Computational Statistics and Data Analysis 103 (2016) 413–425

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric mixture: Continuous scale mixture approach

Sijia Xiang [a], Weixin Yao [b], Byungtae Seo [c,*]

[a] School of Mathematics and Statistics, Zhejiang University of Finance & Economics, Hangzhou, Zhejiang 310018, PR China

[b] Department of Statistics, University of California, Riverside, CA 92887, USA

[c] Department of Statistics, Sungkyunkwan University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form 2 June 2016
Accepted 2 June 2016
Available online 11 June 2016

Keywords:
Mixture models
Semiparametric EM algorithm
Semiparametric mixture models
Continuous normal scale mixture

ABSTRACT

In this article, we propose a new estimation procedure for a class of semiparametric mixture models that is a mixture of unknown location-shifted symmetric distributions. The proposed method assumes that the nonparametric symmetric distribution falls in a rich class of continuous normal scale mixture distributions. With this new modeling approach, we can suitably avoid the misspecification problem in traditional parametric mixture models. In addition, unlike some existing semiparametric methods, the proposed method does not require any modification or smoothing of the likelihood as it can directly estimate parametric and nonparametric components simultaneously in the model. Furthermore, the proposed parameter estimates are robust against outliers. The estimation algorithms are introduced and numerical studies are conducted to examine the finite sample performance of the proposed procedure and to compare it with other existing methods.

1. Introduction

Finite mixture models have a wide application including cluster and latent class analysis, discriminant analysis, image analysis, and survival analysis. They provide extremely flexible descriptive models for distributions in data analysis and inference. For general introduction of mixture models, see Lindsay (1995), Böhning (1999), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006).

A general form of a finite mixture density can be expressed as

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

where $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_m, \pi_1, \ldots, \pi_m)$, $\sum_{j=1}^{m} \pi_j = 1$, $\pi_j > 0$ for $j = 1, \ldots, m$, and $f(x;\lambda_j)$ is the density function for the jth component. The traditional parametric mixture model assumes that the density f belongs to a certain parametric family, such as family of normal distributions or t-distributions. The maximum likelihood estimator (MLE) of the unknown parameter $\boldsymbol{\theta}$ can be then obtained using the Expectation–Maximization (EM) algorithm.

In practice, however, a practitioner might not have prior information about which parametric family one should use for f. The estimate of $\boldsymbol{\theta}$ might be sensitive to the parametric form of f, and in addition, a distributional misspecification of f could lead to wrong or inefficient statistical inference. For this, one can consider a semiparametric model that leaves f completely unspecified. Yet, this causes an identifiability problem as the model is too flexible and not parsimonious enough. For this

* Corresponding author.
E-mail address: seobt@skku.edu (B. Seo).

http://dx.doi.org/10.1016/j.csda.2016.06.001
0167-9473/

identifiability issue in semiparametric mixture models, Bordes et al. (2006) and Hunter et al. (2007) considered the following location-shifted semiparametric model with symmetric nonparametric component densities:

$$p(x;\boldsymbol{\theta},f)=\sum_{j=1}^{m}\pi_j f(x-\mu_j), \tag{1.1}$$

where $\boldsymbol{\theta}=(\pi_1,\mu_1,\ldots,\pi_m,\mu_m)$ and f is an unknown but symmetric density about zero. Bordes et al. (2006) proved the identifiability of model (1.1) for $m=2$. Hunter et al. (2007) further established the identifiability of model (1.1) for both $m=2$ and $m=3$.

Bordes et al. (2007) and Benaglia et al. (2009) proposed a semiparametric EM type algorithm to estimate parameters in (1.1) using a kernel-based estimator for the symmetric nonparametric density f. They demonstrated, through numerical study, its superiority over the methods provided by Hunter et al. (2007) and Bordes et al. (2006). However, the bandwidth selection is not an easy task and sensitive to the model efficiency. In this case, many commonly used methods for bandwidth selection may not be relevant because each component density may have a different impact on the choice of bandwidth and the ideal bandwidth selection depends on whether components are well-separated or not.

In this article, we propose a new method to estimate the model parameters in (1.1) by modeling f as nonparametric scale mixtures. The proposed method is free from bandwidth selection and thus is more reliable and robust to model misspecification. Unlike Bordes et al. (2007) or Benaglia et al. (2009), the new technique relies only on the likelihood function without any modification or smoothing. In addition, it can give a direct legitimate nonparametric estimator of f. Furthermore, the proposed parameter estimates are robust against outliers.

The remainder of this paper is organized as follows. In Section 2, we introduce the new estimation method for the semiparametric mixture model (1.1) and an effective algorithm is introduced to find the proposed estimator in Section 3. In Section 4, we present both a Monte Carlo study and a real data example to compare the proposed new estimator with some other existing methods. Finally, some discussions are given in Section 5.

2. Semiparametric mixtures under continuous scale mixture

For the nonparametric symmetric density f in (1.1), we propose to model f as a continuous normal scale mixture. That is, we assume that f is a member of

$$\mathcal{F}=\left\{f(x)|\int\frac{1}{\sigma}\phi(x/\sigma)dQ(\sigma)\right\}, \tag{2.1}$$

where $\phi(x)$ is the standard normal density, and Q is an unspecified probability measure on $\mathbb{R}^+$. Although we restrict the nonparametric symmetric density f to $\mathcal{F}$, $\mathcal{F}$ is rich enough to contain almost all symmetric unimodal continuous probability densities such as normal, Laplace, t, stable, and so on. Efron and Olshen (1978) and Basu (1996) discussed on how many distributions are contained in $\mathcal{F}$. Kelker (1971) and Andrews and Mallows (1974) also studied necessary and sufficient conditions for a probability density to be a member of $\mathcal{F}$. Recently, Seo and Lee (2015) utilized this class of normal scale mixture densities to efficiently estimate the distribution of innovations as well as parameters in semiparametric generalized autoregressive conditional heteroskedasticity models. Böhning and Ruangroj (2002) discussed the difference between the normal with a free variance parameter and component mixture of normals with the same mean for $m=2$. Böhning and Ruangroj (2002) proved in Theorem 2.3, the difference is an increasing function of the contaminated component variance when other parameters are fixed. However their results are limited to two-component normal scale mixtures and thus cannot be applied to the continuous normal scale mixtures. Since the t-distribution is a special case of the continuous normal scale mixtures, the difference can be very big between normal distribution with a free variance parameter and the continuous normal scale mixtures.

Under this model class, (1.1) can be expressed as

$$\begin{aligned}p(x;\boldsymbol{\theta},Q)&=\sum_{j=1}^{m}\pi_j\left\{\int\frac{1}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right)dQ(\sigma)\right\}\\&=\int\sum_{j=1}^{m}\frac{\pi_j}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right)dQ(\sigma).\end{aligned} \tag{2.2}$$

The identifiability of (2.2) can be shown by combining the identifiability of (1.1) and $\mathcal{F}$ as described in Proposition 2.1. Chee and Wang (2013) used a similar argument for the identifiability of their semiparametric location mixtures.

Proposition 2.1. *The semiparametric model $p(x;\boldsymbol{\theta},Q)$ in (2.2) is identifiable when $m\le 3$, i.e., if $p(x;\boldsymbol{\theta},Q)=p(x;\boldsymbol{\theta}^*,Q^*)$, then $Q=Q^*$ and $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ up to a permutation of component labels.*

Proof. Suppose $p(x;\boldsymbol{\theta}, Q) = p(x;\boldsymbol{\theta}^*, Q^*)$, i.e.,

$$\sum_{j=1}^{m}\pi_j\left\{\int\frac{1}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right)dQ(\sigma)\right\} = \sum_{j=1}^{m}\pi_j^*\left\{\int\frac{1}{\sigma}\phi\left(\frac{x-\mu_j^*}{\sigma}\right)dQ^*(\sigma)\right\}.$$

Since both $\int\frac{1}{\sigma}\phi(x/\sigma)dQ(\sigma)$ and $\int\frac{1}{\sigma}\phi(x/\sigma)dQ^*(\sigma)$ are symmetric distributions, from the identifiability result of (1.1) provided in Hunter et al. (2007), we can get $\pi_j^* = \pi_j$, $\mu_j^* = \mu_j$ up to a permutation of component labels. In addition, we have $\int\frac{1}{\sigma}\phi(x/\sigma)dQ(\sigma) = \int\frac{1}{\sigma}\phi(x/\sigma)dQ^*(\sigma)$. Based on the identifiability of the normal scale mixture density (Lindsay, 1983b), we can get $Q = Q^*$ except on a set of probability zero. Therefore, $p(x;\boldsymbol{\theta}, Q) = p(x;\boldsymbol{\theta}^*, Q^*)$ implies $Q = Q^*$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ up to a permutation of component labels. □

In this article, we mainly focus on estimating model (2.2) assuming m is known, and demonstrate the performance of the proposed method in the simulation for $m = 2$ and $m = 3$. Computationally, our method can be easily extended to the case of $m > 3$, however, it requires more research to establish the identifiability of model (2.2) for $m > 3$.

In the first displayed expression of (2.2), if Q is known, $\int\frac{1}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right)dQ(\sigma)$ can be considered as a parametric component density in a finite location mixture model. On the other hand, in the last displayed expression of (2.2), if (π_j, μ_j)'s are known, $\sum_{j=1}^{m}\frac{\pi_j}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right)$ plays a role of a component density in the nonparametric mixture with unknown mixing distribution Q. Though we model the nonparametric density f using (2.1), (2.2) can still be considered as a class of semiparametric mixture model as (2.1) is a rich class of symmetric distributions with the nonparametric component Q.

When the mixing distribution Q is a distribution with finite support points, we can also interpret (2.2) as a finite location-scale mixture but with some restriction. To explain this, suppose that Q is a discrete probability measure supported by σ_1 and σ_2 with corresponding point masses p_1 and p_2, respectively. $p(x;\boldsymbol{\theta}, Q)$ in (2.2) with $m = 2$ is then

$$\frac{\pi_1 p_1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + \frac{\pi_1 p_2}{\sigma_2}\phi\left(\frac{x-\mu_1}{\sigma_2}\right) + \frac{\pi_2 p_1}{\sigma_1}\phi\left(\frac{x-\mu_2}{\sigma_1}\right) + \frac{\pi_2 p_2}{\sigma_2}\phi\left(\frac{x-\mu_2}{\sigma_2}\right). \tag{2.3}$$

This is just a four-component normal location-scale mixture density but with some restrictions, for example, the location parameters of the first two components are identical and the scale parameters of the first and third components are the same.

One potential problem with finite location-scale mixtures is that the likelihood may not be bounded (Kiefer and Wolfowitz, 1956). In most finite location-scale mixtures, if the scale parameters are common, the likelihood is bounded. However, since (2.3) allows heterogeneous scale parameters, one can easily show that the likelihood is unbounded: let μ_1 be one of data points and $\sigma_1 \downarrow 0$. This in turn implies that the likelihood based on (2.2) is also unbounded because (2.3) is a submodel of (2.2). There have been considerable research efforts in dealing with the unbounded likelihood issue for finite normal mixtures. See, for example, Hathaway (1985), Chen et al. (2008), Yao (2010) and Seo and Kim (2012). To avoid this unbounded likelihood issue, we gave a constraint on the support of Q. That is, instead of assuming that Q is a probability measure on $(0, \infty)$, we restricted the support of Q to $[c, \infty)$, where c is a predetermined positive constant. This simple idea was first proposed by Hathaway (1985) in order to avoid singular problems in finite location-scale mixtures. Later, Tanaka and Takemura (2006) proposed to use $[c_n, \infty)$, where $c_n \downarrow 0$, and showed the strong consistency for general finite mixtures of location-scale distributions with $\log(c_n) = O(-n^d)$, $0 < d < 1$. These simple restrictions also work for our case and almost no significant effect for the estimation of Q and $\boldsymbol{\theta}$.

Although $\mathcal{F}$ is a subclass of the class of all symmetric densities, $\mathcal{F}$ is a quite dense class as it contains almost all unimodal symmetric probability densities. Based on our limited empirical experience (see, also, the numerical examples in Section 4.1), the new estimation procedure still works well even if the true component density does not belong to $\mathcal{F}$. The proposed new method has several important advantages over existing methods. First, unlike other existing methods, the proposed model has an explicit form of the likelihood in which neither artificial modification nor a tuning parameter is required. This can yield highly efficient estimators. Second, the estimation of Q directly gives a legitimate density estimator for each component density as

$$\int\frac{1}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right)dQ, \quad j = 1, \ldots, m.$$

Third, since $\mathcal{F}$ contains many heavy tailed distributions, the estimation of $\boldsymbol{\theta}$ would be very robust against outliers. In fact, if an outlier exists, the NPMLE of Q tends to contain a large support point with a small mass. This can downweight the effect of an outlier so that the estimation of location parameters is not greatly affected by outliers. Recently, Seo et al. (submitted for publication) used $\mathcal{F}$ to model the error distribution in a regression setting and showed that the estimated regression coefficients are quite robust to some severe outliers. We will show this outlier-resistant property empirically in Section 4.

3. Estimating algorithms

For a given random sample $X_1, \ldots, X_n$, the log-likelihood based on (2.2) is given by

$$\ell(\boldsymbol{\theta}, Q) = \sum_{i=1}^{n} \log \left\{ \int \sum_{j=1}^{m} \frac{\pi_j}{\sigma} \phi\left(\frac{x_i - \mu_j}{\sigma}\right) dQ(\sigma) \right\}. \tag{3.1}$$

The simultaneous estimation for the MLE of $\boldsymbol{\theta}$ and Q is quite difficult problem owing to the nonparametric mixing distribution Q. For this problem, we propose to iteratively update $\boldsymbol{\theta}$ and Q in turn until convergence. Since (3.1) can be considered as a mixture of finite and infinite mixtures, as demonstrated in Section 2, we can iteratively exploit the standard EM algorithm for $\boldsymbol{\theta}$ and some existing algorithms for Q. In the following subsections, we explain the procedure to find the NPMLE of Q with fixed $\boldsymbol{\theta}$ and the MLE of $\boldsymbol{\theta}$ with fixed Q.

3.1. NPMLE of Q

In this subsection, we provide the estimation method for the NPMLE of Q with fixed $\boldsymbol{\theta}$. Even with fixed $\boldsymbol{\theta}$, the estimation of Q is not a simple task as it is an optimization problem over an infinite dimensional space. Lindsay (1983a) showed the existence and uniqueness of the NPMLE of Q and also showed that the NPMLE of Q must be discrete with finite support points no more than the number of observations. This enables us to search for the NPMLE of Q on distributions having finite support points. To estimate the mixing distribution Q, one may fix support points of Q on a predetermined grid and estimate the weights for each grid point by exploiting the EM algorithm (Laird, 1978). However, this method requires huge computing time particularly for semiparametric mixture models and the choice of grid is another critical issue.

In this case, some gradient based algorithms would be efficient alternatives. To explain gradient-based algorithms, let us rewrite (3.1) as

$$\ell_n(Q) = \sum_i \log \left(\int g(x_i; \boldsymbol{\theta}, \sigma) dQ(\sigma) \right),$$

where

$$g(x; \boldsymbol{\theta}, \sigma) = \sum_{j=1}^{m} \frac{\pi_j}{\sigma} \phi\left(\frac{x - \mu_j}{\sigma}\right).$$

The gradient function $D_Q(\xi)$ is defined as the directional derivative of $\ell_n(Q)$ at Q toward δ_ξ, where δ_ξ is a Dirac δ distribution having all its mass at ξ:

$$\begin{aligned} D_Q(\xi) &= \frac{d}{d\alpha} \ell_n((1-\alpha)Q + \alpha\delta_\xi)|_{\alpha=0} \\ &= \sum_{i=1}^{n} \frac{g(x_i; \boldsymbol{\theta}, \xi)}{\int g(x_i; \boldsymbol{\theta}, \sigma) dQ(\sigma)} - n. \end{aligned} \tag{3.2}$$

For a given estimator $\hat{Q}_n$, $D_{\hat{Q}_n}(\xi)$ is an important tool to judge if the current estimator $\hat{Q}_n$ attains the NPMLE. That is, if $D_{\hat{Q}_n}(\xi)$ is greater than zero at $\xi = \xi^*$, this means that there exists $\hat{Q}_n^* = (1-\alpha)\hat{Q}_n + \alpha\delta_{\xi^*}$ satisfying $\ell_n(\hat{Q}_n) < \ell_n(\hat{Q}_n^*)$ for some $0 < \alpha < 1$. Hence, $D_{\hat{Q}_n}(\xi) \leq 0$ for all ξ is a necessary condition for $\hat{Q}_n$ to be the NPMLE. Indeed, it is also a sufficient condition (Lindsay, 1995). We call it a gradient condition.

Vertex-Direction-Method (VDM) utilizes this property to find the NPMLE of Q. VDM first searches the maximizer ξ^* of $D_{\hat{Q}_n}(\xi)$ with the current estimator $\hat{Q}_n$. If $D_{\hat{Q}_n}(\xi^*) \leq 0$, the algorithm stops and returns $\hat{Q}_n$ as the NPMLE of Q. Otherwise VDM updates $\hat{Q}_n$ to $(1-\alpha)\hat{Q}_n + \alpha\delta_{\xi^*}$ for some $0 < \alpha < 1$. This type of algorithms was first proposed in the literature of optimal design theory (Wynn, 1970, 1972; Atwood, 1976; Wu, 1978). Böhning (1982) and Lindsay (1983a) showed their connection to the NPMLE problem in continuous mixtures.

Although VDM simplifies the estimation procedure and guarantees the convergence to the NPMLE of Q, it is generally too slow and requires too many iterations until the gradient condition is satisfied. This slow convergence is mainly caused by too many support points because the number of support points in VDM always increases in each iteration. Several algorithms have been suggested to speed up the convergence such as Vertex-Exchange-Method (VEM) (Böhning, 1986), Intra-Simplex-Direction-Method (ISDM) (Lesperance and Kalbfleisch, 1992), and Constrained-Newton method for Multiple supports (CNM) (Wang, 2007). Among these, we the recommend CNM algorithm because it is much faster than other existing algorithms.

We here briefly describe the CNM method to find the NPMLE of Q for fixed $\boldsymbol{\theta}$. Suppose that $\hat{Q}_n^{(t)}$ is the estimator of Q at (t)th iteration. CNM first adds all local maximizers of $D_{\hat{Q}_n^{(t)}}(\xi)$ to the current set of support points of $\hat{Q}_n^{(t)}$. Let this new set of

support points be $\boldsymbol{\xi}^{(t+\frac{1}{2})} = (\xi_1, \ldots, \xi_K)^T$. The corresponding weight vector $\boldsymbol{p}^{(t+\frac{1}{2})}$ is then set to be the minimizer of $\|\boldsymbol{S}\boldsymbol{p} - \boldsymbol{2}\|$ subject to $\boldsymbol{p} \geq 0$ and $\boldsymbol{p}^T\boldsymbol{1} = 1$, where $\boldsymbol{1} = (1, \ldots, 1)^T$, $\boldsymbol{2} = (2, \ldots, 2)^T$, and $\boldsymbol{S}$ is a $n \times K$ matrix having (i, k)th element

$$\frac{g(x_i; \boldsymbol{\theta}, \xi_k)}{\sum_{l=1}^{K} p_l g(x_i; \boldsymbol{\theta}, \xi_l)}.$$

Then, all zero elements in $\boldsymbol{p}^{(t+\frac{1}{2})}$ and the corresponding support points in $\boldsymbol{\xi}^{(t+\frac{1}{2})}$ will be removed, and the remaining support points and weights are set to $\boldsymbol{p}^{(t+1)}$ and $\boldsymbol{\xi}^{(t+1)}$ which determine $\hat{Q}_n^{(t+1)}$. These procedure will be repeated until the gradient condition is fulfilled. For more details, see Wang (2007).

3.2. MLE of $\boldsymbol{\theta}$

When Q is fixed in (2.2), estimating $\boldsymbol{\theta}$ is just a simple application of the EM algorithm for the usual finite mixture models because it is equivalent to estimating parameters in the m-component mixture with component densities $\int \frac{1}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right) dQ$, $j = 1, \ldots, m$. In this case, E-step is reduced to computing the posterior probability that x_i belongs to the jth component with a given current parameter estimate $\boldsymbol{\theta}^{(t)}$. That is, for fixed Q and the current parameter estimate $\boldsymbol{\theta}^{(t)}$, the E-step is equivalent to computing

$$z_{ij}^{(t)} = \frac{\pi_j^{(t)} \int \phi((x_i - \mu_j^{(t)})/\sigma)/\sigma \, dQ(\sigma)}{\sum_{k=1}^{m} \pi_k^{(t)} \int \phi((x_i - \mu_k^{(t)})/\sigma)/\sigma \, dQ(\sigma)}. \tag{3.3}$$

Because the current estimator of Q must be discrete, without loss of generality, we can assume that the fixed Q has support points $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^T$ with corresponding weights $\boldsymbol{p} = (p_1, \ldots, p_K)^T$. Then, $z_{ij}^{(t)}$ can also be expressed as

$$z_{ij}^{(t)} = \frac{\pi_j^{(t)} \sum_{k=1}^{K} p_k \phi((x_i - \mu_j^{(t)})/\xi_k)/\xi_k}{\sum_{l=1}^{m} \pi_l^{(t)} \sum_{k=1}^{K} p_k \phi((x_i - \mu_l^{(t)})/\xi_k)/\xi_k}.$$

In M-step, $\boldsymbol{\theta}^{(t+1)}$ is obtained by maximizing

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}^{(t)} \log\left(\pi_j \sum_{k=1}^{K} p_k \phi((x_i - \mu_j)/\xi_k)/\xi_k\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}^{(t)} \log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}^{(t)} \log\left(\sum_{k=1}^{K} p_k \phi((x_i - \mu_j)/\xi_k)/\xi_k\right) \end{aligned}$$

with respect to π_j and $\mu_j, j = 1, \ldots, m$. From $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, $\pi_j^{(t+1)}$ can be explicitly computed as

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n}. \tag{3.4}$$

However, there is no explicit form for the maximizer of

$$\sum_{i=1}^{n} z_{ij}^{(t)} \log\left(\sum_{k=1}^{K} p_k \phi((x_i - \mu_j)/\xi_k)/\xi_k\right), \tag{3.5}$$

with respect to μ_j. In such a case, standard optimization methods such as the Newton–Raphson algorithm can be used to find $\mu_j^{(t+1)}$. However, this optimization strategy may not be stable due to the nature of the mixture likelihood.

As an alternative, we can use an EM-like algorithm for this optimization problem. The maximization of (3.5) with respect to μ_j is similar to computing the MLE in finite normal scale mixture models, where the common unknown mean parameter is μ_j with fixed $(p_1, \ldots, p_K)$ and $(\xi_1, \ldots, \xi_K)$. This interpretation facilitates the use of the EM algorithm again. In this case, the E-step is to compute

$$u_{ijk} = \frac{p_k \phi((x_i - \mu_j)/\xi_k)/\xi_k}{\sum_{l=1}^{K} p_l \phi((x_i - \mu_j)/\xi_l)/\xi_l}. \tag{3.6}$$

In the M-step, μ_j is updated by

$$\frac{\sum_{i=1}^{n} z_{ij}^{(t)} x_i \sum_{k=1}^{K} \frac{u_{ijk}}{\xi_k^2}}{\sum_{i=1}^{n} z_{ij}^{(t)} \sum_{k=1}^{K} \frac{u_{ijk}}{\xi_k^2}}. \tag{3.7}$$

Now, $\mu_j^{(t+1)}$ is obtained by iterating (3.6) and (3.7) until it converges. For these inner E- and M-steps, five iterations are usually sufficient based on our experience.

Now, we summarize the full iterative algorithm to find the MLE of $\boldsymbol{\theta}$ and Q in Algorithm 3.1.

Algorithm 3.1. 1. For fixed $\boldsymbol{\theta}^{(t)}$, find all local maximizers of $D_{Q^{(t)}}(\xi)$ and add these to the set of support points of $Q^{(t)}$.
2. Using constrained Newton method, update weights corresponding to the new support points obtained in step 1.
3. Discard all support points corresponding to zero weights, and set the remaining support points and weights to $Q^{(t+1)}$.
4. For fixed $Q^{(t+1)}$, compute $\pi_j^{(t+1)}$ and $\mu_j^{(t+1)}$ using the EM algorithm, $j = 1, \ldots, m$.

In Step 4, like the inner E- and M-steps in (3.6) and (3.7), a small number of iterations for the outer E- and M-steps can be used for computational ease.

The likelihood based on the proposed method generally has multiple local maxima and we need to start the algorithm from several initial values. One way to obtain the initial values for $\boldsymbol{\theta}$ and Q is to use the MLE of $\boldsymbol{\theta}$ and σ from $\sum_{j=1}^{m} \frac{\pi_j}{\sigma} \phi\left(\frac{x-\mu_j}{\sigma}\right)$, say $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}$, and set $\boldsymbol{\theta}^{(0)} = \hat{\boldsymbol{\theta}}$ and $Q^{(0)} = \delta_{\hat{\sigma}}$ as the initial estimates for $\boldsymbol{\theta}$ and Q. We can also use some random partitions of data to m clusters to obtain the initial values for $\boldsymbol{\theta}$ and Q. In practice, it is prudent to run the algorithm from multiple initial values and choose the solution which maximizes the log-likelihood (3.1).

4. Numerical examples

4.1. Simulation studies

In this section, we use Monte Carlo simulation studies to illustrate the finite sample performance of the proposed semiparametric mixtures under continuous scale mixtures (SMCSM), and compare it with the MLE based on normality assumption for the component density (MLE) and semiparametric EM algorithm (SPEM) proposed by Benaglia et al. (2009).

For SMCSM and MLE, we use 10 random values and the true value as initial values, and select the converged value which has the largest likelihood. For SPEM, we use MLE as the initial value (note that for SPEM, there is no objective function and thus it is difficult to choose the right root if multiple initial values are used).

The first simulations are based on two-component mixture models. We assume $m = 2$ to be known. A random sample $\{x_1, \ldots, x_n\}$ is generated from a population with density function

$$h(x) = \pi_1 f(x - \mu_1) + (1 - \pi_1) f(x - \mu_2), \tag{4.1}$$

where (π_1, μ_1, μ_2) are unknown parameters and f is an unknown density that is symmetric about zero. We consider the following five cases:

*Case*I: $f(x) \sim N(0, 1)$, $\mu_1 = 0$, $\mu_2 = 3$, $\pi_1 = 0.3$.
*Case*II: $f(x) \sim U(-1, 1)$, $\mu_1 = 0$, $\mu_2 = 1$, $\pi_1 = 0.3$.
*Case*III: $f(x) \sim t_3$, $\mu_1 = 0$, $\mu_2 = 3$, $\pi_1 = 0.3$.
*Case*IV: $f(x) \sim t_5$, $\mu_1 = 0$, $\mu_2 = 3$, $\pi_1 = 0.3$.
*Case*V: $f(x) \sim$ Laplace$(0, 1)$, $\mu_1 = -1$, $\mu_2 = 1$, $\pi_1 = 0.3$.

We use Case I to test the efficiency of our new estimator when the component density is correctly specified by MLE. Case II is included to see the performance of methods when the true component density does not belong to $\mathcal{F}$ in (2.1), since uniform distribution is not a member of $\mathcal{F}$. Cases III and IV demonstrate situations with heavy-tailed component densities. Case III is also the model used by Bordes et al. (2006) and Benaglia et al. (2009) to show the performance of their semiparametric EM algorithm. Case V is used to demonstrate that the new method can be adaptive to non-normal component densities. Note that Cases III, IV and V are scale-mixtures of normals.

For simulation studies, a total of 200 random samples with sample sizes $n = 100$, 300, and 500 were generated from each case. To assess the performance, we compute both the mean and the mean squared error (MSE) of each estimate:

$$\text{mean}(\hat{\boldsymbol{\theta}}) = \bar{\hat{\boldsymbol{\theta}}} =: \frac{1}{N} \sum_{t=1}^{N} \hat{\boldsymbol{\theta}}_t,$$

$$\text{MSE}(\hat{\boldsymbol{\theta}}) =: \frac{1}{N} \sum_{t=1}^{N} (\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta})^2,$$

Table 1
Average (MSE) of point estimates over 200 repetitions with $n = 100$.

Case	TRUE	SMCSM	MLE	SPEM
I	$\pi_1 = 0.3$	0.303(0.004)	0.3038(0.004)	0.286(0.006)
	$\mu_1 = 0$	0.033(0.108)	0.024(0.068)	0.161(0.354)
	$\mu_2 = 3$	3.001(0.031)	3.0161(0.023)	2.875(0.087)
II	$\pi_1 = 0.3$	0.451(0.049)	0.398(0.031)	0.356(0.021)
	$\mu_1 = 0$	0.125(0.071)	0.037(0.037)	0.414(0.289)
	$\mu_2 = 1$	1.179(0.088)	1.130(0.047)	0.831(0.074)
III	$\pi_1 = 0.3$	0.313(0.007)	0.380(0.142)	0.312(0.147)
	$\mu_1 = 0$	0.030(0.306)	−1.159(12.241)	0.091(12.864)
	$\mu_2 = 3$	3.014(0.047)	4.944(26.658)	4.253(26.687)
IV	$\pi_1 = 0.3$	0.314(0.006)	0.297(0.034)	0.263(0.040)
	$\mu_1 = 0$	0.077(0.221)	−0.493(3.274)	0.107(3.852)
	$\mu_2 = 3$	3.004(0.043)	3.249(3.178)	2.819(3.272)
V	$\pi_1 = 0.3$	0.303(0.008)	0.336(0.084)	0.264(0.083)
	$\mu_1 = -1$	−0.993(0.180)	−1.392(2.221)	−0.725(2.334)
	$\mu_2 = 1$	0.999(0.025)	1.405(2.042)	0.916(2.047)

Table 2
Average (MSE) of point estimates over 200 repetitions with $n = 300$.

Case	TRUE	SMCSM	MLE	SPEM
I	$\pi_1 = 0.3$	0.296(0.001)	0.297(0.001)	0.284(0.002)
	$\mu_1 = 0$	0.001(0.030)	−0.011(0.024)	−0.006(0.026)
	$\mu_2 = 3$	2.986(0.009)	3.000(0.008)	2.943(0.014)
II	$\pi_1 = 0.3$	0.431(0.029)	0.419(0.023)	0.320(0.007)
	$\mu_1 = 0$	0.106(0.028)	0.080(0.018)	0.204(0.098)
	$\mu_2 = 1$	1.156(0.046)	1.145(0.032)	0.910(0.020)
III	$\pi_1 = 0.3$	0.305(0.002)	0.322(0.162)	0.330(0.163)
	$\mu_1 = 0$	0.050(0.053)	−3.729(51.876)	−3.451(59.433)
	$\mu_2 = 3$	3.016(0.015)	5.658(46.821)	5.693(50.771)
IV	$\pi_1 = 0.3$	0.303(0.002)	0.291(0.019)	0.259(0.020)
	$\mu_1 = 0$	0.014(0.045)	−0.433(3.609)	−0.202(3.541)
	$\mu_2 = 3$	3.018(0.011)	3.138(1.626)	2.956(1.615)
V	$\pi_1 = 0.3$	0.306(0.002)	0.293(0.048)	0.256(0.052)
	$\mu_1 = -1$	−1.011(0.012)	−1.330(1.150)	−0.966(1.042)
	$\mu_2 = 1$	1.005(0.005)	1.154(1.113)	1.014(1.252)

where all calculations are elementwise, $N = 200$ is the number of repetitions, and $\hat{\boldsymbol{\theta}}_t$ is the estimate based on the tth replicate. $\hat{\boldsymbol{\theta}}$ is either MLE, SPEM, or SMCSM of $\boldsymbol{\theta} = (\pi_1, \mu_1, \mu_2)$.

Note, however, for mixture models, there are well known label switching issues (Celeux et al., 2000; Stephens, 2000; Yao and Lindsay, 2009; Yao, 2015) when doing comparison using the simulation study. There are no widely accepted labeling methods. In our simulation study, we choose the labels by minimizing the distance to the true parameter values.

Tables 1–3 present the mean and MSE of each parameter estimate for $n = 100$, 300, and 500, respectively, based on 200 replications. These three tables clearly show that the new method SMCSM has the best performance among all tested methods in Case III to Case V, but slightly worse in Cases I and II. This is natural because MLE should be the best, especially for large n when the true component density is normal (Case I). But, even in this case, SMCSM is just slightly inferior to MLE. For Cases III to V, SMCSM outperforms the other two methods, and superiority is quite significant. Based on the results of Cases III and IV, unlike MLE, SMCSM is very robust to heavy tailed component density.

For better presentation, we also provide Figs. 1–3, which plot the estimates of each method for Case I, Case III and Case V. It can be easily seen that for Case I, SMCSM and MLE work comparatively, with SPEM having bigger bias for all parameters. The superiority of SMCSM is clear in Cases III and V, where the estimates by MLE and SPEM have much bigger variation.

In the second part of the simulation study, we investigate the performance of different estimates for three-component mixture models. The model settings are similar to the above study design where we continue using normal, uniform, t_3, t_5, and Laplace distributions. Below are the list of cases we consider:

*Case*I: $f(x) \sim N(0, 1), \mu_1 = 0, \mu_2 = 3, \mu_3 = 5, \pi_1 = 0.3, \pi_2 = 0.5.$
*Case*II: $f(x) \sim U(-1, 1), \mu_1 = 0, \mu_2 = 1, \mu_3 = 3, \pi_1 = 0.3, \pi_2 = 0.5.$
*Case*III: $f(x) \sim t_3, \mu_1 = 0, \mu_2 = 3, \mu_3 = 5, \pi_1 = 0.3, \pi_2 = 0.5.$
*Case*IV: $f(x) \sim t_5, \mu_1 = 0, \mu_2 = 3, \mu_3 = 5, \pi_1 = 0.3, \pi_2 = 0.5.$
*Case*V: $f(x) \sim \text{Laplace}(0, 1), \mu_1 = -1, \mu_2 = 1, \mu_3 = 3, \pi_1 = 0.3, \pi_2 = 0.5.$

Table 3
Average (MSE) of point estimates over 200 repetitions with $n = 500$.

Case	TRUE	SMCSM	MLE	SPEM
I	$\pi_1 = 0.3$	0.319(0.001)	0.320(0.001)	0.311(0.000)
	$\mu_1 = 0$	−0.003(0.014)	−0.008(0.012)	0.003(0.013)
	$\mu_2 = 3$	2.992(0.006)	2.999(0.005)	2.955(0.008)
II	$\pi_1 = 0.3$	0.424(0.023)	0.418(0.019)	0.294(0.003)
	$\mu_1 = 0$	0.131(0.026)	0.092(0.016)	0.107(0.024)
	$\mu_2 = 1$	1.127(0.030)	1.141(0.028)	0.948(0.006)
III	$\pi_1 = 0.3$	0.303(0.001)	0.321(0.187)	0.352(0.194)
	$\mu_1 = 0$	0.017(0.026)	−5.205(101.051)	−5.003(116.853)
	$\mu_2 = 3$	3.006(0.007)	6.612(78.698)	7.022(89.414)
IV	$\pi_1 = 0.3$	0.303(0.001)	0.316(0.032)	0.302(0.033)
	$\mu_1 = 0$	0.019(0.023)	−0.312(3.785)	−0.214(4.431)
	$\mu_2 = 3$	3.005(0.007)	3.477(5.426)	3.413(6.038)
V	$\pi_1 = 0.3$	0.300(0.001)	0.263(0.050)	0.239(0.050)
	$\mu_1 = -1$	−1.001(0.008)	−1.552(2.017)	−1.193(1.806)
	$\mu_2 = 1$	0.997(0.002)	1.204(2.091)	1.096(2.213)

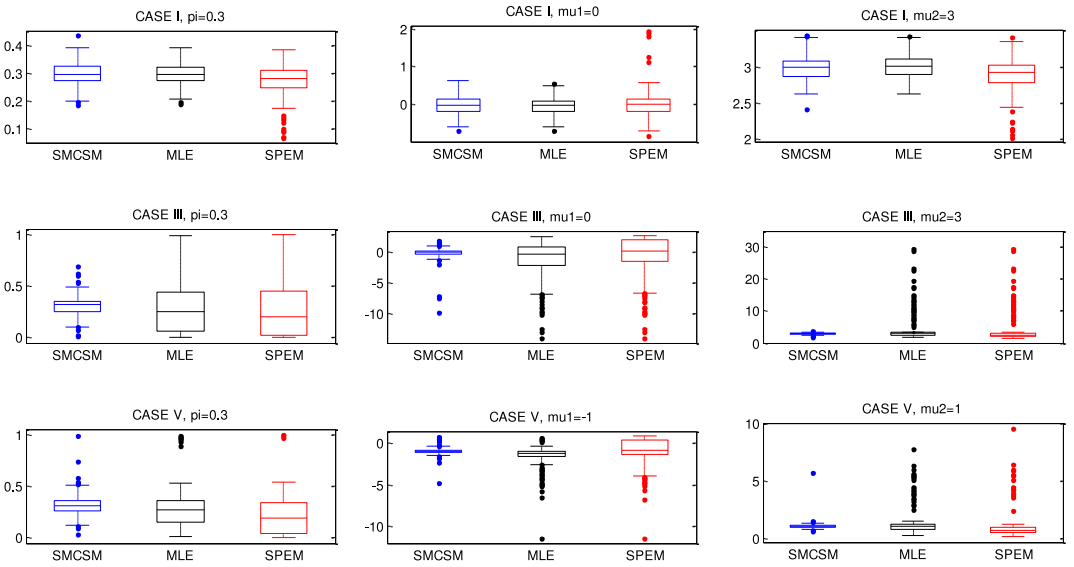

Fig. 1. Box-plot of estimates when $n = 100$.

Similarly, a total of $N = 200$ datasets with sample sizes $n = 100$, $n = 300$ and $n = 500$ were generated from each case. SMCSM, MLE and SPEM are applied to estimate $\boldsymbol{\theta} = (\pi_1, \pi_2, \mu_1, \mu_2, \mu_3)$ in each case, and the mean and MSE of each parameter estimates are summarized in Tables 4–6. SMCSM still show its superiority over MLE and SPEM for all parameter estimates and sample sizes for Case III to Case V. The improvement is especially extraordinary for mean estimates for heavy tailed distributions. For Case I and Case II, unlike the above two-component case, SMCSM performs slightly better than MLE and SPEM for all parameter estimates but a few location parameters.

4.2. Real data example

We apply our methodology to the elbow diameter data, described in Heinz et al. (2003). The dataset contains the elbow diameters of 507 physically active people, and due to the gender difference, it is highly likely that there are two clusters of observations. Of these subjects, 247 men (or 48.72%) have a sample mean of 14.46 and a sample standard deviation of 0.88. The corresponding values for the remaining female subjects are 12.37 and 0.84. Fig. 4 shows the histogram of the data.

To evaluate the number of components of this dataset, we used the likelihood based k-fold cross validation. To be more specific, let $\mathscr{D}$ be the full dataset and $\mathscr{D}_l$ be the lth partition such that $\cup_{l=1}^{k} \mathscr{D}_l = \mathscr{D}$, where k is the total number of partitions. For each partition, we use the training set $\mathscr{D} - \mathscr{D}_l$ to obtain the estimates $\hat{\boldsymbol{\theta}}^{(l)}$ and $\hat{Q}^{(l)}$ and use $\mathscr{D}_l$ for testing. Then, the

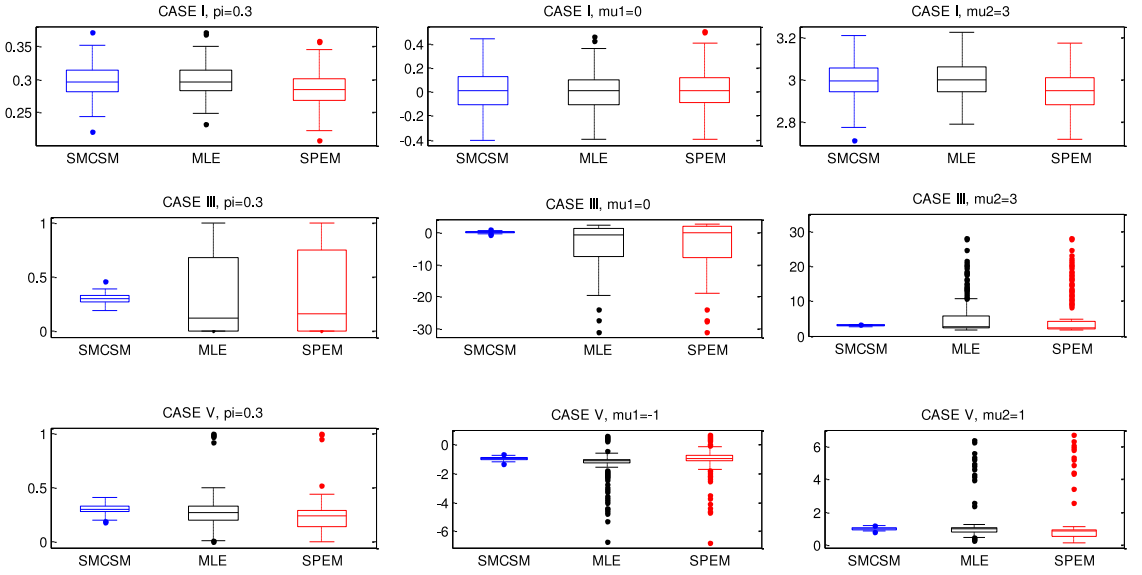

Fig. 2. Box-plot of estimates when $n = 300$.

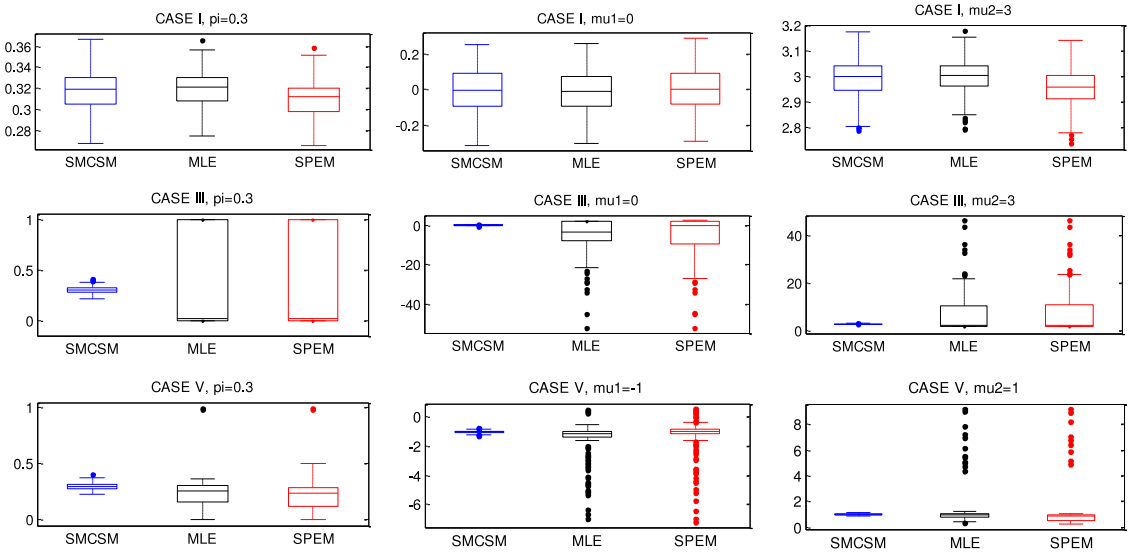

Fig. 3. Box-plot of estimates when $n = 500$.

likelihood version of CV is defined by

$$CV(m) = \sum_{l=1}^{k} \sum_{t \in \mathscr{D}_l} \log \left\{ \int \sum_{j=1}^{m} \frac{\hat{\pi}_j^{(l)}}{\sigma} \phi \left(\frac{x_t - \hat{\mu}_j^{(l)}}{\sigma} \right) d\hat{Q}^{(l)}(\sigma) \right\}. \tag{4.2}$$

We use the 10-fold cross validation and the $CV(m)$ values for $m = 1$, $m = 2$ and $m = 3$ are -90.22, -84.02 and -85.41, respectively, and therefore, $m = 2$ is selected as the number of components.

Table 7 reports the parameter estimates based on different methods, without using the gender information. For bench mark comparison, we also report the oracle value that uses the gender information. For SMCSM and MLE, we use 10 random initial values and select the converged value which has the largest likelihood. For SPEM, we use the MLE as the initial value. We can see that all three methods provide similar parameter estimates for the raw data. In addition, without using the gender information, all three methods based on mixture models can recover the mean elbow diameters for male and female, that is, provide similar estimates to the oracle one. Fig. 5 shows the estimated CDFs of $p(x)$ for different methods and the empirical CDF of $p(x)$. It can be seen that the CDFs of SPEM and SMCSM are closer to the empirical CDF than that of MLE.

Next, we check the robustness of different methods by adding outliers to the original dataset. We consider two outlier cases: Case I. five 21's (the range of original data is from 9.9 to 16.7); Case II. ten randomly generated points from $U(16, 20)$.

Table 4
Average (MSE) of point estimates over 200 repetitions with $n = 100$ for $m = 3$.

Case	TRUE	SMCSM	MLE	SPEM
	$\pi_1 = 0.3$	0.247(0.006)	0.248(0.004)	0.227(0.013)
	$\pi_2 = 0.5$	0.446(0.018)	0.500(0.024)	0.588(0.036)
I	$\mu_1 = 0$	0.103(0.243)	−0.027(0.093)	0.678(1.173)
	$\mu_2 = 3$	3.028(0.376)	3.193(0.723)	3.104(0.180)
	$\mu_3 = 5$	4.216(1.384)	4.845(0.833)	3.636(2.429)
	$\pi_1 = 0.3$	0.372(0.015)	0.468(0.047)	0.367(0.042)
	$\pi_2 = 0.5$	0.463(0.014)	0.437(0.020)	0.527(0.037)
II	$\mu_1 = 0$	0.352(0.199)	0.240(0.233)	0.524(0.324)
	$\mu_2 = 1$	0.819(0.147)	1.150(0.221)	0.848(0.132)
	$\mu_3 = 3$	2.678(0.559)	3.184(0.101)	2.992(0.084)
	$\pi_1 = 0.3$	0.276(0.004)	0.212(0.027)	0.166(0.039)
	$\pi_2 = 0.5$	0.430(0.017)	0.630(0.051)	0.711(0.089)
III	$\mu_1 = 0$	0.030(0.312)	−2.258(26.225)	−1.074(26.537)
	$\mu_2 = 3$	2.838(0.262)	2.925(1.123)	2.806(0.503)
	$\mu_3 = 5$	4.588(0.806)	7.763(26.941)	6.082(26.524)
	$\pi_1 = 0.3$	0.300(0.003)	0.281(0.012)	0.263(0.016)
	$\pi_2 = 0.5$	0.432(0.014)	0.481(0.025)	0.590(0.038)
IV	$\mu_1 = 0$	−0.005(0.219)	−0.486(2.642)	0.613(2.683)
	$\mu_2 = 3$	2.979(0.309)	3.141(1.694)	3.047(0.431)
	$\mu_3 = 5$	4.677(0.725)	5.641(5.544)	4.504(6.398)
	$\pi_1 = 0.3$	0.322(0.003)	0.295(0.017)	0.249(0.029)
	$\pi_2 = 0.5$	0.472(0.009)	0.563(0.020)	0.667(0.061)
V	$\mu_1 = -1$	−0.990(0.051)	−1.485(1.300)	−0.424(1.262)
	$\mu_2 = 1$	0.919(0.089)	0.936(0.217)	0.795(0.141)
	$\mu_3 = 3$	2.448(1.038)	3.396(1.339)	2.285(2.762)

Table 5
Average (MSE) of point estimates over 200 repetitions with $n = 300$ for $m = 3$.

Case	TRUE	SMCSM	MLE	SPEM
	$\pi_1 = 0.3$	0.298(0.001)	0.301(0.000)	0.320(0.002)
	$\pi_2 = 0.5$	0.447(0.017)	0.490(0.016)	0.525(0.010)
I	$\mu_1 = 0$	−0.003(0.041)	−0.027(0.022)	0.289(0.167)
	$\mu_2 = 3$	3.065(0.375)	3.188(0.493)	3.244(0.106)
	$\mu_3 = 5$	4.462(0.851)	4.876(0.556)	4.106(1.242)
	$\pi_1 = 0.3$	0.381(0.017)	0.464(0.036)	0.462(0.059)
	$\pi_2 = 0.5$	0.468(0.011)	0.411(0.015)	0.390(0.045)
II	$\mu_1 = 0$	0.397(0.229)	0.218(0.121)	0.456(0.255)
	$\mu_2 = 1$	0.935(0.097)	1.263(0.145)	0.935(0.065)
	$\mu_3 = 3$	2.963(0.076)	3.160(0.043)	2.966(0.031)
	$\pi_1 = 0.3$	0.277(0.002)	0.175(0.038)	0.142(0.045)
	$\pi_2 = 0.5$	0.440(0.010)	0.702(0.087)	0.745(0.107)
III	$\mu_1 = 0$	0.015(0.072)	−3.639(45.878)	−2.567(46.077)
	$\mu_2 = 3$	3.052(0.199)	3.135(0.789)	2.884(0.622)
	$\mu_3 = 5$	4.790(0.397)	8.960(45.409)	7.411(42.633)
	$\pi_1 = 0.3$	0.315(0.002)	0.307(0.012)	0.304(0.014)
	$\pi_2 = 0.5$	0.453(0.009)	0.544(0.014)	0.562(0.021)
IV	$\mu_1 = 0$	0.017(0.075)	−0.562(4.604)	0.036(4.874)
	$\mu_2 = 3$	3.016(0.153)	3.217(0.631)	3.188(0.481)
	$\mu_3 = 5$	4.862(0.393)	6.241(8.022)	4.982(7.910)
	$\pi_1 = 0.3$	0.308(0.002)	0.263(0.022)	0.199(0.029)
	$\pi_2 = 0.5$	0.505(0.005)	0.581(0.024)	0.702(0.063)
V	$\mu_1 = -1$	−0.982(0.049)	−1.633(2.574)	−0.464(2.218)
	$\mu_2 = 1$	1.006(0.086)	0.868(0.244)	0.791(0.099)
	$\mu_3 = 3$	2.788(0.389)	3.445(1.698)	2.484(2.279)

The results are also reported in Table 7. From the table, we can see that SMCSM with added outliers provides almost the same estimates as SMCSM without outliers. However, both MLE and SPEM are very sensitive to the outliers. MLE tends to fit both types of contaminated data with one component containing only the outliers and the other component containing the rest of the data. SPEM performs similarly to MLE for Case I outliers, and provides a one component fit for Case II outliers.

To check whether MLE and SPEM might pick up the outliers as a third component, we fit the contaminated data with a three-component model and estimate the parameters with MLE or SPEM. The results are reported in Table 8. For outlier

Table 6
Average (MSE) of point estimates over 200 repetitions with $n = 500$ for $m = 3$.

Case	TRUE	SMCSM	MLE	SPEM
	$\pi_1 = 0.3$	0.309(0.001)	0.311(0.000)	0.338(0.002)
	$\pi_2 = 0.5$	0.443(0.014)	0.466(0.012)	0.471(0.006)
I	$\mu_1 = 0$	0.023(0.024)	0.010(0.016)	0.262(0.112)
	$\mu_2 = 3$	3.180(0.495)	3.186(0.434)	3.267(0.114)
	$\mu_3 = 5$	4.624(0.644)	4.840(0.386)	4.387(0.594)
	$\pi_1 = 0.3$	0.374(0.015)	0.464(0.034)	0.512(0.064)
	$\pi_2 = 0.5$	0.444(0.011)	0.386(0.020)	0.304(0.059)
II	$\mu_1 = 0$	0.359(0.200)	0.346(0.294)	0.397(0.189)
	$\mu_2 = 1$	0.924(0.103)	1.183(0.213)	0.970(0.037)
	$\mu_3 = 3$	2.936(0.038)	3.126(0.024)	2.919(0.024)
	$\pi_1 = 0.3$	0.277(0.002)	0.175(0.038)	0.142(0.045)
	$\pi_2 = 0.5$	0.440(0.010)	0.702(0.087)	0.745(0.107)
III	$\mu_1 = 0$	0.015(0.072)	−3.639(45.878)	−2.567(46.077)
	$\mu_2 = 3$	3.052(0.199)	3.135(0.789)	2.884(0.622)
	$\mu_3 = 5$	4.790(0.397)	8.960(45.409)	7.411(42.633)
	$\pi_1 = 0.3$	0.291(0.001)	0.257(0.023)	0.265(0.015)
	$\pi_2 = 0.5$	0.475(0.006)	0.561(0.027)	0.636(0.041)
IV	$\mu_1 = 0$	0.015(0.032)	−1.396(13.151)	−0.514(10.758)
	$\mu_2 = 3$	3.039(0.105)	3.151(1.543)	3.212(0.390)
	$\mu_3 = 5$	4.884(0.277)	6.397(14.045)	5.485(12.951)
	$\pi_1 = 0.3$	0.321(0.001)	0.294(0.028)	0.265(0.035)
	$\pi_2 = 0.5$	0.484(0.002)	0.527(0.014)	0.594(0.041)
V	$\mu_1 = -1$	−0.997(0.010)	−1.543(2.761)	−0.709(2.624)
	$\mu_2 = 1$	1.024(0.047)	0.944(0.412)	0.874(0.113)
	$\mu_3 = 3$	2.912(0.126)	3.403(1.899)	2.603(2.410)

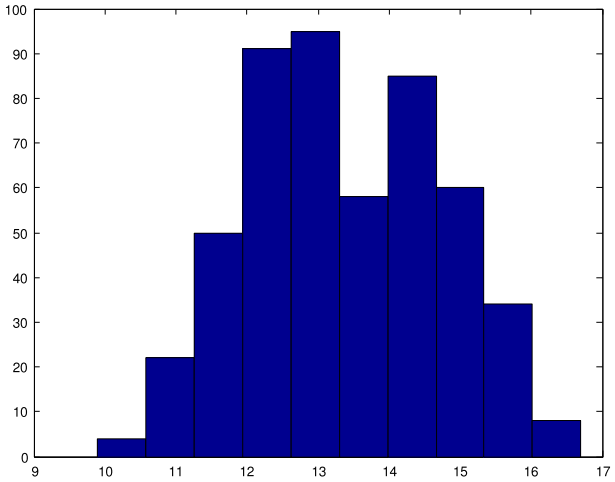

Fig. 4. Histogram of the elbow diameter data.

Case I, MLE and SPEM indeed pick up the outliers as a third component, and can estimate the other two components very well. However, for outlier Case II, neither of the two methods can recover the original two components.

5. Discussion

The method proposed in this paper utilizes nonparametric normal scale mixture models to specify the nonparametric symmetric component density. This enables us to estimate parametric and nonparametric components in the model simultaneously without modifying the likelihood. The existing methods rely on the kernel density estimator, which requires the selection of bandwidth, and can have much effect on model performance. On the other hand, there is no such selection issue for the proposed method. Hence, it can give a more efficient and reliable estimator than others. In addition, the proposed method is robust to any potential outlier as it automatically downweights observations far away from the center of each component.

The choice of the number of components has long been a difficult problem for mixture models. As far as we know, there is no established model selection procedure for semiparametric mixture models. We proposed to use a likelihood based cross

Table 7
Parameter estimates for the elbow diameter data based on different estimation methods.

	Method	π_1	μ_1	μ_2
	Oracle value	0.513	12.37	14.46
Original data				
	SMCSM	0.564	12.47	14.56
	MLE	0.561	12.46	14.56
	SPEM	0.557	12.52	14.47
Data with five "21"				
	SMCSM	0.522	12.42	14.33
	MLE	0.990	13.39	20.99
	SPEM	0.990	13.39	21.00
Data with ten outliers from $U(16, 20)$				
	SMCSM	0.571	12.53	14.61
	MLE	0.984	13.39	18.37
	SPEM	0.869	13.48	13.49

Table 8
Parameter estimates for the elbow diameter data with 3-component model.

	Method	π_1	π_2	π_3	μ_1	μ_2	μ_3
Data with five "21"							
	MLE	0.555	0.435	0.010	12.458	14.569	21.001
	SPEM	0.556	0.434	0.010	12.513	14.505	21.001
Data with ten outliers from $U(16, 20)$							
	MLE	0.551	0.433	0.016	13.38	13.42	18.37
	SPEM	0.549	0.433	0.018	13.39	13.40	18.24

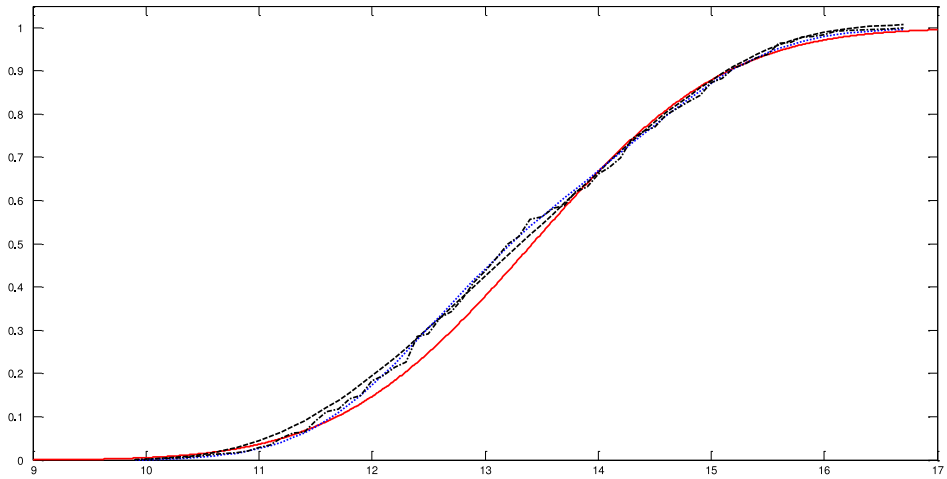

Fig. 5. The CDFs of $p(x)$ for different methods and the empirical CDF of $p(x)$: SMCSM (blue dotted line), MLE (red solid line), SPEM (black dashed line) and empirical CDF (black dash dotted line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

validation method to choose the number of components and demonstrated its empirical success in our real data application in Section 4.2. But it requires more research to find some theoretically justified model selection methods for semiparametric mixture models. Bordes et al. (2006) and Hunter et al. (2007) have investigated extensively the identifiability of the model (1.1) and proved its identifiability when $m \leq 3$. However, their identifiability results cannot be extended to the case where $m > 3$. It requires more research to establish the identifiability of the model (1.1) for $m > 3$.

As another future work, one may extend the proposed method to semiparametric multivariate mixtures. In this case, one can apply nonparametric multivariate scale mixtures for the nonparametric component densities. Based on our description in this paper, it is plausible at least theoretically but may require a huge computing time especially for high dimensional mixtures. In fact, there is no rigorous study for the estimation algorithm in nonparametric multivariate mixture though existing methods may still be applicable. This extension should be further investigated in the future. In addition, we can also extend the proposed method to the two component semiparametric mixture model when one component is known while the other is symmetric but otherwise arbitrary (Bordes et al., 2006; Xiang et al., 2014; Ma and Yao, 2015). This

semiparametric model has wide applications in many areas such as large-scale simultaneous testing/multiple testing, sequential clustering, and robust modeling.

Acknowledgments

The authors wish to thank the Associate Editor and two referees for their helpful comments and suggestions that have led to significant improvements of this paper. Xiang's research is supported by Zhejiang Provincial NSF of China grant LQ16A010002, and Zhejiang Provincial Education Department scientific research project grant Y201534431. Yao's research is supported by National Science Foundation (NSF) grant DMS-1461677. The research of Byungtae Seo was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2057715).

References

Andrews, D.F., Mallows, C.L., 1974. Scale mixtures of normal distributions. J. R. Stat. Soc. Ser. B Stat. Methodol. 36, 99–102.
Atwood, C.L., 1976. Convergent design sequences for sufficiently regular optimality criteria. Ann. Statist. 4, 1124–1138.
Basu, S., 1996. Existence of a normal scale mixture with a given variance and a percentile. Statist. Probab. Lett. 28, 115–120.
Benaglia, T., Chauveau, D., Hunter, D.R., 2009. An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. J. Comput. Graph. Statist. 18, 505–526.
Böhning, D., 1982. Convergence of Simar's algorithm for finding the MLE of a compound Poisson process. Ann. Statist. 10, 1006–1008.
Böhning, D., 1986. A vertex-exchange-method in D-optimal design theory. Metrika 33, 337–347.
Böhning, D., 1999. Computer-Assisted Analysis of Mixtures and Applications. Chapman and Hall/CRC, Boca Raton, FL.
Böhning, D., Ruangroj, R., 2002. A note on the maximum deviation of the scale-contaminated normal to the best normal distribution. Metrika 55, 177–182.
Bordes, L., Chauveau, D., Vandekerkhove, P., 2007. An EM algorithm for a semiparametric mixture model. Comput. Statist. Data Anal. 51, 5429–5443.
Bordes, L., Mottelet, S., Vandekerkhove, P., 2006. Semiparametric estimation of a two-component mixture model. Ann. Statist. 34, 1204–1232.
Celeux, G., Hurn, M., Robert, C.P., 2000. Computational and inferential difficulties with mixture posterior distributions. J. Amer. Statist. Assoc. 95, 957–970.
Chee, C., Wang, Y., 2013. Estimation of finite mixtures with symmetric components. Stat. Comput. 23, 233–249.
Chen, J., Tan, X., Zhang, R., 2008. Inference for normal mixture in mean and variance. Statist. Sincia 18, 443–465.
Efron, B., Olshen, R.A., 1978. How broad is the class of normal scale mixtures? Ann. Statist. 6, 1159–1164.
Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer.
Hathaway, R.J., 1985. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann. Statist. 13, 795–800.
Heinz, G., Peterson, L.J., Johnson, R.W., Kerk, C.J., 2003. Exploring relationships in body dimensions. J. Stat. Educ. 11.
Hunter, D., Wang, S., Hettmansperger, T.P., 2007. Inference for mixtures of symmetric distributions. Ann. Statist. 35, 224–251.
Kelker, D., 1971. Infinite divisibility and variance mixtures of the normal distribution. Ann. Math. Stat. 42, 802–808.
Kiefer, J., Wolfowitz, J., 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Stat. 27, 886–906.
Laird, N.M., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. J. Amer. Statist. Assoc. 73, 805–811.
Lesperance, M.L., Kalbfleisch, J.D., 1992. An algorithm for computing the nonparametric MLE of a mixing distribution. J. Amer. Statist. Assoc. 87, 120–126.
Lindsay, B.G., 1983a. The geometry of mixture likelihoods: a general theory. Ann. Statist. 11, 86–94.
Lindsay, B.G., 1983b. The geometry of mixture likelihoods, Part II: The exponential family. Ann. Statist. 11, 783–792.
Lindsay, B.G., 1995. Mixture Models: Theory, Geometry, and Applications. In: NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 5. Institute of Mathematical Statistics, Hayward, CA.
Ma, Y., Yao, W., 2015. Flexible estimation of a semiparametric two-component mixture model with one parametric component. Electron. J. Stat. 9, 444–474.
McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.
Seo, B., Kim, D., 2012. Root selection in normal mixture models. Comput. Statist. Data Anal. 56 (8), 2454–2470.
Seo, B., Lee, T., 2015. A new algorithm for maximum likelihood estimation in normal scale-mixture generalized autoregressive conditional heteroskedastic models. J. Stat. Comput. Simul. 85, 202–215.
Seo, B., Noh, J., Lee, T., Yoon, Y., Adaptive robust regression with continuous Gaussian scale mixture errors. submitted for publication.
Stephens, M., 2000. Dealing with label switching in mixture models. J. R. Stat. Soc. Ser. B 62, 795–809.
Tanaka, K., Takemura, A., 2006. Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when the scale parameters are exponentially small. Bernoulli 12, 1003–1017.
Wang, Y., 2007. On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. J. R. Stat. Soc. Ser. B Stat. Methodol. 69, 185–198.
Wu, C.F., 1978. Some algorithmic aspects of the theory of optimal designs. Ann. Statist. 6, 1286–1301.
Wynn, H.P., 1970. The sequential generation of D-optimum experimental designs. Ann. Math. Statist. 41, 1655–1664.
Wynn, H.P., 1972. Results in the theory and construction of D-optimum experimental designs. J. R. Stat. Soc. Ser. B Stat. Methodol. 34, 133–147.
Xiang, S., Yao, W., Wu, J., 2014. Minimum profile hellinger distance estimation for a semiparametric mixture model. Canad. J. Statist. 42 (2), 246–267.
Yao, W., 2010. A profile likelihood method for normal mixture with unequal variance. J. Statist. Plann. Inference 140, 2089–2098.
Yao, W., 2015. Label switching and its simple solutions for frequentist mixture models. J. Stat. Comput. Simul. 85, 1000–1012.
Yao, W., Lindsay, B.G., 2009. Bayesian mixture labeling by highest posterior density. J. Amer. Statist. Assoc. 104, 758–767.