begin{aligned}
&h_{\boldsymbol{\theta}}(\boldsymbol{X}_1, \boldsymbol{X}_2) \\
&= \lambda^2(\|\boldsymbol{X}_1 + \boldsymbol{X}_2 - 2\boldsymbol{\mu}_1\| - \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|) \\
&\quad + 2\lambda(1-\lambda)\big(\|\boldsymbol{X}_1 + \boldsymbol{X}_2 - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\| \\
&\qquad\qquad - \|\boldsymbol{X}_1 - \boldsymbol{X}_2 + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\|\big) \\
&\quad + (1-\lambda)^2(\|\boldsymbol{X}_1 + \boldsymbol{X}_2 - 2\boldsymbol{\mu}_2\| - \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|) \\
&\leq \lambda^2(\|\boldsymbol{X}_1 + \boldsymbol{X}_2\| - \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| + 2\|\boldsymbol{\mu}_1\|) \\
&\quad + 2\lambda(1-\lambda)(\|\boldsymbol{X}_1 + \boldsymbol{X}_2\| - \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| + 2\|\boldsymbol{\mu}_1\| + 2\|\boldsymbol{\mu}_2\|) \\
&\quad + (1-\lambda)^2(\|\boldsymbol{X}_1 + \boldsymbol{X}_2\| - \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| + 2\|\boldsymbol{\mu}_2\|) \\
&\leq \|\boldsymbol{X}_1 + \boldsymbol{X}_2\| - \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| + 2(\|\boldsymbol{\mu}_1\| + \|\boldsymbol{\mu}_2\|).
\end{aligned}$$

If $E\{\max(\|\boldsymbol{X}_1\|, \|\boldsymbol{X}_2\|)\} < \infty$ and $\max(\|\boldsymbol{\mu}_1\|, \|\boldsymbol{\mu}_2\|) < \infty$, according to Lemma 18(ii) of [16], $\mathcal{H}$ is a VC class. Then, we have $\sup_{\boldsymbol{\theta}\in\Theta} |\mathcal{D}_n(\boldsymbol{\theta}) - \mathcal{D}(\boldsymbol{\theta})| \to 0$, a.s. according to Theorem 7 of [16]. □

APPENDIX C. ASYMPTOTIC NORMALITY PROOF

Proof of Theorem 3.4. Denote $\hat{\Delta}_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ and $\tilde{\Delta}_n = 2\sqrt{n}V_n^{(1)}\eta$. First, we show that both of the sequences $\hat{\Delta}_n$ and $\tilde{\Delta}_n$ are stochastically bounded. Define $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + A^{-1}\tilde{\Delta}_n/\sqrt{n}$ and using these results we can obtain the equations

$$\begin{aligned}
nV_n^{(2)}(h_{\hat{\boldsymbol{\theta}}_n} - h_{\boldsymbol{\theta}_0}) &= -\frac{1}{2}\hat{\Delta}_n^\top A\hat{\Delta}_n + \hat{\Delta}_n\tilde{\Delta}_n + o_p(1), \\
nV_n^{(2)}(h_{\tilde{\boldsymbol{\theta}}_n} - h_{\boldsymbol{\theta}_0}) &= \frac{1}{2}\tilde{\Delta}_n^\top A^{-1}\tilde{\Delta}_n + o_p(1).
\end{aligned}$$

By the definition of $\hat{\boldsymbol{\theta}}_n$, the left side of the second equation is not less than that of the first equation. Then take the difference, complete the square and we obtain

$$\frac{1}{2}\big(\hat{\Delta}_n - A^{-1}\tilde{\Delta}_n\big)^\top A\big(\hat{\Delta}_n - A^{-1}\tilde{\Delta}_n\big) + o_p(1) \geq 0.$$

Furthermore, $\|\hat{\Delta}_n - A^{-1}\tilde{\Delta}_n\| \to 0$ in probability. By the central limit theorem, $\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges to $N(\boldsymbol{0}, 4A^{-1}BA^{-1})$ in distribution, and the result follows. More technical details about this proof can refer to Theorem 2.1 in [2]. □

ACKNOWLEDGEMENT

Xueqin Wang's research is partially supported by NSFC for Excellent Young Scholar(11322108), NCET(12-0559), NSFC(11001280), RFDP(20110171110037). Heping Zhang's research is partially supported by grant R01DA016750-08 from the U.S. National Institute on Drug Abuse, a 1,000-plan scholarship from the Chinese Government, and the overseas and Hong Kong, Macau Young Scholars Collaborative Research Fund from the National Natural Science Foundation of China (11328103).

Received 20 September 2013

REFERENCES

[1] Allman, E., Matias, C. and Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

[2] Arcones, M., Chen, Z. and Gine, E. (1994). Estimators related to u-processes with applications to multivariate medians: Asymptotic normality. *Ann. Statist.* **22** 1460–1477. MR1311984

[3] Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232. MR2278356

[4] Cruz-Medina, I., Hettmansperger, T. and Thomas, H. (2004). Semiparametric mixture models and repeated measures: The multinomial cut point model. *J. R. Statist. Soc. C* **53** 463–474. MR2075502

[5] Fang, K., Kotz, S. and Ng, K. (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London. MR1071174

[6] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York. MR2265601

[7] Hall, P., Neeman, A., Pakyari, R. and Elmore, R. (2005). Nonparametric inference in multivariate mixtures. *Biometrika* **92** 667–678. MR2202653

[8] Hall, P. and Zhou, X. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31** 201–224. MR1962504

[9] Hettmansperger, T. and Thomas, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *J. R. Statist. Soc. B* **62** 811–825. MR1796294

[10] Hunter, D., Wang, S. and Hettmansperger, T. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275

[11] Komárek, A. (2009). A new R package for bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Comput. Stat. Data. An.* **53** 3932–3947. MR2744295

[12] Kruskal, J. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **18** 95–138. MR0444690

[13] Levine, M., Hunter, D. and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98** 403–416. MR2806437

[14] Lindsay, B. (1995). *Mixture Models: Theory, Geometry, and Applications*. Institue for Mathematical Statistics, Hayward.

[15] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

[16] Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *Ann. Statist.* **15** 780–799. MR0888439

[17] Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930

[18] Schlattmann, P. (2009). *Medical Applications of Finite Mixture models*. Springer, Berlin.

[19] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

[20] Serfling, R. (2006). Multivariate symmetry and asymmetry. *Encyclopedia of Statistical Sciences*.

[21] Székely, G. and Rizzo, M. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *J. Classification* **22** 151–183. MR2231170

[22] Székely, G. and Rizzo, M. (2013). Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference* **143** 1249–1272. MR3055745

[23] Titterington, D., Smith, A. and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York. MR0838090

[24] Weyl, H. (1983). *Symmety*. Princeton University Press, New Jersey.

Wenxiu Ge
Department of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, 510275
China
E-mail address: wenxiuge@gmail.com

Xiaobo Guo
Department of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, 510275
China
E-mail address: mc03gxb@126.com

Xueqin Wang
Department of Statistical Science
School of Mathematics and Computational Science
State Key Laboratory of Ophthalmology
Zhongshan Ophthalmic Center
Zhongshan Medical School
Xinhua College
Sun Yat-Sen University
Guangzhou, 510275
China
E-mail address: wangxq88@mail.sysu.edu.cn

Heping Zhang
Department of Statistical Science
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou, 510275
China

Department of Biostatistics
Yale University School of Public Health
New Haven, CT 06520-8034
USA
E-mail address: heping.zhang@yale.edu