Advances in Data Analysis and Classification (2020) 14:261–292
https://doi.org/10.1007/s11634-020-00392-w

REGULAR ARTICLE

Semiparametric mixtures of regressions with single-index for model based clustering

Sijia Xiang[1] · **Weixin Yao**[2]

Received: 25 October 2018 / Revised: 27 January 2020 / Accepted: 6 March 2020 /
Published online: 23 April 2020

Abstract
In this article, we propose two classes of semiparametric mixture regression models with single-index for model based clustering. Unlike many semiparametric/nonparametric mixture regression models that can only be applied to low dimensional predictors, the new semiparametric models can easily incorporate high dimensional predictors into the nonparametric components. The proposed models are very general, and many of the recently proposed semiparametric/nonparametric mixture regression models are indeed special cases of the new models. Backfitting estimates and the corresponding modified EM algorithms are proposed to achieve optimal convergence rates for both parametric and nonparametric parts. We establish the identifiability results of the proposed two models and investigate the asymptotic properties of the proposed estimation procedures. Simulation studies are conducted to demonstrate the finite sample performance of the proposed models. Two real data applications using the new models reveal some interesting findings.

Keywords EM algorithm · Kernel regression · Mixture regression model · Model based clustering · Single-index model

Mathematics Subject Classification 62G08 · 62E20

1 Introduction

Mixtures of regression models are commonly used as "model based clustering" methods to reveal the relationship among variables of interest if the population consists

✉ Sijia Xiang
sjxiang@zufe.edu.cn

Weixin Yao
weixin.yao@ucr.edu

1 School of Data Sciences, Zhejiang University of Finance & Economics, Hangzhou, Zhejiang 310018, People's Republic of China

2 Department of Statistics, University of California, Riverside, CA, USA

of several homogeneous subgroups. This type of application is commonly seen in econometrics, where it is also known as switching regression models, and in various other fields, see, for example, in econometrics (Wedel and DeSarbo 1993; Frühwirth-Schnatter 2001), and in epidemiology (Green and Richardson 2002). Another wide application of finite mixture of regressions is in outlier detection or robust regression estimation (Young and Hunter 2010). Traditional mixture of linear regression models require strong parametric assumptions: linear component regression functions, constant component variances, and constant component proportions. The fully parametric hierarchical mixtures of experts model (Jordan and Jacobs 1994) has been proposed to allow the component proportions to depend on the covariates in machine learning. Recently, many semiparametric and nonparametric mixture regression models have been proposed to relax the parametric assumptions of mixture of regression models. See, for example, Young and Hunter (2010), Huang and Yao (2012), Cao and Yao (2012),Huang et al. (2013, 2014), Hu et al. (2017), Xiang and Yao (2018), among others. Xiang et al. (2019) provided a good review of many semiparametric regression models. However, most of those existing semiparametric or nonparametric mixture regressions can only be applied to low dimensional predictors due to the "curse of dimensionality". It will be desirable to be able to relax parametric assumptions of traditional mixtures of regression models when the dimension of predictors is high.

In this article, we propose a mixture of single-index models (MSIM) and a mixture of regression models with varying single-index proportions (MRSIP) to reduce the dimension of high dimensional predictors before modeling them nonparametrically. Many existing popular models can be considered as special cases of the proposed two models. Huang et al. (2013) proposed the nonparametric mixture of regression models

$$f(y|x,\pi,m,\sigma^2)=\sum_{j=1}^{k}\pi_j(x)\phi(y|m_j(x),\sigma_j^2(x)),$$

where $\pi_j(x)$, $m_j(x)$, and $\sigma_j^2(x)$ are unknown smoothing functions, and $\phi(y|\mu,\sigma^2)$ is the normal density with mean μ and variance σ^2. Their proposed model can drastically reduce the modeling bias when the strong parametric assumption of traditional mixture of linear regression models does not hold. However, the above model is not applicable to high dimensional predictors due to the kernel estimation used for nonparametric parts. To solve the above problem, we propose a *mixture of single-index models*

$$f(y|x,\boldsymbol{\alpha},\pi,m,\sigma^2)=\sum_{j=1}^{k}\pi_j(\boldsymbol{\alpha}^\top\boldsymbol{x})\phi(y|m_j(\boldsymbol{\alpha}^\top\boldsymbol{x}),\sigma_j^2(\boldsymbol{\alpha}^\top\boldsymbol{x})), \tag{1}$$

in which the single index $\boldsymbol{\alpha}^\top\boldsymbol{x}$ transfers the high dimensional nonparametric problem to a univariate nonparametric problem. When $k=1$, model (1) reduces to a single index model (Ichimura 1993; Härdle et al. 1993). If $\boldsymbol{x}$ is a scalar, then model (1) reduces to the nonparametric mixture of regression models proposed by Huang et al. (2013). Zeng (2012) also applied the single index idea to the component means and variances

and assumed that component proportions do not depend on the predictor $\boldsymbol{x}$. However, Zeng (2012) did not give any theoretical properties of their proposed estimates.

Young and Hunter (2010) and Huang and Yao (2012) proposed a semiparametric mixture of regression models

$$f(y|x,\pi,\boldsymbol{\beta},\sigma^2)=\sum_{j=1}^{k}\pi_j(\boldsymbol{x})\phi(y|\boldsymbol{x}^\top\boldsymbol{\beta}_j,\sigma_j^2),$$

where $\pi_j(\boldsymbol{x})$'s are unknown smoothing functions, to combine nice properties of both nonparametric mixture regression models and traditional parametric mixture regression models. Their semiparametric mixture models assume that component proportions depend on covariates nonparametrically to reduce the modeling bias while component regression functions are still assumed to be linear to have better model interpretation. However, their estimation procedures cannot be applied if the dimension of predictors $\boldsymbol{x}$ is high due to the kernel estimation used for $\pi_j(\boldsymbol{x})$. We propose a *mixture of regression models with varying single-index proportions*

$$f(y|x,\boldsymbol{\alpha},\pi,\boldsymbol{\beta},\sigma^2)=\sum_{j=1}^{k}\pi_j(\boldsymbol{\alpha}^\top\boldsymbol{x})\phi(y|\boldsymbol{x}^\top\boldsymbol{\beta}_j,\sigma_j^2), \tag{2}$$

which uses the idea of single index to model the nonparametric effect of predictors on component proportions, while allowing easy interpretation of linear component regression functions. When $k=1$, model (2) reduces to the traditional linear regression model. If $\boldsymbol{x}$ is a scalar, then model (2) reduces to the semiparametric mixture models considered by Young and Hunter (2010) and Huang and Yao (2012). Modeling component proportions nonparametrically can reduce the modeling bias and better cluster the data when the traditional parametric assumptions of component proportions do not hold (Young and Hunter 2010; Huang and Yao 2012).

We prove the identifiability results of the two models under some mild conditions. We propose a modified EM algorithm, which combines the ideas of backfitting algorithm, kernel estimation, and local likelihood, to estimate both the global parameters and the nonparametric functions. In addition, the asymptotic properties of the proposed estimation procedures are also investigated. Simulation studies are conducted to demonstrate the finite sample performance of the proposed models. Two real data applications reveal some new interesting findings.

The rest of the paper is organized as follows. In Sect. 2, we introduce the MSIM and study its identifiability result. A one-step and a fully-iterated backfitting estimate are proposed, and their asymptotic properties are also studied. In Sect. 3, the MRSIP is proposed. The identifiability result and the asymptotic properties of the proposed estimates are given. In Sects. 4 and 5, Monte Carlo studies and two real data examples are illustrated to demonstrate the finite sample performance of the two models. A discussion section is given in Sect. 6 and we defer the technical conditions and proofs to the "Appendix".

2 Mixture of single-index models

2.1 Model definition and identifiability

Assume that $\{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$ is a random sample from the population $(\boldsymbol{x}, Y)$, where $\boldsymbol{x}$ is p-dimensional and Y is univariate. Let $\mathscr{C}$ be a latent variable, and has a discrete distribution $P(\mathscr{C} = j|\boldsymbol{x}) = \pi_j(\boldsymbol{\alpha}^\top \boldsymbol{x})$ for $j = 1, \ldots, k$. Conditional on $\mathscr{C} = j$ and $\boldsymbol{x}$, Y follows a normal distribution with mean $m_j(\boldsymbol{\alpha}^\top \boldsymbol{x})$ and variance $\sigma_j^2(\boldsymbol{\alpha}^\top \boldsymbol{x})$. Without observing $\mathscr{C}$, the conditional distribution of Y given $\boldsymbol{x}$ can be written as:

$$f(y|x, \boldsymbol{\alpha}, \pi, m, \sigma^2) = \sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^\top \boldsymbol{x}) \phi(y|m_j(\boldsymbol{\alpha}^\top \boldsymbol{x}), \sigma_j^2(\boldsymbol{\alpha}^\top \boldsymbol{x})).$$

The above model is the proposed mixture of single-index models. Throughout the paper, we assume that k is fixed, and refer to model (1) as a finite semiparametric mixture of regression models, since $\pi_j(\cdot)$, $m_j(\cdot)$ and $\sigma_j^2(\cdot)$ are all nonparametric. In the model (1), we use the same index $\boldsymbol{\alpha}$ for all components. But our proposed estimation procedure and asymptotic results can be easily extended to the cases where components have different index $\boldsymbol{\alpha}$.

Compared to Huang et al. (2013), the appeal of the proposed MSIM is that by using an index $\boldsymbol{\alpha}^\top \boldsymbol{x}$, the so-called "curse of dimensionality" in fitting multivariate nonparametric regression functions is avoided. It is of dimension-reduction structure in the sense that, given the estimate of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}$, we can use the univariate $\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}$ as the covariate and simplify model (1) by the nonparametric mixture regression model proposed by Huang et al. (2013). Therefore, model (1) is a reasonable compromise between fully parametric and fully nonparametric modeling.

Identifiability is a major concern for most mixture models. Some well known identifiability results of finite mixture models include: mixture of univariate normals is identifiable up to relabeling (Titterington et al. 1985) and finite mixture of regression models is identifiable up to relabeling provided that covariates have a certain level of variability (Henning 2000). Wang et al. (2014) established some general identifiability results for many existing nonparametric or semiparametric mixture regression models. The following theorem establishes the identifiability result of the model (1) and its proof is given in the "Appendix".

Theorem 1 *Assume that*

1. $\pi_j(z)$, $m_j(z)$, *and* $\sigma_j^2(z)$ *are differentiable and not constant on the support of* $\boldsymbol{\alpha}^\top \boldsymbol{x}$, $j = 1, \ldots, k$;
2. *The* $\boldsymbol{x}$ *is continuously distributed random variable that has a joint probability density function;*
3. *The support of* $\boldsymbol{x}$ *is not contained in any proper linear subspace of* $\mathbb{R}^p$;
4. $\|\boldsymbol{\alpha}\| = 1$ *and the first nonzero element of* $\boldsymbol{\alpha}$ *is positive;*
5. *For any* $1 \leq i \neq j \leq k$,

$$\sum_{l=0}^{1} \|m_i^{(l)}(z) - m_j^{(l)}(z)\|^2 + \sum_{l=0}^{1} \|\sigma_i^{(l)}(z) - \sigma_j^{(l)}(z)\|^2 \neq 0,$$

for any z where $g^{(l)}$ is the lth derivative of g and equal to g if $l = 0$.

Then, model (1) is identifiable.

2.2 Estimation procedure

In this subsection, we propose a one-step estimation procedure and a backfitting algorithm to estimate the nonparametric functions and the single index of the model (1).

Let $\ell^{*(1)}(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha})$ be the log-likelihood of the collected data $\{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$ from the model (1). That is:

$$\ell^{*(1)}(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^\top \boldsymbol{x}_i) \phi(Y_i | m_j(\boldsymbol{\alpha}^\top \boldsymbol{x}_i), \sigma_j^2(\boldsymbol{\alpha}^\top \boldsymbol{x}_i)) \right\}, \quad (3)$$

where $\phi(y|\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2, $\boldsymbol{\pi}(\cdot) = \{\pi_1(\cdot), \ldots, \pi_{k-1}(\cdot)\}^\top$, $\boldsymbol{m}(\cdot) = \{m_1(\cdot), \ldots, m_k(\cdot)\}^\top$, and $\boldsymbol{\sigma}^2(\cdot) = \{\sigma_1^2(\cdot), \ldots, \sigma_k^2(\cdot)\}^\top$. Since $\boldsymbol{\pi}(\cdot), \boldsymbol{m}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ consist of nonparametric functions, (3) is not ready for maximization.

Note that for the model (1), the space spanned by the single index $\boldsymbol{\alpha}$ is in fact the central mean subspace of $Y|\boldsymbol{x}$ (Cook and Li 2002) in the literature of sufficient dimension reduction. Therefore, we can employ existing sufficient dimension reduction methods to find an initial estimate of $\boldsymbol{\alpha}$. Please see, for example, Li (1991), Li et al. (2005), Wang and Xia (2008), Luo et al. (2009), Wang and Yao (2012), Ma and Zhu (2012, 2013), Yao et al. (2019). In this article, we will simply employ sliced inverse regression (Li 1991) to obtain an initial estimate of $\boldsymbol{\alpha}$, denoted by $\tilde{\boldsymbol{\alpha}}$.

Given the estimated single index $\tilde{\boldsymbol{\alpha}}$, the nonparametric functions $\boldsymbol{\pi}(z)$, $\boldsymbol{m}(z)$ and $\boldsymbol{\sigma}^2(z)$ can then be estimated by maximizing the following local log-likelihood function:

$$\ell_1^{(1)}(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i) \phi(Y_i | m_j(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i), \sigma_j^2(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)) \right\} K_h(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z), \quad (4)$$

where $K_h(z) = \frac{1}{h} K(\frac{z}{h})$, $K(\cdot)$ is a kernel density function, and h is a tuning parameter. Let $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$ and $\hat{\boldsymbol{\sigma}}^2(\cdot)$ be the estimates that maximize (4). The above estimates are the proposed *one-step estimate*.

We propose a modified EM-type algorithm to maximize $\ell_1^{(1)}$. In practice, we usually want to evaluate unknown functions at a set of grid points, which in this case, requires us to maximize local log-likelihood functions at a set of grid points. If we simply employ

the EM algorithm separately for each grid point, the labels in the EM algorithm may change at different grid points, and we may not be able to get smoothed estimated curves (Huang and Yao 2012). Therefore, we propose the following modified EM-type algorithm, which estimates the nonparametric functions simultaneously at a set of grid points, say $\{u_t, t = 1, \ldots, N\}$, and provides a unified label for each observation across all grid points.

Algorithm 1 *Modified EM-type algorithm to maximize* (4) *given the single index estimate* $\tilde{\boldsymbol{\alpha}}$.

E-step: Calculate the expectations of component labels based on estimates from lth iteration:

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)\phi(Y_i|m_j^{(l)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i), \sigma_j^{2(l)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i))}{\sum_{j=1}^k \pi_j^{(l)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)\phi(Y_i|m_j^{(l)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i), \sigma_j^{2(l)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i))}, \tag{5}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, k$.

M-step: Update the estimates

$$\pi_j^{(l+1)}(z) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z)}{\sum_{i=1}^n K_h(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z)}, \tag{6}$$

$$m_j^{(l+1)}(z) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} Y_i K_h(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z)}, \tag{7}$$

$$\sigma_j^{2(l+1)}(z) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - m_j^{(l+1)}(z))^2 K_h(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z)}, \tag{8}$$

for $z \in \{u_t, t = 1, \ldots, N\}$ *and* $j = 1, \ldots, k$. *We then update* $\pi_j^{(l+1)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)$, $m_j^{(l+1)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)$ *and* $\sigma_j^{2(l+1)}(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)$, $i = 1, \ldots, n$, *by linear interpolating* $\pi_j^{(l+1)}(u_t)$, $m_j^{(l+1)}(u_t)$ *and* $\sigma_j^{2(l+1)}(u_t)$, $t = 1, \ldots, N$, *respectively.*

Note that in the M-step, the nonparametric functions are estimated simultaneously at a set of grid points, and therefore, the classification probabilities in the the E-step can be estimated globally to avoid the label switching problem (Stephens 2000; Yao and Lindsay 2009). If the sample size n is not too large, one can also take all $\{\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i, i = 1, \ldots, n\}$ as grid points for z in the M-step.

The initial estimate $\tilde{\boldsymbol{\alpha}}$ by SIR does not make use of the mixture information and thus is not efficient. Given one step estimate $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$ and $\hat{\boldsymbol{\sigma}}^2(\cdot)$, we can further improve the estimate of $\boldsymbol{\alpha}$ by maximizing

$$\ell_2^{(1)}(\boldsymbol{\alpha}) = \sum_{i=1}^n \log \left\{ \sum_{j=1}^k \hat{\pi}_j(\boldsymbol{\alpha}^\top \boldsymbol{x}_i)\phi(Y_i|\hat{m}_j(\boldsymbol{\alpha}^\top \boldsymbol{x}_i), \hat{\sigma}_j^2(\boldsymbol{\alpha}^\top \boldsymbol{x}_i)) \right\}, \tag{9}$$

with respect to $\boldsymbol{\alpha}$. The proposed *fully iterative backfitting estimator* of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}$, iterates the above two steps until convergence.

Algorithm 2 *Fully iterative backfitting estimator* (*FIB*)

Step 1: Apply sliced inverse regression (*SIR*) *to obtain an initial estimate of the single index parameter* $\boldsymbol{\alpha}$*, denoted by* $\tilde{\boldsymbol{\alpha}}$.

Step 2: Given $\tilde{\boldsymbol{\alpha}}$*, apply the modified EM-algorithm* (5)–(8) *to maximize* $\ell_1^{(1)}$ *in* (4) *to obtain the estimates* $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$, *and* $\hat{\boldsymbol{\sigma}}^2(\cdot)$.

Step 3: Given $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$, *and* $\hat{\boldsymbol{\sigma}}^2(\cdot)$ *from Step 2, update the estimate of* $\boldsymbol{\alpha}$ *by maximizing* $\ell_2^{(1)}$ *in* (9).

Step 4: Iterate Steps 2–3 until convergence.

2.3 Asymptotic properties

The asymptotic properties of the proposed estimates are investigated below. Let $\boldsymbol{\theta}(z) = (\boldsymbol{\pi}^\top(z), \boldsymbol{m}^\top(z), (\boldsymbol{\sigma}^2)^\top(z))^\top$. Define

$$\begin{aligned}
\ell(\boldsymbol{\theta}(z), y) &= \log \sum_{j=1}^{k} \pi_j(z)\phi\{y|m_j(z), \sigma_j^2(z)\},\\
q_1(z) &= \frac{\partial \ell(\boldsymbol{\theta}(z), y)}{\partial \theta},\\
q_2(z) &= \frac{\partial^2 \ell(\boldsymbol{\theta}(z), y)}{\partial \theta \partial \theta^\top},\\
\mathscr{I}_\theta^{(1)}(z) &= -E[q_2(Z)|Z = z],\\
\Lambda_1(u|z) &= E[q_1(z)|Z = u].
\end{aligned}$$

Under further conditions defined in the "Appendix", the asymptotic properties of the one-step estimates $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$, and $\hat{\boldsymbol{\sigma}}^2(\cdot)$ are given in the following theorem.

Theorem 2 *Assume that conditions* (*C*1)–(*C*7) *in the "Appendix" hold. Then, as* $n \to \infty$, $h \to 0$ *and* $nh \to \infty$, *we have*

$$\sqrt{nh}\{\hat{\boldsymbol{\theta}}(z) - \boldsymbol{\theta}(z) - \mathscr{B}_1 + o_p(h^2)\} \xrightarrow{D} N\{0, \nu_0 f^{-1}(z)\mathscr{I}_\theta^{(1)}(z)\}, \tag{10}$$

where

$$\mathscr{B}_1(z) = \mathscr{I}_\theta^{(1)-1}\left\{\frac{f'(z)\Lambda_1'(z|z)}{f(z)} + \frac{1}{2}\Lambda_1''(z|z)\right\}\kappa_2 h^2,$$

with $f(\cdot)$ *the marginal density function of* $\boldsymbol{\alpha}^\top \boldsymbol{x}$, $\kappa_l = \int t^l K(t)dt$ *and* $\nu_l = \int t^l K^2(t)dt$.

Note that the asymptotic variance of $\hat{\boldsymbol{\theta}}(z)$ is the same as those given in Huang et al. (2013). Thus, the nonparametric functions can be estimated with the same accuracy

as it would have if the single index $\boldsymbol{\alpha}^\top \boldsymbol{x}$ were known. This is expected since the index $\boldsymbol{\alpha}$ can be estimated at a root n convergence rate which is faster than $\hat{\boldsymbol{\theta}}(z)$. In addition, note that the one-step estimates of $\boldsymbol{\theta}(z)$ have the same asymptotic variance (up to the first order) as the fully iterative backfitting algorithm but with much less computations. Our simulation results in Sect. 4 further confirm this result.

The next theorem gives the asymptotic results of the $\hat{\boldsymbol{\alpha}}$ given by the fully iterative backfitting algorithm.

Theorem 3 *Assume that conditions* $(C1)$–$(C8)$ *in the "Appendix" hold. Then, as* $n \to \infty$, $nh^4 \to 0$, *and* $nh^2/\log(1/h) \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{D} N(0, \boldsymbol{Q}_1^{-1}), \tag{11}$$

where

$$\boldsymbol{Q}_1 = E\left[\{\boldsymbol{x}\boldsymbol{\theta}'(Z)\}q_2(Z)\{\boldsymbol{x}\boldsymbol{\theta}'(Z)\}^\top - \boldsymbol{x}\boldsymbol{\theta}'(Z)q_2(Z)\mathscr{I}_\theta^{(1)-1}(Z)E\{q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^\top|Z\}\right].$$

3 Mixtures of regression models with varying single-index proportions

3.1 Model definition and identifiability

The MRSIP assumes that $P(\mathscr{C} = j|\boldsymbol{x}) = \pi_j(\boldsymbol{\alpha}^\top \boldsymbol{x})$ for $j = 1, \ldots, k$, and conditional on $\mathscr{C} = j$ and $\boldsymbol{x}$, Y follows a normal distribution with mean $\boldsymbol{x}^\top \boldsymbol{\beta}_j$ and variance σ_j^2. That is,

$$f(y|x, \boldsymbol{\alpha}, \pi, \boldsymbol{\beta}, \sigma^2) = \sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^\top \boldsymbol{x})\phi(y|\boldsymbol{x}^\top \boldsymbol{\beta}_j, \sigma_j^2).$$

Since $\pi_j(\cdot)$'s are nonparametric, model (2) is also a finite semiparametric mixture of regression models. The linear component regression functions $\boldsymbol{x}^\top \boldsymbol{\beta}_j$ enjoy simple interpretation, while nonparametric functions $\pi_j(\boldsymbol{\alpha}^\top \boldsymbol{x})$ can incorporate the effects of predictors on component proportions more flexibly to reduce the modeling bias. See Young and Hunter (2010), Huang et al. (2013) for more information. We first prove the identifiability result of model (2) in the following theorem and defer its proof to the "Appendix".

Theorem 4 *Assume that*

1. $\pi_j(z) > 0$ *are differentiable and not constant on the support of* $\boldsymbol{\alpha}^\top \boldsymbol{x}$, $j = 1, \ldots, k$;
2. *The components of* $\boldsymbol{x}$ *are continuously distributed random variables that have a joint probability density function;*
3. *The support of* $\boldsymbol{x}$ *contains an open set in* $\mathbb{R}^p$ *and is not contained in any proper linear subspace of* $\mathbb{R}^p$;
4. $\|\boldsymbol{\alpha}\| = 1$ *and the first nonzero element of* $\boldsymbol{\alpha}$ *is positive;*
5. $(\boldsymbol{\beta}_j, \sigma_j^2)$, $j = 1, \ldots, k$, *are distinct pairs.*

Then, model (2) *is identifiable.*

3.2 Estimation procedure

The log-likelihood of the collected data for model (2) is:

$$\ell^{*(2)}(\boldsymbol{\pi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^\top \boldsymbol{x}_i) \phi(Y_i | \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2) \right\}, \tag{12}$$

where $\boldsymbol{\pi}(\cdot) = \{\pi_1(\cdot), \ldots, \pi_{k-1}(\cdot)\}^\top$, $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \ldots, \sigma_k^2\}^\top$, and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\}^\top$. Since $\boldsymbol{\pi}(\cdot)$ consists of nonparametric functions, (12) is not ready for maximization. We propose a backfitting algorithm to iterate between estimating the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ and the nonparametric functions $\boldsymbol{\pi}(\cdot)$.

Given the estimates of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$, say $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2)$, then $\boldsymbol{\pi}(\cdot)$ can be estimated locally by maximizing the following local log-likelihood function:

$$\ell_1^{(2)}(\boldsymbol{\pi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i) \phi(Y_i | \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2) \right\} K_h(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z). \tag{13}$$

Let $\hat{\boldsymbol{\pi}}(\cdot)$ be the estimate that maximizes (13). We can then further update the estimate of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ by maximizing

$$\ell_2^{(2)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \hat{\pi}_j(\boldsymbol{\alpha}^\top \boldsymbol{x}_i) \phi(Y_i | \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2) \right\}. \tag{14}$$

The backfitting algorithm by iterating the above two steps can be summarized as follows.

Algorithm 3 *Backfitting algorithm to estimate the model (2).*

Step 1: *Obtain an initial estimate of* $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$.

Step 2: *Given* $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2)$*, use the following modified EM-type algorithm to maximize* $\ell_1^{(2)}$ *in (13).*

E-step: *Calculate the expectations of component labels based on estimates from* l*th iteration:*

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)}(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i) \phi(Y_i | \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)}{\sum_{j=1}^{k} \pi_j^{(l)}(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i) \phi(Y_i | \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)}, \tag{15}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, k$.

M-step: *Update the estimate*

$$\pi_j^{(l+1)}(z) = \frac{\sum_{i=1}^{n} p_{ij}^{(l+1)} K_h(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z)}{\sum_{i=1}^{n} K_h(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z)} \tag{16}$$

for $z \in \{u_t, t = 1, \ldots, N\}$. *We then update* $\pi_j^{(l+1)}(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)$, $i = 1, \ldots, n$ *by linear interpolating* $\pi_j^{(l+1)}(u_t)$, $t = 1, \ldots, N$.

Step 3: *Given* $\hat{\boldsymbol{\pi}}(\cdot)$ *from Step 2, update* $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ *by maximizing* (14). *We propose to iterate between updating* $\boldsymbol{\alpha}$ *and* $(\boldsymbol{\beta}, \boldsymbol{\sigma})$.

Step 3.1: *Given* $\hat{\boldsymbol{\alpha}}$, *update* $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$.
E-step: *Calculate the classification probabilities:*

$$p_{ij}^{(l+1)} = \frac{\hat{\pi}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(l)}, \sigma_j^{2(l)})}{\sum_{j=1}^k \hat{\pi}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(l)}, \sigma_j^{2(l)})}, \quad j = 1, \ldots, k. \tag{17}$$

M-step: *Update* $\boldsymbol{\beta}$ *and* $\boldsymbol{\sigma}^2$*:*

$$\boldsymbol{\beta}_j^{(l+1)} = (\boldsymbol{S}^\top \boldsymbol{R}_j^{(l+1)} \boldsymbol{S})^{-1} \boldsymbol{S}^\top \boldsymbol{R}_j^{(l+1)} \boldsymbol{y}, \tag{18}$$

$$\sigma_j^{2(l+1)} = \frac{\sum_{i=1}^n p_{ij}^{(l+1)}(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j^{(l+1)})^2}{\sum_{i=1}^n p_{ij}^{(l+1)}}, \tag{19}$$

where $j = 1, \ldots, k$, $\boldsymbol{R}_j^{(l+1)} = diag\{p_{ij}^{(l+1)}, \ldots, p_{nj}^{(l+1)}\}$, *and* $\boldsymbol{S} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$.

Step 3.2: *Given* $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$, *update* $\boldsymbol{\alpha}$ *by maximizing the following log-likelihood*

$$\ell_3^{(2)}(\boldsymbol{\alpha}) = \sum_{i=1}^n \log \left\{ \sum_{j=1}^k \hat{\pi}_j(\boldsymbol{\alpha}^\top \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2) \right\}.$$

Step 3.3: *Iterate Steps 3.1–3.2 until convergence.*
Step 4: *Iterate Steps 2–3 until convergence.*

There are many ways to obtain an initial estimate of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$. In our numerical studies, we get an initial estimate of $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ by fitting traditional mixtures of linear regression models. Using the resulting hard-clustering results as new response variable, we then apply the SIR to get an initial estimate for $\boldsymbol{\alpha}$.

3.3 Asymptotic properties

Let $(\hat{\boldsymbol{\pi}}(z), \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ be the resulting estimate of Algorithm 3. In this section, we investigate their asymptotic properties. Let $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, (\boldsymbol{\sigma}^2)^\top)^\top$ and $\boldsymbol{\lambda} = (\boldsymbol{\alpha}^\top, \boldsymbol{\eta}^\top)^\top$. Define

$$\ell(\boldsymbol{\pi}(z), \boldsymbol{\lambda}, \boldsymbol{x}, y) = \log \sum_{j=1}^{k} \pi_j(z)\phi\{y|\boldsymbol{x}^\top \boldsymbol{\beta}_j, \sigma_j^2\},$$
$$q_\pi(z) = \frac{\partial \ell(\boldsymbol{\pi}(z), \lambda, x, y)}{\partial \boldsymbol{\pi}},$$
$$q_{\pi\pi}(z) = \frac{\partial^2 \ell(\boldsymbol{\pi}(z), \lambda, x, y)}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}^\top}.$$

Similarly, define q_λ, $q_{\lambda\lambda}$, and $q_{\pi\eta}$. Denote $\mathscr{I}_\pi^{(2)}(z) = -E[q_{\pi\pi}(Z)|Z = z]$ and $\Lambda_2(u|z) = E[q_\pi(z)|Z = u]$.

Under some regularity conditions, the asymptotic properties of $\hat{\boldsymbol{\pi}}(z)$ are given in the following theorem and its proof is given in the "Appendix".

Theorem 5 *Assume that conditions* $(C1)$–$(C4)$ *and* $(C9)$–$(C11)$ *in the "Appendix" hold. Then, as* $n \to \infty$, $h \to 0$ *and* $nh \to \infty$, *we have*

$$\sqrt{nh}\{\hat{\boldsymbol{\pi}}(z) - \boldsymbol{\pi}(z) - \mathscr{B}_2(z) + o_p(h^2)\} \xrightarrow{D} N\{0, \nu_0 f^{-1}(z)\mathscr{I}_\pi^{(2)}(z)\}, \tag{20}$$

where

$$\mathscr{B}_2(z) = \mathscr{I}_\pi^{(2)-1}\left\{\frac{f'(z)\Lambda_2'(z|z)}{f(z)} + \frac{1}{2}\Lambda_2''(z|z)\right\}\kappa_2 h^2.$$

The asymptotic property of the parametric estimate $\hat{\boldsymbol{\lambda}}$ is given in the following theorem.

Theorem 6 *Assume that conditions* $(C1)$–$(C4)$ *and* $(C9)$–$(C12)$ *in the "Appendix" hold. Then, as* $n \to \infty$, $nh^4 \to 0$, *and* $nh^2/\log(1/h) \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \xrightarrow{D} N(0, \boldsymbol{Q}_2^{-1}),$$

where,

$$\boldsymbol{Q}_2 = E\left[q_{\pi\pi}(Z)\begin{pmatrix}\boldsymbol{x}\boldsymbol{\pi}'(Z)\\ \boldsymbol{I}\end{pmatrix}\left\{\begin{pmatrix}\boldsymbol{x}\boldsymbol{\pi}'(Z)\\ \boldsymbol{I}\end{pmatrix} - \begin{pmatrix}\mathscr{I}_\pi^{(2)-1}(Z)E\{q_{\pi\pi}(Z)(\boldsymbol{x}\boldsymbol{\pi}'(Z))^\top|Z\}\\ \mathscr{I}_\pi^{(2)-1}(Z)E\{q_{\pi\eta}(Z)|Z\}\end{pmatrix}\right\}^\top\right].$$

4 Simulation studies

In this section, we conduct simulation studies to test the performance of the proposed models and estimation procedures.

The performance of the estimates are measured via the absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE), which is $(AB^2 + SD^2)^{1/2}$. Specifically, the AB and SD of the mean functions $m_j(\cdot)$'s in the model (1) is defined as:

$$AB = \left[\frac{1}{N} \sum_{j=1}^{k} \sum_{t=1}^{N} \left\{ E\hat{m}_j(u_t) - m_j(u_t) \right\}^2 \right]^{1/2},$$

and

$$SD = \left[\frac{1}{N} \sum_{j=1}^{k} \sum_{t=1}^{N} E\left\{ \hat{m}_j(u_t) - E\hat{m}_j(u_t) \right\}^2 \right]^{1/2},$$

where $\hat{m}_j(\cdot)$ is an estimator of $m_j(\cdot)$, and $E\hat{m}_j(u_t)$ are estimated by their replicates of studies. Similarly, we can define the AB and SD for variance functions $\sigma_j^2(\cdot)$'s and proportion functions $\pi_j(\cdot)$'s. In the following numerical studies, we set $N = 100$.

In addition, a conditional bootstrap procedure is applied to estimate the standard error of estimates and construct confidence intervals for the parameters.

Example 1 Generate data from the following two-component MSIM:

$$\begin{aligned}
\pi_1(z) &= 0.5 + 0.3\sin(\pi z) \quad \text{and} \quad \pi_2(z) = 1 - \pi_1(z), \\
m_1(z) &= 6 - \sin(2\pi z/\sqrt{3}) \quad \text{and} \quad m_2(z) = \cos(\sqrt{3}\pi z) - 1, \\
\sigma_1(z) &= 0.4 + \sin(3\pi z)/15 \quad \text{and} \quad \sigma_2(z) = 0.3 + \cos(1.3\pi z)/10,
\end{aligned}$$

where $z_i = \boldsymbol{\alpha}^\top \boldsymbol{x}_i, \boldsymbol{x}_i$ is an eight-dimensional random vector with independent uniform (0,1) components, and the direction parameter is $\boldsymbol{\alpha} = (1, 1, 1, 0, 0, 0, 0, 0)^\top/\sqrt{3}$. The sample sizes $n = 200$ and $n = 400$ are conducted over 500 repetitions. To estimate $\boldsymbol{\alpha}$, we use sliced inverse regression (SIR) and the fully iterative backfitting estimate (MSIM). To estimate the nonparametric functions, we apply the one-step estimate (OS) and MSIM. For MSIM, we use both true value (T) and SIR (S) as the initial values.

First, a proper bandwidth for estimating $\boldsymbol{\pi}(\cdot)$, $\boldsymbol{m}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ is selected. Based on Theorem 2, one can calculate the theoretical optimal bandwidth by minimizing asymptotic mean squared errors. However, the theoretical optimal bandwidth depends on many unknown quantities, which are not easy to estimate in practice. In our examples, we propose to use the following cross-validation (CV) method to choose the bandwidth. Let $\mathscr{D}$ be the full data set, and divide $\mathscr{D}$ into a training set $\mathscr{R}_l$ and a test set $\mathscr{T}_l$. That is, $\mathscr{R}_l \cup \mathscr{T}_l = \mathscr{D}$ for $l = 1, \ldots, L$. We use the training set $\mathscr{R}_l$ to obtain the estimates $\{\hat{\boldsymbol{\pi}}(\cdot), \hat{\boldsymbol{m}}(\cdot), \hat{\boldsymbol{\sigma}}^2(\cdot), \hat{\boldsymbol{\alpha}}\}$, then evaluate $\boldsymbol{\pi}(\cdot)$, $\boldsymbol{m}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ for the test data set $\mathscr{T}_l$. For each $(\boldsymbol{x}_t, y_t) \in \mathscr{T}_l$, calculate the classification probability as

$$\hat{p}_{tj} = \frac{\hat{\pi}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_t)\phi(y_t|\hat{m}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_t), \hat{\sigma}_j^2(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_t))}{\sum_{j=1}^{k} \hat{\pi}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_t)\phi(y_t|\hat{m}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_t), \hat{\sigma}_j^2(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_t))}, \tag{21}$$

Table 1 Simulation results for Example 1, $n = 200$. Absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE) of global and local parameter estimates using the proposed methods

	OS			MSIM(S)			MSIM(T)		
	AB	SD	RMSE	AB	SD	RMSE	AB	SD	RMSE
α_1	0.1377	1.2297	1.2374	0.0429	0.0883	0.0982	0.0383	0.0769	0.0859
α_2	0.0186	1.1108	1.1110	0.0517	0.0905	0.1043	0.0443	0.0781	0.0898
α_3	0.0757	1.1200	1.1225	0.0568	0.0983	0.1135	0.0454	0.0913	0.1020
α_4	0.1601	1.4358	1.4447	0.1383	0.1108	0.1772	0.1201	0.1064	0.1605
α_5	0.0445	0.9646	0.9656	0.1222	0.1159	0.1684	0.1161	0.1086	0.1590
α_6	0.0579	0.8136	0.8157	0.1276	0.1164	0.1727	0.1013	0.1082	0.1482
α_7	0.0416	0.8206	0.8216	0.1133	0.1086	0.1569	0.1109	0.1007	0.1498
α_8	0.0872	0.8396	0.8442	0.1241	0.1092	0.1653	0.1272	0.1020	0.1631
π	0.1922	0.6384	0.6667	0.2030	0.1394	0.2463	0.1858	0.1266	0.2248
μ	1.1762	1.8898	2.2259	1.0708	0.5394	1.1989	1.0222	0.5199	1.1468
σ^2	0.2568	1.2091	1.2361	0.2952	0.3706	0.4738	0.2870	0.2881	0.4066

for $j = 1, \ldots, k$. The regular CV is considered, which is defined by

$$CV(h) = \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_l} (y_t - \hat{y}_t)^2,$$

where $\hat{y}_t = \sum_{j=1}^{k} \hat{p}_{tj} \hat{m}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_t)$. We also implemented the likelihood based cross validation to choose the bandwidth and the results are similar but with more computations.

Throughout the simulation, $L = 10$ and the data are randomly partitioned. The procedure is repeated 30 times, and the average of the selected bandwidth is taken as the optimal bandwidth, denoted by $\hat{h}$.

Tables 1 and 2 present the AB, SD, and RMSE for the estimates. We summarize our findings as follows. The AB, SD, and RMSE of the index parameter estimates from MSIM are consistently smaller than the OS estimates, which indicates that the MSIM is superior to the OS. This is reasonable, since the MSIM makes use of mixture information while SIR does not. Furthermore, the performance of MSIM(T) is slightly better than MSIM(S), but the improvement is negligible, indicating that the sliced inverse regression provides good initial values for our method. In addition, the analysis results about nonparametric function estimates demonstrate that MSIM works slightly better than OS. This verifies the theoretical results stated in Sect. 2.3.

In addition, to see how the selected bandwidth works, we also did simulation for two other bandwidths, $\hat{h} \times n^{-2/15}$ and $1.5\hat{h}$, which correspond to the under-smoothing and over-smoothing conditions, respectively. The results are summarized in Table 3 and 4. We can see that the proposed bandwidth selection procedure works reasonably well since the corresponding AB, SD, and RMSE are often the smallest.

Table 2 Simulation results for Example 1, $n = 400$. Absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE) of global and local parameter estimates using the proposed methods

	OS			MSIM(S)			MSIM(T)		
	AB	SD	RMSE	AB	SD	RMSE	AB	SD	RMSE
α_1	0.0908	0.9818	0.9860	0.0069	0.0347	0.0353	0.0115	0.0410	0.0426
α_2	0.1316	0.9763	0.9852	0.0182	0.0426	0.0463	0.0175	0.0405	0.0441
α_3	0.0708	0.9963	0.9988	0.0152	0.0383	0.0412	0.0142	0.0404	0.0428
α_4	0.0610	0.6902	0.6929	0.0668	0.0649	0.0931	0.0679	0.0672	0.0955
α_5	0.0251	0.5203	0.5209	0.0722	0.0569	0.0920	0.0702	0.0586	0.0914
α_6	0.0121	0.4714	0.4716	0.0706	0.0589	0.0920	0.0728	0.0663	0.0984
α_7	0.0082	0.5459	0.5460	0.0618	0.0636	0.0887	0.0633	0.0626	0.0890
α_8	0.0459	0.5097	0.5117	0.0690	0.0568	0.0894	0.0731	0.0622	0.0960
π	0.1867	0.4134	0.4536	0.1094	0.0907	0.1421	0.1119	0.0939	0.1461
μ	1.1864	1.3448	1.7933	0.6825	0.4297	0.8065	0.6931	0.4467	0.8246
σ^2	0.3219	0.8432	0.9026	0.1905	0.1698	0.2552	0.1905	0.1761	0.2594

Next, we test the accuracy of the standard error estimation and the confidence interval constructed for α via a conditional bootstrap procedure. Given the covariate x, the response Y can be generated from the estimated distribution

$$\sum_{j=1}^{k} \hat{\pi}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x})\phi(y|\hat{m}_j(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}), \hat{\sigma}_j^2(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x})).$$

For the simplicity of presentation, only the results from MSIM(S) are reported. We apply the proposed estimation procedure to each of the 200 bootstrap samples, and further obtain the confidence intervals. Table 5 summarizes the performance of the bootstrap procedure. The average and standard deviation of the 500 estimated standard errors are reported, and are denoted by SE and STD, respectively. The actual coverage probabilities based on the constructed confidence intervals are also reported. From Table 5, we can see that the actual coverage probabilities are usually very close to the nominal coverage probabilities, although are conservative for some parameters.

In the following, we explore the performance of MSIM in even higher dimensional spaces. To be more specific, $\boldsymbol{\alpha}$ is now set to be a 50-dimensional vector, whose first three elements are $1/\sqrt{3}$ and the others are zeros. $\boldsymbol{x}_i$ and the nonparametric functions are generated in the similar manner as before. That is to say, we are now exploring the cases when $n = 200$, $p = 50$ and $n = 400$, $p = 50$. The AD, SD, and RMSE for the nonparametric functions are reported in Table 6. For the global vector $\boldsymbol{\alpha}$, we reported the pooled AB, SD, and RMSE, which are defined by

$$AB_\alpha = \left[\frac{1}{p}\sum_{t=1}^{p}\left(E\hat{\alpha}_t - \alpha_t\right)^2\right]^{1/2}, \quad \text{and} \quad SD_\alpha = \left[\frac{1}{p}\sum_{t=1}^{p} E\left(\hat{\alpha}_t - E\hat{\alpha}_t\right)^2\right]^{1/2}.$$

Table 3 Absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE) of global and local parameter estimates using the proposed methods with various bandwidths, $n = 200$

	Under-smoothing			Appropriate smoothing			Over-smoothing		
	AB	SD	RMSE	AB	SD	RMSE	AB	SD	RMSE
α_1	0.0295	0.0991	0.1034	0.0429	0.0883	0.0982	0.0616	0.1032	0.1202
α_2	0.0457	0.1063	0.1157	0.0517	0.0905	0.1043	0.0781	0.1026	0.1289
α_3	0.0496	0.1107	0.1213	0.0568	0.0983	0.1135	0.0772	0.1145	0.1381
α_4	0.1180	0.1495	0.1905	0.1383	0.1108	0.1772	0.1589	0.1253	0.2024
α_5	0.1065	0.1480	0.1823	0.1222	0.1159	0.1684	0.1583	0.1353	0.2083
α_6	0.1011	0.1539	0.1841	0.1276	0.1164	0.1727	0.1533	0.1231	0.1966
α_7	0.0936	0.1665	0.1910	0.1133	0.1086	0.1569	0.1462	0.1207	0.1896
α_8	0.1139	0.1560	0.1932	0.1241	0.1092	0.1653	0.1649	0.1193	0.2035
π	0.2045	0.2382	0.3139	0.2030	0.1394	0.2463	0.2167	0.1130	0.2444
μ	1.1443	0.9375	1.4793	1.0708	0.5394	1.1989	1.1237	0.4176	1.1988
σ^2	0.2923	0.6448	0.7079	0.2952	0.3706	0.4738	0.4137	0.4523	0.6130

Table 4 Absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE) of global and local parameter estimates using the proposed methods with various bandwidths, $n = 400$

	Under-smoothing			Appropriate smoothing			Over-smoothing		
	AB	SD	RMSE	AB	SD	RMSE	AB	SD	RMSE
α_1	0.0068	0.0484	0.0489	0.0069	0.0347	0.0353	0.0161	0.0522	0.0546
α_2	0.0140	0.0453	0.0475	0.0182	0.0426	0.0463	0.0236	0.0533	0.0583
α_3	0.0143	0.0538	0.0556	0.0152	0.0383	0.0412	0.0250	0.0511	0.0569
α_4	0.0572	0.0876	0.1046	0.0668	0.0649	0.0931	0.0860	0.0779	0.1161
α_5	0.0578	0.0870	0.1044	0.0722	0.0569	0.0920	0.0878	0.0689	0.1116
α_6	0.0533	0.0867	0.1018	0.0706	0.0589	0.0920	0.0863	0.0723	0.1126
α_7	0.0371	0.0950	0.1020	0.0618	0.0636	0.0887	0.0835	0.0778	0.1142
α_8	0.0536	0.0868	0.1020	0.0690	0.0568	0.0894	0.0898	0.0730	0.1157
π	0.0942	0.1370	0.1663	0.1094	0.0907	0.1421	0.1305	0.0906	0.1588
μ	0.6046	0.6502	0.8878	0.6825	0.4297	0.8065	0.8045	0.4161	0.9057
σ^2	0.1408	0.3370	0.3653	0.1905	0.1698	0.2552	0.3020	0.1811	0.3521

The results are based on MSIM(S). The results from the previous $n = 200$, $p = 8$ and $n = 400$, $p = 8$ cases are also reported for comparison. It can be seen that when we increase p to 50, the MSIM still provides reasonable, although slightly worse, estimates.

Table 5 Standard errors and coverage probabilities for the global index parameter $\boldsymbol{\alpha}$ in Example 1

	$n = 200$		$n = 400$	
	SE(STD)	95%	SE(STD)	95%
α_1	0.1394 (0.0294)	98.50	0.0594 (0.0089)	100.00
α_2	0.1393 (0.0285)	98.00	0.0608 (0.0107)	98.50
α_3	0.1400 (0.0297)	97.50	0.0604 (0.0095)	99.50
α_4	0.1441 (0.0278)	94.00	0.0858 (0.0119)	95.00
α_5	0.1419 (0.0290)	94.50	0.0850 (0.0106)	94.00
α_6	0.1438 (0.0273)	94.00	0.0848 (0.0124)	94.00
α_7	0.1428 (0.0287)	97.50	0.0831 (0.0117)	93.00
α_8	0.1440 (0.0280)	95.00	0.0840 (0.0119)	95.50

Table 6 Example 1 in higher dimensional case

	AB	SD	RMSE		AB	SD	RMSE
$n = 200, p = 8$				$n = 400, p = 8$			
α	0.1012	0.1053	0.1460	α	0.0546	0.0532	0.0763
π	0.2030	0.1394	0.2463	π	0.1094	0.0907	0.1421
μ	1.0708	0.5394	1.1989	μ	0.6825	0.4297	0.8065
σ^2	0.2952	0.3706	0.4738	σ^2	0.1905	0.1698	0.2552
$n = 200, p = 50$				$n = 400, p = 50$			
α	0.0383	0.0951	0.1025	α	0.0164	0.0551	0.0575
π	0.2021	0.2613	0.3303	π	0.1935	0.1561	0.2486
μ	1.0437	0.7285	1.2729	μ	1.0468	0.6642	1.2398
σ^2	0.3031	0.3915	0.4951	σ^2	0.2418	0.2830	0.3723

Absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE) of global and local parameter estimates using MSIM(S)

Example 2 Next, we consider a two-component MRSIP:

$$\pi_1(z) = 0.5 - 0.35\sin(\pi z) \quad \text{and} \quad \pi_2(z) = 1 - \pi_1(z),$$
$$m_1(\boldsymbol{x}) = 3 + 3x_2 \quad \textit{and} \quad m_2(\boldsymbol{x}) = -3 + 2x_1 + 3x_3,$$
$$\sigma_1^2 = 0.6 \quad \text{and} \quad \sigma_2^2 = 0.4,$$

where $m_1(\boldsymbol{x})$ and $m_2(\boldsymbol{x})$ are the regression functions for the first and second components, respectively, and $\boldsymbol{x}_i$ is an eight-component random vector with independent uniform (0,1) components. Let $\boldsymbol{\beta}_1 = (3, 0, 3, 0, 0, 0, 0, 0, 0)^\top$, $\boldsymbol{\beta}_2 = (-3, 2, 0, 3, 0, 0, 0, 0, 0)^\top$, and $\boldsymbol{\alpha} = (1, 1, 1, 0, 0, 0, 0, 0)^\top/\sqrt{3}$. MRSIP with true value (T) and SIR (S) as initial values are used to fit the data, and the results are compared to the traditional mixture of linear regression models (MLR). The bandwidth for MRSIP is chosen based on the cross-validation, similar to Example 1.

Tables 7 and 8 report the AB, SD, and RMSE for the estimates. From both tables, we can see that MRSIP works comparable to MLR when the sample size is small,

Table 7 Simulation results for Example 2, $n = 200$

	MRSIP(S)			MRSIP(T)			MLR		
	AB	SD	RMSE	AB	SE	RMSE	AB	SD	RMSE
β_{01}	0.1420	0.5537	0.5716	0.1649	0.6426	0.6634	0.4617	1.4625	1.5337
β_{11}	0.0608	0.3399	0.3453	0.0572	0.3364	0.3412	0.1939	0.7771	0.8009
β_{21}	0.0448	0.3872	0.3898	0.0407	0.3838	0.3859	0.1011	0.9011	0.9068
β_{31}	0.0405	0.3249	0.3274	0.0642	0.5240	0.5279	0.1874	0.9366	0.9552
β_{41}	0.0189	0.3266	0.3271	0.0280	0.3818	0.3828	0.0624	1.0480	1.0498
β_{51}	0.0299	0.3556	0.3569	0.0240	0.4383	0.4390	0.0859	0.7774	0.7821
β_{61}	0.0330	0.3331	0.3348	0.0533	0.4862	0.4891	0.0316	0.8222	0.8228
β_{71}	0.0133	0.3578	0.3580	0.0181	0.4503	0.4507	0.0516	0.9714	0.9728
β_{81}	0.0357	0.3803	0.3819	0.0481	0.4027	0.4055	0.0301	0.7091	0.7098
β_{02}	0.0011	0.3669	0.3669	0.0178	0.4112	0.4116	0.2119	0.7298	0.7599
β_{12}	0.0038	0.2672	0.2672	0.0406	0.5078	0.5094	0.2117	0.7577	0.7867
β_{22}	0.0156	0.3206	0.3210	0.0269	0.3795	0.3805	0.2717	0.9662	1.0037
β_{32}	0.0251	0.2661	0.2673	0.0273	0.2820	0.2833	0.1896	0.5939	0.6234
β_{42}	0.0177	0.2609	0.2615	0.0297	0.2920	0.2935	0.0305	0.4154	0.4166
β_{52}	0.0147	0.2548	0.2553	0.0112	0.3210	0.3211	0.0362	0.3730	0.3747
β_{62}	0.0061	0.2566	0.2567	0.0041	0.3140	0.3140	0.0033	0.4068	0.4068
β_{72}	0.0050	0.2663	0.2663	0.0088	0.2750	0.2752	0.0345	0.4096	0.4111
β_{82}	0.0101	0.2614	0.2616	0.0099	0.2875	0.2877	0.0142	0.3587	0.3590
σ_1^2	0.0053	0.1115	0.1116	0.0281	0.2920	0.2934	0.1131	0.6941	0.7033
σ_2^2	0.0825	0.1132	0.1401	0.1023	0.1919	0.2175	0.1811	0.8156	0.8355
α_1	0.0035	0.1016	0.1017	0.0004	0.0048	0.0048	–	–	–
α_2	0.0556	0.1491	0.1591	0.0003	0.0047	0.0047	–	–	–
α_3	0.0452	0.1441	0.1510	0.0001	0.0044	0.0044	–	–	–
α_4	0.0019	0.0983	0.0983	0.0004	0.0059	0.0059	–	–	–
α_5	0.0130	0.1019	0.1027	0.0001	0.0057	0.0057	–	–	–
α_6	0.0050	0.1056	0.1057	0.0007	0.0056	0.0057	–	–	–
α_7	0.0054	0.0997	0.0999	0.0002	0.0057	0.0057	–	–	–
α_8	0.0032	0.1116	0.1117	0.0006	0.0058	0.0058	–	–	–
π	0.0082	0.1872	0.1874	0.0071	0.1900	0.1901	0.0516	0.9714	0.9728

Absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE) of global and local parameter estimates using the proposed methods

and outperforms MLR when the sample size is large. It is clear that the MRSIP provides better estimates of component proportions than MLR since the constant assumption of component proportions by MLR is violated. By reducing the modeling bias of component proportions, MRSIP is able to better classify observations into two components and thus provide better component regression parameters. In addition, we can see that MRSIP(S) provides similar results to MRSIP(T), which demonstrates that SIR provides good initial values for MRSIP.

Table 8 Simulation results for Example 2, $n = 400$

	MRSIP(S)			MRSIP(T)			MLR		
	AB	SD	RMSE	AB	SD	RMSE	AB	SD	RMSE
β_{01}	0.0405	0.3785	0.3806	0.0323	0.3794	0.3807	0.1018	0.8683	0.8742
β_{11}	0.0094	0.2393	0.2395	0.0084	0.2359	0.2360	0.0649	0.4331	0.4379
β_{21}	0.0157	0.2346	0.2351	0.0071	0.2327	0.2328	0.0180	0.7057	0.7060
β_{31}	0.0021	0.2428	0.2428	0.0054	0.2361	0.2362	0.0670	0.3974	0.4030
β_{41}	0.0038	0.2425	0.2425	0.0048	0.2422	0.2422	0.0188	0.4072	0.4076
β_{51}	0.0171	0.2522	0.2527	0.0162	0.2539	0.2545	0.0365	0.6270	0.6281
β_{61}	0.0093	0.2449	0.2451	0.0099	0.2448	0.2450	0.0055	0.3642	0.3643
β_{71}	0.0065	0.2271	0.2272	0.0081	0.2309	0.2311	0.0132	0.4773	0.4775
β_{81}	0.0168	0.2464	0.2470	0.0163	0.2484	0.2489	0.0262	0.3277	0.3287
β_{02}	0.0120	0.2674	0.2677	0.0198	0.2712	0.2720	0.1211	0.5956	0.6078
β_{12}	0.0082	0.1846	0.1848	0.0045	0.1872	0.1872	0.1124	0.4199	0.4347
β_{22}	0.0267	0.1888	0.1907	0.0221	0.1876	0.1889	0.1253	0.6704	0.6821
β_{32}	0.0184	0.1963	0.1972	0.0094	0.1985	0.1987	0.1097	0.3891	0.4043
β_{42}	0.0116	0.1764	0.1768	0.0108	0.1792	0.1795	0.0060	0.2454	0.2454
β_{52}	0.0163	0.1726	0.1734	0.0164	0.1777	0.1784	0.0257	0.2754	0.2766
β_{62}	0.0058	0.1876	0.1877	0.0054	0.1915	0.1916	0.0023	0.2593	0.2593
β_{72}	0.0119	0.1726	0.1730	0.0117	0.1786	0.1790	0.0090	0.2249	0.2251
β_{82}	0.0067	0.1677	0.1678	0.0085	0.1725	0.1727	0.0109	0.1937	0.1940
σ_1^2	0.0621	0.0838	0.1043	0.0606	0.0815	0.1016	0.0706	0.6667	0.6705
σ_2^2	0.1075	0.0718	0.1293	0.1068	0.0703	0.1279	0.0529	0.4638	0.4668
α_1	0.0231	0.0907	0.0936	0.0001	0.0045	0.0045	–	–	–
α_2	0.0407	0.1289	0.1352	0.0007	0.0047	0.0048	–	–	–
α_3	0.0444	0.1326	0.1398	0.0004	0.0047	0.0048	–	–	–
α_4	0.0062	0.0707	0.0710	0.0005	0.0057	0.0058	–	–	–
α_5	0.0039	0.0709	0.0710	0.0007	0.0054	0.0055	–	–	–
α_6	0.0012	0.0717	0.0717	0.0002	0.0060	0.0060	–	–	–
α_7	0.0029	0.0710	0.0711	0.0006	0.0059	0.0060	–	–	–
α_8	0.0013	0.0687	0.0687	0.0003	0.0062	0.0062	–	–	–
π	0.0076	0.1943	0.1944	0.0064	0.1996	0.1997	0.0132	0.4773	0.4775

Absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE) of global and local parameter estimates using the proposed methods

Next, the standard error estimation and the confidence interval construction for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2$ and $\boldsymbol{\alpha}$ are reported in Table 9 via the conditional bootstrap procedure introduced in Example 1. The results are based on MRSIP(S). From Table 9, we can see that the actual coverage probabilities are usually close to the nominal coverage probabilities.

We now illustrate the performance of the MRSIP model for a high dimensional setting when p is increased to 50. In this case, $\boldsymbol{\alpha} = (1, 1, 1, 0, \ldots, 0)^\top/\sqrt{3}$ is a 50 dimensional vector, $\boldsymbol{\beta}_1 = (3, 0, 3, 0, \ldots, 0)^\top$ and $\boldsymbol{\beta}_2 = (-3, 2, 0, 3, 0, \ldots, 0)^\top$ are

Table 9 Standard errors and coverage probabilities for global parameters in Example 2

	$n = 200$		$n = 400$	
	SE(STD)	95%	SE(STD)	95%
β_{01}	0.7427 (0.1616)	99.00	0.4264 (0.0716)	94.00
β_{11}	0.5339 (0.1428)	99.50	0.2825 (0.0533)	97.00
β_{21}	0.5094 (0.1126)	99.00	0.3045 (0.0567)	97.50
β_{31}	0.5296 (0.1458)	98.50	0.2863 (0.0576)	97.50
β_{41}	0.4848 (0.1206)	99.50	0.2731 (0.0541)	96.50
β_{51}	0.4898 (0.1279)	97.50	0.2745 (0.0528)	96.00
β_{61}	0.4819 (0.1277)	97.50	0.2712 (0.0525)	98.50
β_{71}	0.4883 (0.1298)	98.50	0.2846 (0.0753)	97.50
β_{81}	0.5018 (0.1516)	97.50	0.2754 (0.0462)	96.50
β_{02}	0.4832 (0.1044)	99.00	0.2805 (0.0305)	97.00
β_{12}	0.4159 (0.1447)	98.50	0.2128 (0.0360)	98.00
β_{22}	0.4350 (0.1295)	96.50	0.2108 (0.0451)	95.00
β_{32}	0.3793 (0.1040)	99.00	0.2114 (0.0304)	95.00
β_{42}	0.3550 (0.1112)	98.00	0.2016 (0.0483)	96.00
β_{52}	0.3570 (0.1059)	98.00	0.1973 (0.0211)	96.50
β_{62}	0.3487 (0.1070)	97.50	0.1966 (0.0221)	96.00
β_{72}	0.3580 (0.1089)	98.00	0.2018 (0.0453)	95.50
β_{82}	0.3599 (0.1138)	98.00	0.2037 (0.0604)	98.00
σ_1^2	0.2658 (0.1186)	99.50	0.1196 (0.0488)	97.50
σ_2^2	0.3278 (0.1359)	98.50	0.1141 (0.0764)	94.00
α_1	0.1098 (0.0441)	91.00	0.0645 (0.0230)	92.50
α_2	0.2358 (0.0922)	98.00	0.1616 (0.0529)	91.50
α_3	0.1290 (0.0763)	81.00	0.0675 (0.0406)	86.00
α_4	0.0940 (0.0117)	91.50	0.0638 (0.0059)	91.00
α_5	0.0954 (0.0138)	92.00	0.0651 (0.0066)	93.00
α_6	0.0953 (0.0136)	92.00	0.0647 (0.0067)	90.50
α_7	0.0958 (0.0146)	94.00	0.0646(0.0067)	93.50
α_8	0.0962 (0.0147)	91.00	0.0646 (0.0064)	92.50

51-dimensional vectors. $\pi_j(\cdot)$ and σ_j^2, $j = 1, 2$ are still the same as before. The results are summarized in Table 10, where the AB and SD of $\boldsymbol{\beta}$ are defined as

$$AB_\beta = \left[\frac{1}{p+1} \sum_{j=1}^{k} \sum_{t=1}^{p+1} \left(E\hat{\beta}_{jt} - \beta_{jt} \right)^2 \right]^{1/2}, \quad \text{and}$$

$$SD_\beta = \left[\frac{1}{p+1} \sum_{j=1}^{k} \sum_{t=1}^{p+1} E\left(\hat{\beta}_{jt} - E\hat{\beta}_{jt} \right)^2 \right]^{1/2}.$$

The results are based on MRSIP(S). From the table, we can see that, given the same sample size, increasing the number of predictors would downgrade the performance

Table 10 Example 2 in higher dimensional case

	AB	SD	RMSE		AB	SD	RMSE
$n = 200$, $p = 8$				$n = 400$, $p = 8$			
β	0.0743	0.5309	0.5361	β	0.0226	0.3237	0.3245
σ^2	0.0926	0.2415	0.2587	σ^2	0.1242	0.1104	0.1661
α	0.1677	0.1279	0.2109	α	0.0230	0.0918	0.0946
π	0.0094	0.1701	0.1703	π	0.0076	0.1943	0.1944
$n = 400$, $p = 50$				$n = 800$, $p = 50$			
β	0.2246	1.0463	1.0701	β	0.0254	0.2913	0.2924
σ^2	0.4320	1.0913	1.1737	σ^2	0.1115	0.3565	0.3735
α	0.0729	0.0598	0.0943	α	0.0747	0.0416	0.0855
π	0.0211	0.1725	0.1738	π	0.0098	0.1613	0.1616

Absolute bias (AB), standard deviation (SD), and root mean squared error (RMSE) of global and local parameter estimates using MRSIP(S)

of MRSIP. However, if we increase the sample size to 800, even with 50 predictors, the MRSIP still works very well.

5 Real data examples

Example 1 (NBA data) We illustrate the proposed methodology by an analysis of "The effectiveness of National Basketball Association guards". There are many ways to measure the (statistical) performance of guards in the NBA. Of interest is how the height of the player (Height), minutes per game (MPG) and free throw percentage (FTP) affect points per game (PPM) (Chatterjee et al. 1995).

The data set contains some descriptive statistics for all 105 guards for the 1992–1993 season. Since players playing very few minutes are quite different from those who play a sizable part of the season, we only look at those players playing 10 or more minutes per game and appearing in 10 or more games. In addition, Michael Jordan is an outlier, so we also omit him from our data analysis. These exclude 10 players (Chatterjee et al. 1995). We divide each variable by its corresponding standard deviation, so that they have comparable numerical scales.

To evaluate the prediction performance of the proposed models and compared them to the linear regression model (Linear), the mixture of linear regression models (MLR), the nonparametric mixture of regression models (MNP, Huang et al. 2013), the mixture of regression models with varying mixing proportions (VaryPr1, Huang and Yao 2012), and the mixtures of regressions with predictor-dependent mixing proportions (VaryPr2, Young and Hunter 2010), we used d-fold cross-validation with $d = 5$, 10, and the Monte-Carlo cross-validation (MCCV) with $d = 10$, 20 (Shao 1993). In MCCV, the data were partitioned 500 times into disjoint training subsets (with size $n - d$) and test subsets (with size d). The mean squared prediction error evaluated at the test data sets are reported as boxplots in Fig. 1b. Apparently, the MSIM has superior prediction power than the rest of the models, followed by MNP. The aver-

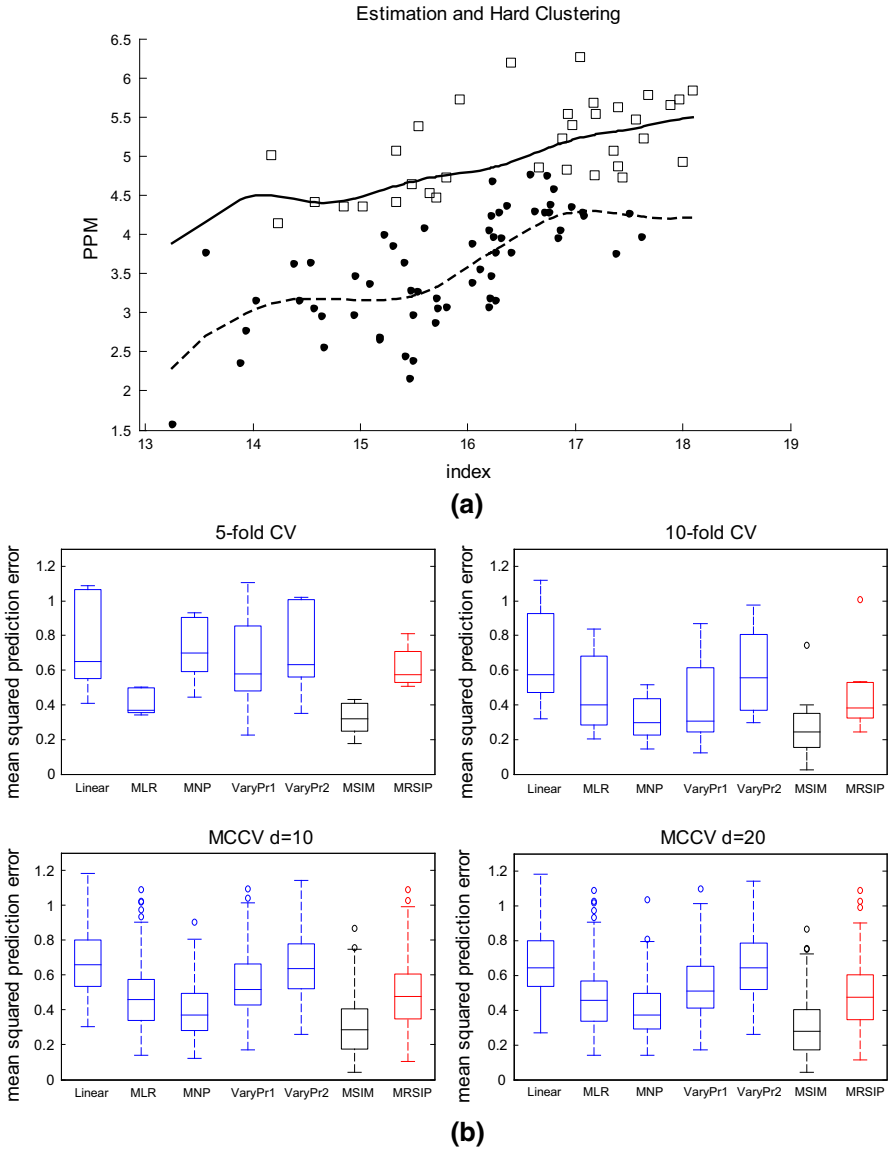

Fig. 1 NBA data: **a** estimated mean functions and a hard-clustering result; **b** mean squared prediction error: five-fold CV; ten-fold CV; MCCV d = 10; MCCV d = 20

age improving rates of MSIM over MNP, for the four cross-validation methods, are 46.64%, 15.16%, 22.87%, and 25.22%, respectively.

Next, we give detailed analysis of this dataset through the MSIM. An optimal bandwidth is selected at 0.344 by CV procedure. Figure 1a contains the estimated mean functions and hard-clustering results, denoted by dots and squares, respectively. The 95% confidence interval for $\hat{\boldsymbol{\alpha}}$ are (0.134,0.541), (0.715,0.949) and (0.202,0.679). Therefore, MPG is the most influential factor on PPM. This might be partly explained

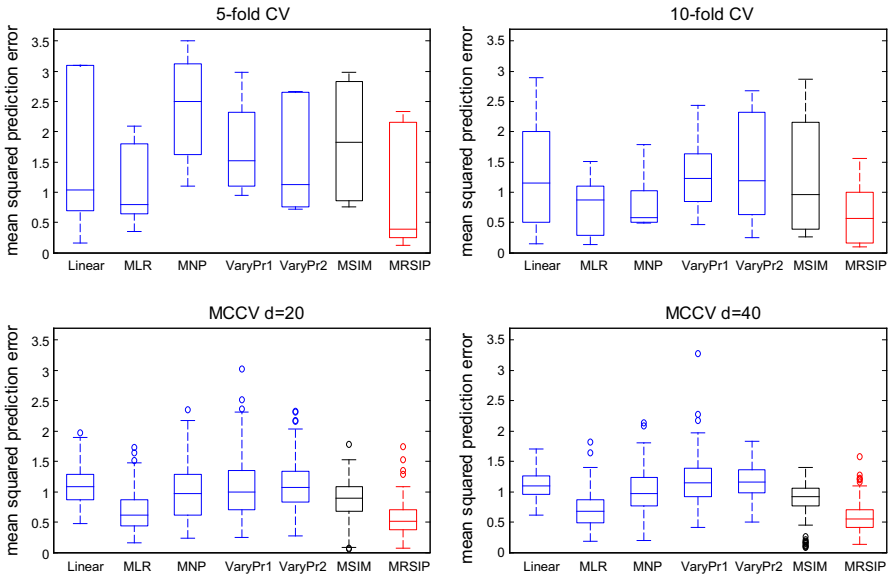

Fig. 2 Corporations core competition data: mean squared prediction error: five-fold CV; ten-fold CV; MCCV d = 20; MCCV d = 40

by that coaches tend to let good players with higher PPM play longer minutes per game (i.e., higher MPG). The two groups of guards our new models found might be explained by the difference between shooting guards and passing guards.

Example 2 (Corporations core competition data) We now analyze a corporations core competition dataset, which includes descriptive statistics and financial data of 196 manufacturing listed companies for the year of 2017. Of interest is what determines the value of a corporation. The response variable is the market value Y, and the independent variables are general assets (X_1), sales revenue (X_2), sales revenue growth rate (X_3), income per capita (X_4), earning cash flow (X_5), inventory turnover ratio (X_6), accounts receivable turnover (X_7), earning per share (X_8), return on equity (X_9), research and development expenditure (X_{10}), proportion of scientific research personnel (X_{11}), proportions of stuffs with undergraduate degrees (X_{12}), rate of production equipment updates (X_{13}), sales revenue within industry (X_{14}), and market share (X_{15}). Each variable is scaled before further analysis.

Similar to the previous example, we compare the prediction performance of the proposed models to the five existing models. Both CV and MCCV are applied to this dataset and the mean squared prediction errors are shown in Fig. 2. We can see that the MRSIP and the MLR are the best two models for this dataset. To better compare these two methods, we also compute the average improving rates of MRSIP over MLR, for the four cross-validation methods, which are 11.19%, 6.2%, 6.8%, and 27.10%, respectively. As a conclusion, MRSIP is the best choice for this dataset, followed by mixture of linear regressions.

Table 11 shows the standard errors and 95% confidence intervals for α, β_1, and β_2, assuming the MRSIP. It can be seen that X_1, X_2, and X_4 have significant effects on mixing proportions. Note that there are variables like X_3, which does not have

Table 11 Standard errors and 95% confidence intervals for global parameters of MRSIP for corporations core competition data

	SE	95% CI		SE	95% CI		SE	95% CI
α_1	0.2594	(0.0497, 1.0667)	β_{11}	0.2804	(−0.1375, 0.4233)	β_{12}	0.2458	(0.1299, 0.6215)
α_2	0.2268	(−0.8898, −0.0009)	β_{21}	0.2822	(−1.0299, −0.4655)	β_{22}	0.1597	(−0.0220, 0.2974)
α_3	0.2069	(−0.3101, 0.5008)	β_{31}	0.1598	(−0.1446, 0.1750)	β_{32}	0.1126	(−0.2032, 0.0221)
α_4	0.1893	(0.0177, 0.7597)	β_{41}	0.1325	(−0.0561, 0.2089)	β_{42}	0.1122	(−0.1365, 0.0880)
α_5	0.1867	(−0.5279, 0.2042)	β_{51}	0.1289	(−0.0455, 0.2124)	β_{52}	0.1176	(−0.1983, 0.0369)
α_6	0.2066	(−0.7443, 0.0657)	β_{61}	0.1425	(−0.1117, 0.1732)	β_{62}	0.1328	(−0.0772, 0.1885)
α_7	0.2156	(−0.2425, 0.6025)	β_{71}	0.1451	(0.0586, 0.3487)	β_{72}	0.1147	(−0.2819, −0.0525)
α_8	0.2746	(−0.3576, 0.7189)	β_{81}	0.1927	(−0.5882, −0.2027)	β_{82}	0.1280	(−0.1370, 0.1189)
α_9	0.2566	(−0.6350, 0.3707)	β_{91}	0.1712	(−0.5251, −0.1827)	β_{92}	0.1181	(−0.1018, 0.1343)
α_{10}	0.2153	(−0.4150, 0.4291)	β_{101}	0.2339	(−0.7801, −0.3124)	β_{102}	0.1440	(0.1207, 0.4086)
α_{11}	0.2080	(−0.3582, 0.4572)	β_{111}	0.1766	(−0.6027, −0.2494)	β_{112}	0.1203	(−0.0055, 0.2350)
α_{12}	0.1773	(−0.5517, 0.1433)	β_{121}	0.1625	(−0.5193, −0.1943)	β_{122}	0.1241	(−0.0164, 0.2318)
α_{13}	0.2033	(−0.4094, 0.3874)	β_{131}	0.1289	(−0.0022, 0.2555)	β_{132}	0.1257	(−0.1346, 0.1169)
α_{14}	0.2063	(−0.2437, 0.5649)	β_{141}	0.2521	(0.3225, 0.8268)	β_{142}	0.1682	(−0.5348, −0.1984)
α_{15}	0.3414	(−0.4794, 0.8588)	β_{151}	0.2563	(−0.3166, 0.1961)	β_{152}	0.2019	(−0.5091, −0.1053)

significant effect on α or β's, and therefore, variable selection methods might be helpful to further improve the accuracy and homogeneity of the model estimation. Hard clustering shows that most companies in the first component are innovative companies, who spend huge amount of money on recruiting high-end talents and developing new products, while the second component contains companies that are mostly more traditional and conservative.

6 Discussion

In this paper, we propose two finite semiparametric mixture of regression models and provide the modified EM algorithms to estimate them. We establish the identifiability results of the new models and investigate the asymptotic properties of the proposed estimation procedures. Throughout the article, we assume that the number of components is known and fixed, but it requires more research to select the number of components for the proposed semiparametric mixture models. It will be interesting to know whether the recently proposed EM test (Chen and Li 2009; Li and Chen 2010) can be extended to the proposed semiparametric mixture models. In addition, it is also interesting to build some formal model selection procedure to compare different semiparametric mixture models. In the real data applications, we use the cross-validation criteria to compare different models. When the models are nested, one might use generalized likelihood ratio statistic proposed by Fan et al. (2001) to test any parametric assumption for the semiparametric models. Furthermore, the assumption of fixed dimension of predictors can be relaxed and the proposed models can be extended to the cases where the dimension of predictors p also diverges with the sample size n. This might be done by using the idea of penalized local likelihood if the sparsity assumption is added on the predictors. We employed kernel regression method to estimate the nonparametric functions. One can also use local polynomial regression method (such as local linear) to possibly reduce the bias and boundary effects of the estimators.

Acknowledgements The authors are grateful to the editor, the guest editor, and two referees for numerous helpful comments during the preparation of the article. Funding was provided by National Natural Science Foundation of China (Grant No. 11601477), Natural Science Foundation (USA) (Grant No. DMS-1461677), Department of Energy (Grant No. 10006272), the First Class Discipline of Zhejiang - A (Zhejiang University of Finance and Economics-Statistics), China (Grant No. NA) and Natural Science Foundation of Zhejiang Province (Grant No. LY19A010006).

Appendix A

Technical conditions

(C1) The sample $\{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$ is independent and identically distributed from its population $(\boldsymbol{x}, Y)$. The support for $\boldsymbol{x}$, denoted by $\mathscr{X}$, is a compact subset of $\mathbb{R}^3$.

(C2) The marginal density of $\boldsymbol{\alpha}^\top \boldsymbol{x}$, denoted by $f(\cdot)$, is twice continuously differentiable and positive at the point z.

(C3) The kernel function $K(\cdot)$ has a bounded support, and satisfies that

$$\int K(t)dt = 1, \qquad \int tK(t)dt = 0, \qquad \int t^2K(t)dt < \infty,$$
$$\int K^2(t)dt < \infty, \qquad \int |K^3(t)|dt < \infty.$$

(C4) $h \to 0$, $nh \to 0$, and $nh^5 = O(1)$ as $n \to \infty$.

(C5) The third derivative $|\partial^3 \ell(\boldsymbol{\theta}, y)/\partial\theta_i\partial\theta_j\partial\theta_k| \le M(y)$ for all y and all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}(z)$, and $E[M(y)] < \infty$.

(C6) The unknown functions $\boldsymbol{\theta}(z)$ have continuous second derivative. For $j = 1, \ldots, k$, $\sigma_j^2(z) > 0$, and $\pi_j(z) > 0$ for all $\boldsymbol{x} \in \mathscr{X}$.

(C7) For all i and j, the following conditions hold:

$$E\left[\left|\frac{\partial \ell(\boldsymbol{\theta}(z), Y)}{\partial\theta_i}\right|^3\right] < \infty \qquad E\left[\left(\frac{\partial^2 \ell(\boldsymbol{\theta}(z), Y)}{\partial\theta_i\partial\theta_j}\right)^2\right] < \infty$$

(C8) $\boldsymbol{\theta}_0''(\cdot)$ is continuous at the point z.

(C9) The third derivative $|\partial^3 \ell(\boldsymbol{\pi}, y)/\partial\pi_i\partial\pi_j\partial\pi_k| \le M(y)$ for all y and all $\boldsymbol{\pi}$ in a neighborhood of $\boldsymbol{\pi}(z)$, and $E[M(y)] < \infty$.

(C10) The unknown functions $\boldsymbol{\pi}(z)$ have continuous second derivative. For $j = 1, \ldots, k$, $\pi_j(z) > 0$ for all $\boldsymbol{x} \in \mathscr{X}$.

(C11) For all i and j, the following conditions hold:

$$E\left[\left|\frac{\partial \ell(\boldsymbol{\pi}(z), Y)}{\partial\pi_i}\right|^3\right] < \infty \qquad E\left[\left(\frac{\partial^2 \ell(\boldsymbol{\pi}(z), Y)}{\partial\pi_i\partial\pi_j}\right)^2\right] < \infty$$

(C11) $\boldsymbol{\pi}''(\cdot)$ is continuous at the point z.

Proof of Theorem 1 Ichimura (1993) have shown that under conditions (i)–(iv), $\boldsymbol{\alpha}$ is identifiable. Further, Huang et al. (2013) showed that with condition (v), the nonparametric functions are identifiable. Thus completes the proof. □

Proof of Theorem 2 Let

$$\hat{\pi}_j^* = \sqrt{nh}\{\hat{\pi}_j - \pi_j(z)\}, \qquad j = 1, \ldots, k-1.$$
$$\hat{m}_j^* = \sqrt{nh}\{\hat{m}_j - m_j(z)\}, \qquad j = 1, \ldots, k,$$
$$\hat{\sigma}_j^{2*} = \sqrt{nh}\{\hat{\sigma}_j^2 - \sigma_j^2(z)\}, \qquad j = 1, \ldots, k.$$

Define $\hat{\boldsymbol{\pi}}^* = (\hat{\pi}_1^*, \ldots, \hat{\pi}_{k-1}^*)^\top$, $\hat{\boldsymbol{m}}^* = (\hat{m}_1^*, \ldots, \hat{m}_k^*)^\top$, $\hat{\boldsymbol{\sigma}}^* = (\hat{\sigma}_1^*, \ldots, \hat{\sigma}_k^*)^\top$ and denote $\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\pi}}^{*T}, \hat{\boldsymbol{m}}^{*T}, (\hat{\boldsymbol{\sigma}}^{*2})^\top)^\top$. Let $a_n = (nh)^{-1/2}$ and

$$\ell(\boldsymbol{\theta}(z), \tilde{\boldsymbol{\alpha}}, \boldsymbol{x}_i, Y_i) = \log\left\{\sum_{j=1}^{k} \pi_j(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i)\phi(Y_i|m_j(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i), \sigma_j^2(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i))\right\} K_h(\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i - z).$$

If $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{\sigma}}^2)^\top$ maximizes (4), then $\hat{\boldsymbol{\theta}}^*$ maximizes

$$\ell_n^*(\boldsymbol{\theta}^*) = h \sum_{i=1}^{n} [\ell(\boldsymbol{\theta}(z) + a_n \boldsymbol{\theta}^*, \tilde{\boldsymbol{\alpha}}, \boldsymbol{x}_i, Y_i) - \ell(\boldsymbol{\theta}(z), \tilde{\boldsymbol{\alpha}}, \boldsymbol{x}_i, Y_i)] K_h(\hat{Z}_i - z) \tag{22}$$

with respect to $\boldsymbol{\theta}^*$. By a Taylor expansion,

$$\ell_n^*(\boldsymbol{\theta}^*) = \boldsymbol{W}_{1n}^\top \boldsymbol{\theta}^* + \frac{1}{2} \boldsymbol{\theta}^{*T} \boldsymbol{A}_{1n} \boldsymbol{\theta}^* + o_p(1), \tag{23}$$

where

$$\boldsymbol{W}_{1n} = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}(z), \tilde{\boldsymbol{\alpha}}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\theta}} K_h(\hat{Z}_i - z),$$

and

$$\boldsymbol{A}_{2n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{\theta}(z), \tilde{\boldsymbol{\alpha}}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} K_h(\hat{Z}_i - z).$$

By WLLN, it can be shown that $\boldsymbol{A}_{1n} = -f(z) \mathscr{I}_\theta^{(1)}(z) + o_p(1)$. Therefore,

$$\ell_n^*(\boldsymbol{\theta}^*) = \boldsymbol{W}_{1n}^\top \boldsymbol{\theta}^* - \frac{1}{2} f(z) \boldsymbol{\theta}^{*T} \mathscr{I}_\theta^{(1)}(z) \boldsymbol{\theta}^* + o_p(1). \tag{24}$$

Using the quadratic approximation lemma (see, for example, Fan and Gijbels 1996), we have that

$$\hat{\boldsymbol{\theta}}^* = f(z)^{-1} \mathscr{I}_\theta^{(1)}(z)^{-1} \boldsymbol{W}_{1n} + o_p(1). \tag{25}$$

Note that

$$\boldsymbol{W}_{1n} = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}(z), \boldsymbol{\alpha}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\theta}} K_h(Z_i - z) + D_{1n} + O_p\left(\sqrt{\frac{h}{n}} \|\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|^2\right)$$

where

$$D_{1n} = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \left\{ \frac{\partial^2 \ell(\boldsymbol{\theta}(z), \boldsymbol{\alpha}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} [\boldsymbol{x}_i \boldsymbol{\theta}'(Z_i)]^\top K_h(Z_i - z) \right\} (\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}).$$

Since $\sqrt{n}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = O_p(1)$, it can be shown that

$$D_{1n} = -\sqrt{h} f(z) E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}(z), \boldsymbol{\alpha}, \boldsymbol{x}, Y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} [\boldsymbol{x} \boldsymbol{\theta}'(Z)]^\top \right] = o_p(1),$$

and

$$O_p\left(\sqrt{\frac{h}{n}} \|\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|^2\right) = o_p(1).$$

Therefore,

$$\boldsymbol{W}_{1n} = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\theta}} K_h(Z_i - z) + o_p(1).$$

To complete the proof, we now calculate the mean and variance of $\boldsymbol{W}_n$. Note that

$$\begin{aligned} E(\boldsymbol{W}_{1n}) &= \sqrt{nh} E\left[E\left[\frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\theta}} K_h(Z_i - z) | Z = z_0 \right] \right] \\ &= \sqrt{nh} \left[\frac{1}{2} f(z) \Lambda_1^{''}(z|z) + f'(z) \Lambda_1^{'}(z|z) \right] \kappa_2 h^2. \end{aligned} \tag{26}$$

Similarly, we can show that

$$\mathrm{Cov}(\boldsymbol{W}_{1n}) = f(z) \mathscr{I}_{\theta}^{(1)}(z) \nu_0 + o_p(1),$$

where $\kappa_l = \int t^l K(t) dt$ and $\nu_l = \int t^l K^2(t) dt$. The rest of the proof follows a standard argument. □

Proof of Theorem 3 Denote $Z = \boldsymbol{\alpha}^\top \boldsymbol{x}$ and $\hat{Z} = \hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}$. Let $\ell(\boldsymbol{\theta}(z), X, Y) = \log \sum_{j=1}^{k} \pi_j(z) \phi(Y | m_j(z), \sigma_j^2(z))$. If $\hat{\boldsymbol{\theta}}(z_0; \hat{\boldsymbol{\alpha}})$ maximizes (4), then it solves

$$\boldsymbol{0} = n^{-1} \sum_{i=1}^{n} \frac{\partial \ell(\hat{\boldsymbol{\theta}}(z_0; \hat{\boldsymbol{\alpha}}), X_i, Y_i)}{\partial \boldsymbol{\theta}} K_h(\hat{Z}_i - z_0).$$

Apply a Taylor expansion and use the conditions on h, we obtain

$$\begin{aligned} \boldsymbol{0} &= n^{-1} \sum_{i=1}^{n} q_{1i}(Z_i) K_h(Z_i - z_0) \\ &\quad + n^{-1} \sum_{i=1}^{n} [q_{2i}(Z_i) K_h(Z_i - z_0)] \, (\, hat\boldsymbol{\theta}(z_0; \hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(z_0)) \\ &\quad + n^{-1} \sum_{i=1}^{n} q_{2i}(Z_i) [\boldsymbol{x}_i \boldsymbol{\theta}'(Z_i)]^\top K_h(Z_i - z_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + o_p(n^{-1/2}) + O_p(h^2) \end{aligned}$$

By similar argument as in the previous proof,

$$\begin{aligned} \hat{\boldsymbol{\theta}}(z_0; \hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(z_0) &= n^{-1} f^{-1}(z_0) \mathscr{I}_{\theta}^{(1)-1}(z_0) \sum_{i=1}^{n} q_{1i}(Z_i) K_h(Z_i - z_0) \\ &\quad - \mathscr{I}_{\theta}^{(1)-1}(z_0) E\{q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^\top | Z = z_0\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + o_p(n^{-1/2}). \end{aligned} \tag{27}$$

Note that

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i) &= \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i)\\ &= (\hat{\boldsymbol{\theta}}'(\boldsymbol{\alpha}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}))^\top(\hat{\boldsymbol{\alpha}}^\top - \boldsymbol{\alpha}^\top)\boldsymbol{x}_i + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}_0^\top \boldsymbol{x}_i) + o_p(n^{-1/2})\\ &= (\boldsymbol{\theta}'(\boldsymbol{\alpha}^\top \boldsymbol{x}_i))^\top(\hat{\boldsymbol{\alpha}}^\top - \boldsymbol{\alpha}^\top)\boldsymbol{x}_i + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i) + o_p(n^{-1/2}), \qquad (28)\end{aligned}$$

where the second part is handled by (27).

Since $\hat{\boldsymbol{\alpha}}$ maximizes (9), it is the solution to

$$\boldsymbol{0} = \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \hat{\boldsymbol{\theta}}'(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) \frac{\partial \ell(\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}), X_i, Y_i)}{\partial \boldsymbol{\theta}},$$

where λ is the Lagrange multiplier. By the Taylor expansion and using (28), we have that

$$\begin{aligned}\boldsymbol{0} &= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{1i}(Z_i)\\ &\quad + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{2i}(Z_i)[\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x}_i) - \boldsymbol{\theta}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i)] + o_p(1)\\ &= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{1i}(Z_i)\\ &\quad + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{2i}(Z_i)(\boldsymbol{x}_i \boldsymbol{\theta}'(Z_i))^\top(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\\ &\quad + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{2i}(Z_i)[\hat{\boldsymbol{\theta}}(Z_i) - \boldsymbol{\theta}(Z_i)]) + o_p(1).\end{aligned}$$

Define

$$A_\alpha = E\{[\boldsymbol{x}\boldsymbol{\theta}'(Z)] q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^\top\},$$

and apply (27),

$$\begin{aligned}\boldsymbol{0} &= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{1i}(Z_i) + n^{1/2} A_\beta(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\\ &\quad - n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{2i}(Z_i) \mathscr{I}_\theta^{-1}(Z_i) E\{q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^\top | Z = Z_i\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\\ &\quad + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{2i}(Z_i) n^{-1} f^{-1}(Z_i) \mathscr{I}_\theta^{-1}(Z_i)\end{aligned}$$

$$\times \sum_{t=1}^{n} q_{1t}(Z_t) K_h(Z_t - Z_i) + o_p(1)$$
$$= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{1i}(Z_i) + \boldsymbol{Q}_1 n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$$
$$+ n^{-1/2} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) q_{2i}(Z_i) n^{-1} f^{-1}(Z_i) \mathscr{I}_{\theta}^{(1)-1}(Z_i)$$
$$\times \sum_{t=1}^{n} q_{1t}(Z_t) K_h(Z_t - Z_i) + o_p(1). \tag{29}$$

Interchanging the summations in the last term, we get

$$n^{-1/2} \sum_{i=1}^{n} \left[n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{\theta}'(Z_t) q_{2t}(Z_t) K_h(Z_t - Z_i) f^{-1}(Z_t) \mathscr{I}_{\theta}^{-1}(Z_t) q_{1i}(Z_i) \right]$$
$$= n^{-1/2} \sum_{i=1}^{n} E[\boldsymbol{x}\boldsymbol{\theta}'(Z) q_2(Z) | Z_i] \mathscr{I}_{\theta}^{(1)-1}(Z_i) q_{1i}(Z_i) + o_p(1). \tag{30}$$

Let $\Gamma_\alpha = I - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top + o_p(1)$. Combining (29) and (30), and multiply by Γ_α, we have

$$\Gamma_\alpha \boldsymbol{Q}_1 n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = n^{-1/2} \sum_{i=1}^{n} \Gamma_\alpha \{\boldsymbol{x}_i \boldsymbol{\theta}'(Z_i)$$
$$+ E[\boldsymbol{x}\boldsymbol{\theta}'(Z) q_2(Z) | Z_i] \mathscr{I}_{\theta}^{(1)-1}(Z_i)\} q_{1i}(Z_i) + o_p(1). \tag{31}$$

It can be shown that the right-hand side of (31) has the covariance matrix $\Gamma_\alpha \boldsymbol{Q}_1 \Gamma_\alpha$, and therefore, completes the proof. □

Proof of Theorem 4 Ichimura (1993) have shown that under conditions (i)–(iv), $\boldsymbol{\alpha}$ is identifiable. Furthermore, Huang and Yao (2012) showed that with condition (v), $(\boldsymbol{\pi}(\cdot), \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ are identifiable. Thus completes the proof. □

Proof of Theorem 5 This proof is similar to the proof of Theorem 2.

Let $\hat{\pi}_j^* = \sqrt{nh}\{\hat{\pi}_j - \pi_j(z)\}$, $j = 1, \ldots, k-1$, and $\hat{\boldsymbol{\pi}}^* = (\hat{\pi}_1^*, \ldots, \hat{\pi}_{k-1}^*)^\top$. It can be shown that

$$\hat{\boldsymbol{\pi}}^* = f(z)^{-1} \mathscr{I}_{\pi}^{(2)-1}(z) \boldsymbol{W}_{2n} + o_p(1),$$

where

$$\boldsymbol{W}_{2n} = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\pi}(z), \hat{\boldsymbol{\lambda}}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\pi}} K_h(\hat{Z}_i - z).$$

To complete the proof, notice that

$$E(\boldsymbol{W}_{2n}) = \sqrt{nh}E\left\{E[\frac{\partial \ell(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\pi}} K_h(Z_i - z)|Z = z_0]\right\}$$
$$= \sqrt{nh}[\frac{1}{2} f(z)\Lambda_2''(z|z) + f'(z)\Lambda_2'(z|z)]\kappa_2 h^2,$$

and $\mathrm{Cov}(\boldsymbol{W}_{2n}) = f(z)\mathscr{I}_\pi^{(2)}(z)\nu_0 + o_p(1)$. The rest of the proof follows a standard argument. □

Proof of Theorem 6s The proof is similar to the proof of Theorem 3. It can be shown that

$$\hat{\boldsymbol{\pi}}(z_0; \hat{\boldsymbol{\lambda}}) - \boldsymbol{\pi}(z_0) = n^{-1} f^{-1}(z_0)\mathscr{I}_\pi^{(2)-1}(z_0) \sum_{i=1}^{n} q_{\pi i}(Z_i) K_h(Z_i - z_0)$$
$$-\mathscr{I}_\pi^{(2)-1}(z_0)E\{q_{\pi\pi}(Z)[\boldsymbol{x}\boldsymbol{\pi}'(Z)]^\top | Z = z_0\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - \mathscr{I}_\pi^{(2)-1}(z_0)E\{q_{\pi\eta}(Z)|Z$$
$$= z_0\}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + o_p(n^{-1/2}),$$

and therefore,

$$\hat{\boldsymbol{\pi}}(\hat{Z}_i; \hat{\boldsymbol{\lambda}}) - \boldsymbol{\pi}(Z_i) = \{\boldsymbol{x}_i \boldsymbol{\pi}'(Z_i)\}^\top (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \hat{\boldsymbol{\pi}}(Z_i; \hat{\boldsymbol{\lambda}})$$
$$-\boldsymbol{\pi}(Z_i) + o_p(n^{-\frac{1}{2}}). \tag{32}$$

Since $\hat{\boldsymbol{\lambda}}$ maximizes (14), it is the solution to

$$\boldsymbol{0} = \gamma \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \boldsymbol{0} \end{pmatrix} + n^{-\frac{1}{2}} \sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{x}_i \hat{\boldsymbol{\pi}}'(\hat{Z}_i; \hat{\boldsymbol{\lambda}}) \\ \mathbf{I} \end{pmatrix} q_\pi(\hat{\boldsymbol{\pi}}(\hat{Z}_i; \hat{\boldsymbol{\lambda}}), \hat{\boldsymbol{\lambda}}),$$

where γ is the Lagrange multiplier. By Taylor series and (32)

$$\boldsymbol{0} = \gamma \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \boldsymbol{0} \end{pmatrix} + n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{1i} q_{\pi i}(Z_i) + n^{\frac{1}{2}} \boldsymbol{Q}_2 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix}$$
$$+ n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{1i} q_{\pi\pi i}(Z_i) n^{-1} f^{-1}(Z_i)\mathscr{I}_\pi^{(2)-1}(Z_i)$$
$$\times \sum_{j=1}^{n} q_{\pi j}(Z_j) K_h(Z_j - Z_i) + o_p(1)$$
$$= \gamma \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \boldsymbol{0} \end{pmatrix} + n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{1i} q_{\pi i}(Z_i) + n^{\frac{1}{2}} \boldsymbol{Q}_2 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix}$$
$$+ n^{-\frac{1}{2}} \sum_{i=1}^{n} E[\boldsymbol{\Lambda}_{1i} q_{\pi\pi}(Z_i)]\mathscr{I}_\pi^{(2)-1}(Z_i) q_{\pi i}(Z_i) + o_p(1), \tag{33}$$

where $\boldsymbol{\Lambda}_{1i} = \begin{pmatrix} \boldsymbol{x}_i \boldsymbol{\pi}'(Z_i) \\ \mathbf{I} \end{pmatrix}$, and the last equation is the result of interchanging the summations. Let $\Gamma_\alpha = \begin{pmatrix} \boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} + o_p(1)$. By (33), and multiply by Γ_α, we have

$$n^{\frac{1}{2}} \Gamma_\alpha \boldsymbol{Q}_2 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \Gamma_\alpha \left\{ \boldsymbol{\Lambda}_{1i} - \mathscr{I}_\pi^{(2)-1}(Z_i) E[\boldsymbol{\Lambda}_{1i}(Z_i) q_{\pi\pi}(Z_i) | Z_i] \right\} \times q_{\pi i}(Z_i) + o_p(1). \tag{34}$$

It can be shown that the right-hand side of (34) has the covariance matrix $\Gamma_\alpha \boldsymbol{Q}_2 \Gamma_\alpha$, and thus, completes the proof. □

References

Cao J, Yao W (2012) Semiparametric mixture of binomial regression with a degenerate component. Statistica Sinica 22:27–46

Chatterjee S, Handcock MS, Simmonoff JS (1995) A casebook for a first course in statistics and data analysis. Wiley, New York

Chen J, Li P (2009) Hypothesis test for normal mixture models: the EM approach. Ann Stat 37:2523–2542

Cook RD, Li B (2002) Dimension reduction for conditional mean in regression. Ann Stat 30:455–474

Fan J, Zhang C, Zhang J (2001) Generalized likelihood ratio statistics and Wilks phenomenon. Ann Stat 29:153–193

Frühwirth-Schnatter S (2001) Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. J Am Stat Assoc 96:194–209

Green PJ, Richardson S (2002) Hidden markov models and disease mapping. J Am Stat Assoc 97:1055–1070

Härdle W, Hall P, Ichimura H (1993) Optimal smoothing in single-index models. Ann Stat 21:157–178

Henning C (2000) Identifiability of models for clusterwise linear regression. J Classif 17:273–296

Hu H, Yao W, Wu Y (2017) The robust EM-type algorithms for log-concave mixtures of regression models. Comput Stat Data Anal 111:14–26

Huang M, Yao W (2012) Mixture of regression models with varying mixing proportions: a semiparametric approach. J Am Stat Assoc 107:711–724

Huang M, Li R, Wang S (2013) Nonparametric mixture of regression models. J Am Stat Assoc 108:929–941

Huang M, Li R, Wang H, Yao W (2014) Estimating mixture of Gaussian processes by kernel smoothing. J Bus Econ Stat 32:259–270

Ichimura H (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J Econom 58:71–120

Jordan MI, Jacobs RA (1994) Hierarchical mixtures of experts and the EM algorithm. Neural Comput 6:181–214

Li K (1991) Sliced inverse regression for dimension reduction. J Am Stat Assoc 86(414):316–327

Li P, Chen J (2010) Testing the order of a finite mixture. J Am Stat Assoc 105:1084–1092

Li B, Zha H, Chiaromonte F (2005) Contour regression: a general approach to dimension reduction. Ann Stat 33:1580–1616

Luo R, Wang H, Tsai CL (2009) Contour projected dimension reduction. Ann Stat 37:3743–3778

Ma Y, Zhu L (2012) A semiparametric approach to dimension reduction. J Am Stat Assoc 107(497):168–179

Ma Y, Zhu L (2013) Efficient estimation in sufficient dimension reduction. Ann Stat 41:250–268

Shao J (1993) Linear models selection by cross-validation. J Am Stat Assoc 88:486–494

Stephens M (2000) Dealing with label switching in mixture models. J R Stat Soc B 62:795–809

Titterington D, Smith A, Makov U (1985) Statistical analysis of finite mixture distribution. Wiley, New York

Wang H, Xia Y (2008) Sliced regression for dimension reduction. J Am Stat Assoc 103:811–821

Wang Q, Yao W (2012) An adaptive estimation of MAVE. J Multivar Anal 104:88–100

Wang S, Yao W, Huang M (2014) A note on the identiability of nonparametric and semiparametric mixtures of GLMs. Stat Probab Lett 93:41–45
Wedel M, DeSarbo WS (1993) A latent class binomial logit methodology for the analysis of paired comparison data. Decis Sci 24:1157–1170
Xiang S, Yao W (2018) Semiparametric mixtures of nonparametric regressions. Ann Inst Stat Math 70:131–154
Xiang S, Yao W, Yang G (2019) An overview of semiparametric extensions of finite mixture models. Stat Sci 34:391–404
Yao W, Lindsay BG (2009) Bayesian mixture labeling by highest posterior density. J Am Stat Assoc 104:758–767
Yao W, Nandy D, Lindsay B, Chiaromonte F (2019) Covariate information matrix for sufficient dimension reduction. J Am Stat Assoc 114:1752–1764
Young DS, Hunter DR (2010) Mixtures of regressions with predictors dependent mixing proportions. Comput Stat Data Anal 54:2253–2266
Zeng P (2012) Finite mixture of heteroscedastic single-index models. Open J Stat 2:12–20