Mixture of Regression Models with Single-index

Sijia Xiang [*], and Weixin Yao [†]

Abstract

In this article, we propose a class of semiparametric mixture regression models with single-index. We argue that many recently proposed semiparametric/nonparametric mixture regression models can be considered as special cases of the proposed model. However, unlike existing semiparametric mixture regression models, the new proposed model can easily incorporate multivariate predictors into the nonparametric components. Backfitting estimates and the corresponding algorithms have been proposed to achieve the optimal convergence rate for both the parameters and the nonparametric functions. We show that nonparametric functions can be estimated with the same asymptotic accuracy as if the parameters were known and the index parameters can be estimated with the traditional parametric root n convergence rate. Simulation studies and an application of NBA data have been conducted to demonstrate the finite sample performance of the proposed models.

Key words: EM algorithm, Kernel regression, Mixture regression model, Single-index models.

[*]Corresponding Author, School of Mathematics and Statistics, Zhejiang University of Finance and Economics. E-mail: sjxiang@zufe.edu.cn. Xiang's research is supported by Zhejiang Provincial NSF of China grant LQ16A010002.

[†]Department of Statistics, University of California, Riverside. E-mail: weixin.yao@ucr.edu. Yao's research is supported by NSF grant DMS-1461677.

1 Introduction

Mixtures of regression models are commonly used to reveal the relationship among interested variables if the whole population is not homogeneous and consists of several homogeneous subgroups. It has been widely used in many areas such as econometrics, biology, and epidemiology. Recently, many semiparametric mixture models have been proposed. See, for example, Young and Hunter (2010), Huang and Yao (2012), Huang et al. (2013), Cao and Yao (2012), Xiang and Yao (2015), among others.

In this article, we apply the idea of single-index model to mixture of regression models, and propose a mixture of single-index models (MSIM) and a mixture of regression models with varying single-index proportions (MRSIP). Huang et al. (2013) proposed the nonparametric mixture of regression models $Y|_{X=x} \sim \sum_{j=1}^{k} \pi_j(x)\phi(Y_i|m_j(x), \sigma_j^2(x))$, and developed an estimation procedure by employing kernel regression. However, the above model is not very applicable to multivariate predictors due to the so called "curse of dimensionality". The proposed mixture of single-index models can naturally incorporate the multivariate predictors and relax the traditional parametric assumption of mixture of regression models.

In some cases, we might want to assume linearity in the mean functions. Therefore, the proposed MRSIP keeps the easy interpretation of the linear component regression functions while assuming that the mixing proportions are smooth functions of an index $\boldsymbol{\alpha}^T\boldsymbol{x}$.

We show the identifiability of each model under some regularity conditions. To achieve the optimal convergence rate for the global parameters and nonparametric functions, we propose backfitting estimates using the kernel regression technique. We have shown that the nonparametric functions can be estimated with the same rate as if the parameters were known, and the parameters can be estimated with the same rate of convergence, $n^{-1/2}$, that is achieved in a parametric model. Numerical studies are used to

demonstrate the effectiveness of the proposed new models, and we discuss the selection of the two models in the real data analysis.

The rest of the paper is organized as follows. In Section 2, we introduce the MSIM and study its identifiability result. A one-step estimate and a fully-iterated backfitting estimate have been proposed, and their asymptotic properties are studied. Section 3 discusses the MRSIP and its identifiability. A fully-iterated estimate and its asymptotic properties are also studied. In Section 4, we use Monte Carlo studies and a real data example to demonstrate the finite sample performance of the proposed estimates. A discussion section ends the paper.

2 Mixture of Single-index Models (MSIM)

2.1 Model Definition and Identifiability

Assume that $\{(\boldsymbol{x}_i, Y_i), i = 1, ..., n\}$ is a random sample from population $(\boldsymbol{x}, Y)$. Throughout this article, we assume that $\boldsymbol{x}$ is p-dimensional and Y is univariate. Let $\mathcal{C}$ be a latent variable, and we assume that conditional on $\boldsymbol{x}$, $\mathcal{C}$ has a discrete distribution $P(\mathcal{C} = j|\boldsymbol{x}) = \pi_j(\boldsymbol{\alpha}^T\boldsymbol{x})$ for $j = 1, ..., k$. Conditional on $\mathcal{C} = j$ and $\boldsymbol{x}$, Y follows a normal distribution with mean $m_j(\boldsymbol{\alpha}^T\boldsymbol{x})$ and variance $\sigma_j^2(\boldsymbol{\alpha}^T\boldsymbol{x})$. We assume that $\pi_j(\cdot)$, $m_j(\cdot)$, and $\sigma_j^2(\cdot)$ are unknown but smooth functions, and therefore, without observing $\mathcal{C}$, the conditional distribution of Y given $\boldsymbol{x}$ can be written as:

$$Y|\boldsymbol{x} \sim \sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^T\boldsymbol{x})\phi(Y_i|m_j(\boldsymbol{\alpha}^T\boldsymbol{x}), \sigma_j^2(\boldsymbol{\alpha}^T\boldsymbol{x})), \tag{2.1}$$

where $\phi(y|\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. Throughout the paper, we assume that k is fixed, and refer to model (2.1) as a finite semiparametric mixture of regression models, since $\pi_j(\cdot)$, $m_j(\cdot)$ and $\sigma_j^2(\cdot)$ are all nonparametric. When

$k = 1$ and $\pi_j(\cdot)$ and $\sigma_j^2(\cdot)$ are constant, model (2.1) reduces to a single index model (Ichimura, 1993; Härdle et al., 1993). If $\pi_j(\cdot)$ and $\sigma_j^2(\cdot)$ are constant, and $m_j(\cdot)$ are identity functions, then model (2.1) reduces to a finite mixture of linear regression models (Goldfeld and Quandt, 1973). If $\boldsymbol{x}$ is a scalar, then model (2.1) reduces to the nonparametric mixture of regression model proposed by Huang et al. (2013). Therefore, the proposed model (2.1) is a natural generalization of many existing popular models.

Compared to Huang et al. (2013), the appeal of the proposed MSIM is that by focusing on an index $\boldsymbol{\alpha}^T\boldsymbol{x}$, the so-called "curse of dimensionality" in fitting multivariate nonparametric regression functions is avoided. It is of dimension-reduction structure in the sense that, if we can estimate the index $\boldsymbol{\alpha}$ efficiently, then we can use the univariate $\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}$ as the covariate and simplify the model (2.1) to the nonparametric mixture regression model proposed by Huang et al. (2013), and thus avoid the curse of dimensionality when nonparametric smoothing is employed. Therefore, model (2.1) is a reasonable compromise between fully parametric and fully nonparametric modeling.

Identifiability is a major concern for most mixture models. Some well known results for identifiability 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). The following theorem gives the result on identifiability of model (2.1) and its proof is given in Section 7.

Theorem 2.1. *Assume that (i) $\pi_j(z)$, $m_j(z)$, and $\sigma_j^2(z)$ are differentiable and not constant on the support of $\boldsymbol{\alpha}^T\boldsymbol{x}$, $j = 1, ..., k$; (ii) The component of $\boldsymbol{x}$ are continuously distributed random variables that have a joint probability density function; (iii) The support of $\boldsymbol{x}$ is not contained in any proper linear subspace of $\mathbb{R}^p$; (iv) $\|\boldsymbol{\alpha}\| = 1$ and the first nonzero element of $\boldsymbol{\alpha}$ is positive; (v) Any two curves $(m_i(z), \sigma_i^2(z))$ and $(m_j(z), \sigma_j^2(z))$, $i \neq j$, are transversal. Then, model (2.1) is identifiable.*

The transversality of two smooth curves (Huang et al., 2013) implies that the mean and variance functions of any two components cannot be tangent to each other.

2.2 Estimation Procedure and Asymptotic Properties

In this subsection, we propose a one-step estimate and a fully iterative backfitting estimate to achieve the optimal convergence rate for both the index parameter and nonparametric functions.

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

$$\ell^{*(1)}(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\phi(Y_i | m_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i), \sigma_j^2(\boldsymbol{\alpha}^T \boldsymbol{x}_i))\}, \tag{2.2}$$

where $\boldsymbol{\pi}(\cdot) = \{\pi_1(\cdot), ..., \pi_{k-1}(\cdot)\}^T$, $\boldsymbol{m}(\cdot) = \{m_1(\cdot), ..., m_k(\cdot)\}^T$, and $\boldsymbol{\sigma}^2(\cdot) = \{\sigma_1^2(\cdot), ..., \sigma_k^2(\cdot)\}^T$. Since $\boldsymbol{\pi}(\cdot)$, $\boldsymbol{m}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ consist of nonparametric functions, (2.2) is not ready for maximization.

If $\hat{\boldsymbol{\alpha}}$ is an estimate of $\boldsymbol{\alpha}$, then $\boldsymbol{\pi}(\cdot)$, $\boldsymbol{m}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ can be estimated locally by maximizing the following local log-likelihood function:

$$\ell_1^{(1)}(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i | m_j(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i), \sigma_j^2(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i))\} K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z), \tag{2.3}$$

where $K_h(z) = \frac{1}{h} K(\frac{z}{h})$ and $K(\cdot)$ is a kernel density function.

Let $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$ and $\hat{\boldsymbol{\sigma}}^2(\cdot)$ be the result of maximizing (2.3). We can then further update the estimate of $\boldsymbol{\alpha}$ by maximizing

$$\ell_2^{(1)}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \hat{\pi}_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\phi(Y_i | \hat{m}_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i), \hat{\sigma}_j^2(\boldsymbol{\alpha}^T \boldsymbol{x}_i))\}, \tag{2.4}$$

with respect to $\boldsymbol{\alpha}$.

2.2.1 Computing Algorithm

We now propose two effective algorithms to calculate the estimates.

One-step Estimator (OS)

Step 1: Obtain an estimate of the index parameter $\boldsymbol{\alpha}$.

Apply sliced inverse regression (Li, 1991) to obtain the estimate of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}$.

Step 2: Modified EM-type algorithm to maximize $\ell_1^{(1)}$ in (2.3).

In Step 2, we propose a modified EM-type algorithm to maximize $\ell_1^{(1)}$ and obtain the estimators $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$ and $\hat{\boldsymbol{\sigma}}^2(\cdot)$. 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 imply an EM algorithm, 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 . Let $\{u_t, t = 1, ..., N\}$ be a set of grid points where some unknown functions are evaluated, and N be the number of grid points.

E-step:

Calculate the expectations of component labels based on estimates from l^{th} iteration:

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i|m_j^{(l)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i), \sigma_j^{2(l)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i))}{\sum_{j=1}^{k} \pi_j^{(l)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i|m_j^{(l)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i), \sigma_j^{2(l)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i))}. \tag{2.5}$$

M-step:

Update the estimates

$$\pi_j^{(l+1)}(z) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z)}{\sum_{i=1}^n K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z)}, \tag{2.6}$$

$$m_j^{(l+1)}(z) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} Y_i K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z)}, \tag{2.7}$$

$$\sigma_j^{2(l+1)}(z) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - m_j^{(l+1)}(z))^2 K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z)}, \tag{2.8}$$

for $z \in \{u_t, t = 1, ..., N\}$. We then update $\pi_j^{(l+1)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)$, $m_j^{(l+1)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)$ and $\sigma_j^{2(l+1)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)$, $i = 1, ..., n$ by linear interpolating $\pi_j^{(l+1)}(u_t)$, $m_j^{(l+1)}(u_t)$ and $\sigma_j^{2(l+1)}(u_t)$, $t = 1, ..., N$, 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 (Yao and Lindsay, 2009).

Fully Iterative Backfitting Estimator (FIB)

To improve the estimation efficiency, we propose the following *fully iterative backfitting estimator*.

Step 1: Obtain an initial estimate of the index parameter $\boldsymbol{\alpha}$.

Apply sliced inverse regression to obtain an initial estimate of the index parameter $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}$.

Step 2: Modified EM-type algorithm to maximize $\ell_1^{(1)}$ in (2.3).

With $\hat{\boldsymbol{\alpha}}$, apply the modified EM-algorithm proposed above to obtain the estimators $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$, and $\hat{\boldsymbol{\sigma}}^2(\cdot)$.

Step 3: Updating the estimate of $\boldsymbol{\alpha}$ by maximizing $\ell_2^{(1)}$ in (2.4).

Given $\hat{\boldsymbol{\pi}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$, and $\hat{\boldsymbol{\sigma}}^2(\cdot)$ from Step 2, update the estimate of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}$, which maximizes $\ell_2^{(1)}$ defined in (2.4) using some numerical methods.

Step 4: Iterate Step 2 - 3 until convergence.

2.2.2 Asymptotic Properties

The asymptotic properties of the proposed estimates are investigated below.

Let $\boldsymbol{\theta}(z) = (\boldsymbol{\pi}^T(z), \boldsymbol{m}^T(z), (\boldsymbol{\sigma}^2)^T(z))^T$. Define $\ell(\theta(z), y) = \log \sum_{j=1}^{k} \pi_j(z)\phi\{y|m_j(z), \sigma_j^2(z)\}$, $q_1(z) = \frac{\partial \ell(\theta(z),y)}{\partial \theta}$, $q_2(z) = \frac{\partial^2 \ell(\theta(z),y)}{\partial \theta \partial \theta^T}$ and $\mathcal{I}_\theta^{(1)}(z) = -E[q_2(Z)|Z = z]$, $\Lambda_1(u|z) = E[q_1(z)|Z = u]$.

Under further conditions defined in Section 7, the properties of the one-step estimator when $\boldsymbol{\alpha}$ is estimated to the order of $O_p(n^{-1/2})$ (i.e., at the usual parametric rate) is demonstrated in the following theorem.

Theorem 2.2. *Assume that conditions (C1)-(C7) in Section 7 hold. Then, as $n \to \infty$, $h \to 0$ and $nh \to \infty$, we have*

$$\sqrt{nh}\{\hat{\boldsymbol{\theta}}(z) - \boldsymbol{\theta}(z) - \mathcal{B}_1 + o_p(h^2)\} \xrightarrow{D} N\{0, \nu_0 f^{-1}(z)\mathcal{I}_\theta^{(1)}(z)\}, \tag{2.9}$$

where $\mathcal{B}_1(z) = \mathcal{I}_\theta^{(1)-1}\left\{\frac{f'(z)\Lambda_1^{'}(z|z)}{f(z)} + \frac{1}{2}\Lambda_1^{''}(z|z)\right\}\kappa_2 h^2$, with $f(\cdot)$ the marginal density function of $\boldsymbol{\alpha}^T\boldsymbol{x}$, $\kappa_l = \int t^l K(t)dt$ and $\nu_l = \int t^l K^2(t)dt$.

Remark 1. The fully iterative backfitting estimator is at least as efficient as the one-step estimator, but the one-step estimator achieves the same efficiency in some important applications with added computational convenience. This information lower bound turns out to be the same as in Huang et al. (2013). Thus, the nonparametric functions can be estimated with the same rate of convergence as it would have if the one-dimension quantity $\boldsymbol{\alpha}^T\boldsymbol{x}$ were observable.

The next theorem shows that under further conditions, $\boldsymbol{\alpha}$ can be estimated at the usual parametric rate using the fully iterated algorithm.

Theorem 2.3. *Assume that conditions (C1)-(C8) in Section 7 hold. Then, as $n \to \infty$,*

$nh^4 \to 0$, *and* $nh^2/\log(1/h) \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{D} N(0, \boldsymbol{Q}_1^{-1}), \tag{2.10}$$

where $\boldsymbol{Q}_1 = E\left\{[\boldsymbol{x}\boldsymbol{\theta}'(Z)]q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^T\right\} - E\left[\boldsymbol{x}\boldsymbol{\theta}'(Z)q_2(Z)\mathcal{I}_\theta^{(1)-1}(Z)E\{q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^T|Z\}\right]$.

3 Mixture of Regression Models with Varying Single-Index Proportions (MRSIP)

3.1 Model Definition and Identifiability

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

$$Y|\boldsymbol{x} \sim \sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^T\boldsymbol{x})N(\boldsymbol{x}^T\boldsymbol{\beta}_j, \sigma_j^2). \tag{3.1}$$

Since $\pi_j(\cdot)$'s are nonparametric, model (3.1) is also a finite semiparametric mixture of regression models.

Theorem 3.1. *Assume that (i)* $\pi_j(z) > 0$ *are differentiable and not constant on the support of* $\boldsymbol{\alpha}^T\boldsymbol{x}$, $j = 1, ..., k$; *(ii) The component of* $\boldsymbol{x}$ *are continuously distributed random variables that have a joint probability density function; (iii) The support of* $\boldsymbol{x}$ *contains an open set in* $\mathbb{R}^p$ *and is not contained in any proper linear subspace of* $\mathbb{R}^p$; *(iv)* $\|\boldsymbol{\alpha}\| = 1$ *and the first nonzero element of* $\boldsymbol{\alpha}$ *is positive; (v)* $(\boldsymbol{\beta}_j, \sigma_j^2)$, $j = 1, ..., k$, *are distinct pairs. Then, model (3.1) is identifiable.*

3.2 Estimation Procedure and Asymptotic Properties

The log-likelihood of the collected data is:

$$\ell^{*(2)}(\boldsymbol{\pi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2)\}, \tag{3.2}$$

where $\boldsymbol{\pi}(\cdot) = \{\pi_1(\cdot), ..., \pi_{k-1}(\cdot)\}^T$, $\boldsymbol{\sigma}^2 = \{\sigma_1^2, ..., \sigma_k^2\}^T$, and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k\}^T$. Since $\boldsymbol{\pi}(\cdot)$ consists of nonparametric functions, (3.2) is not ready for maximization.

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

$$\ell_1^{(2)}(\boldsymbol{\pi}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\}K_h(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_i - z). \tag{3.3}$$

Let $\hat{\boldsymbol{\pi}}(\cdot)$ be the result of maximizing (3.3). We can then further update the estimate of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ by maximizing

$$\ell_2^{(2)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \hat{\pi}_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2)\}. \tag{3.4}$$

3.2.1 Computing Algorithm

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

Step 2: Modified EM-type algorithm to maximize $\ell_1^{(2)}$ in (3.3).

E-step:

Calculate the expectations of component labels based on estimates from l^{th} iteration:

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)}{\sum_{j=1}^{k} \pi_j^{(l)}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)}, j = 1, ..., k. \tag{3.5}$$

M-step:

Update the estimate

$$\pi_j^{(l+1)}(z) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z)}{\sum_{i=1}^n K_h(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i - z)} \tag{3.6}$$

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

Step 3: Update $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2)$ by maximizing (3.4).

Step 3.1: Given $\hat{\boldsymbol{\alpha}}$, update $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$.

E-step:

Calculate the expectations of component identities:

$$p_{ij}^{(l+1)} = \frac{\hat{\pi}_j(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i | \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(l)}, \sigma_j^{2(l)})}{\sum_{j=1}^k \hat{\pi}_j(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i | \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(l)}, \sigma_j^{2(l)})}, j = 1, ..., k. \tag{3.7}$$

M-step:

Update $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$:

$$\boldsymbol{\beta}_j^{(l+1)} = (\boldsymbol{S}^T \boldsymbol{R}_j^{(l+1)} \boldsymbol{S})^{-1} \boldsymbol{S}^T \boldsymbol{R}_j^{(l+1)} \boldsymbol{y}, \tag{3.8}$$

$$\sigma_j^{2(l+1)} = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(l+1)})^2}{\sum_{i=1}^n p_{ij}^{(l+1)}}, \tag{3.9}$$

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

Step 3.2: Given $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2)$, update $\boldsymbol{\alpha}$.

Given $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2)$, maximize $\ell_3^{(2)}(\boldsymbol{\alpha}) = \sum_{i=1}^n \log\{\sum_{j=1}^k \hat{\pi}_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\phi(Y_i | \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\}$ to updates the estimate of $\boldsymbol{\alpha}$, using some numerical methods.

Step 3.3: Iterate Step 3.1-3.2 until convergence.

Step 4: Iterate Step 2-3 until convergence.

3.2.2 Asymptotic Properties

Define $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, (\boldsymbol{\sigma}^2)^T)^T$, $\boldsymbol{\lambda} = (\boldsymbol{\alpha}^T, \boldsymbol{\eta}^T)^T$, and $\ell(\boldsymbol{\pi}(z), \boldsymbol{\lambda}, \boldsymbol{x}, y) = \log \sum_{j=1}^{k} \pi_j(z)\phi\{y|\boldsymbol{x}^T\boldsymbol{\beta}_j, \sigma_j^2\}$. Let $q_\pi(z) = \frac{\partial \ell(\pi(z),\lambda,x,y)}{\partial \pi}$, $q_{\pi\pi}(z) = \frac{\partial^2 \ell(\pi(z),\lambda,x,y)}{\partial \pi \partial \pi^T}$, and similarly, define q_λ, $q_{\lambda\lambda}$, and $q_{\pi\eta}$. Denote $\mathcal{I}_\pi^{(2)}(z) = -E[q_{\pi\pi}(Z)|Z = z]$ and $\Lambda_2(u|z) = E[q_\pi(z)|Z = u]$.

Under further conditions, the properties of the estimator when $\boldsymbol{\lambda}$ is estimated to the order of $O_p(n^{-1/2})$ (i.e., at the usual parametric rate) is demonstrated in the following theorem.

Theorem 3.2. *Assume that conditions (C1)-(C4) and (C9)-(C11) in Section 7 hold. Then, as $n \to \infty$, $h \to 0$ and $nh \to \infty$, we have*

$$\sqrt{nh}\{\hat{\boldsymbol{\pi}}(z) - \boldsymbol{\pi}(z) - \mathcal{B}_2(z) + o_p(h^2)\} \xrightarrow{D} N\{0, \nu_0 f^{-1}(z)\mathcal{I}_\pi^{(2)}(z)\}, \tag{3.10}$$

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

Theorem 3.3. *Assume that conditions (C1)-(C4) and (C9)-(C12) in Section 7 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,

$$\boldsymbol{Q}_2 = E\left[q_{\pi\pi}(Z)\begin{pmatrix}\boldsymbol{x}\boldsymbol{\pi}'(Z)\\ \boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\boldsymbol{\pi}'(Z)\\ \boldsymbol{I}\end{pmatrix}^T - q_{\pi\pi}(Z)\begin{pmatrix}\boldsymbol{x}\boldsymbol{\pi}'(Z)\\ \boldsymbol{I}\end{pmatrix}\begin{pmatrix}\mathcal{I}_\pi^{(2)-1}(Z)E\{q_{\pi\pi}(Z)(\boldsymbol{x}\boldsymbol{\pi}'(Z))^T|Z\}\\ \mathcal{I}_\pi^{(2)-1}(Z)E\{q_{\pi\eta}(Z)|Z\}\end{pmatrix}^T\right].$$

4 Numerical Studies

4.1 Simulation Study

In this section, we conduct simulation studies to test the performance of the proposed methodologies.

The performance of the estimates of the mean functions $m_j(\cdot)$'s in model (2.1) is measured by the square root of the average square errors (RASE)

$$RASE_m^2 = N^{-1} \sum_{j=1}^{k} \sum_{t=1}^{N} [\hat{m}_j(u_t) - m_j(u_t)]^2.$$

In this simulation, we set $N = 100$, and take the grid points are on the range. Similarly, we can define the $RASE$ for the variance functions $\sigma_j^2(\cdot)$'s and proportion functions $\pi_j(\cdot)$'s, denoted by $RASE_{\sigma^2}$ and $RASE_\pi$, respectively.

To apply the proposed methodologies, we use cross-validation (CV) to select a proper bandwidth for estimating the nonparametric functions.

4.1.1 Example 1.

We conduct a simulation for a 2-component MSIM:

$$\pi_1(z) = 0.5 + 0.3\sin(\pi z) \text{ and } \pi_2(z) = 1 - \pi_1(z),$$
$$m_1(z) = 3 - \sin(2\pi z/\sqrt{3}) \text{ and } m_2(z) = \cos(\sqrt{3}\pi z),$$
$$\sigma_1(z) = 0.7 + \sin(3\pi z)/15 \text{ and } \sigma_2(z) = 0.3 + \cos(1.3\pi z)/10.$$

where $z_i = \boldsymbol{\alpha}^T \boldsymbol{x}_i$, $\boldsymbol{x}_i$ are trivariate with independent uniform (0,1) components, and the direction parameter is $\boldsymbol{\alpha} = (1, 1, 1)/\sqrt{3}$. The sample sizes $n = 200$, $n = 400$, and $n = 800$ are conducted over 500 repetitions. To estimate $\boldsymbol{\alpha}$, we use sliced inverse regression (SIR) and the fully iterative backfitting estimate (FIB). To estimate the nonparametric functions, we apply the one-step estimate (OS) and FIB. For FIB, we use both true value (T) and SIR (S) as the initial values.

We first select a proper bandwidth for estimating $\boldsymbol{\pi}(\cdot)$, $\boldsymbol{m}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$. There are ways to calculate theoretical optimal bandwidth, but in practice, data driven methods, such as cross-validation (CV), are popularly used. 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 $l = 1, ..., L$. We use

the training set $\mathscr{R}_l$ to obtain the estimates $\{\hat{\boldsymbol{\pi}}(\cdot), \hat{\boldsymbol{m}}(\cdot), \hat{\boldsymbol{\sigma}}^2(\cdot), \hat{\boldsymbol{\alpha}}\}$. We then evaluate $\boldsymbol{\pi}(\cdot)$, $\boldsymbol{m}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ at the data in the corresponding training set. Then, for $(\boldsymbol{x}_t, y_t) \in \mathscr{T}_l$, we calculate the classification probability as

$$\hat{p}_{tj} = \frac{\hat{\pi}_j(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_t)\phi(y_t|\hat{m}_j(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_t), \hat{\sigma}_j^2(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_t))}{\sum_{j=1}^k \hat{\pi}_j(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_t)\phi(y_t|\hat{m}_j(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_t), \hat{\sigma}_j^2(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_t))} \tag{4.1}$$

for $j = 1, ..., k$. We then consider the regular CV, which is defined by

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

where $\hat{y}_t = \sum_{j=1}^k \hat{p}_{tj}\hat{m}_j(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_t)$.

We set $L = 10$ and randomly partition the data. We repeat the procedure 30 times, and take the average of the selected bandwidth as the optimal bandwidth, denoted by $\hat{h}$. In the simulation, we consider three different bandwidth, $\hat{h} \times n^{-2/15}$, $\hat{h}$ and $1.5\hat{h}$, which correspond to the under-smoothing, appropriate smoothing and over-smoothing condition, respectively.

Table 1 reports the MSEs of $\hat{\boldsymbol{\alpha}}$ (true value times 100). From Table 1, we can see that the fully iterative estimates give better results than SIR. We further notice that FIB(S) provides similar results to FIB(T), and therefore, SIR provides good initial values for other estimates.

Table 2 contains the mean and standard deviation of $RASE_\pi$, $RASE_m$, and $RASE_{\sigma^2}$. We see that the fully iterative estimate is not sensitive to initial values.

4.1.2 Example 2.

We conduct a simulation for a 2-component MRSIP:

$$\pi_1(z) = 0.5 - 0.35\sin(\pi z) \text{ and } \pi_2(z) = 1 - \pi_1(z),$$

Table 1: MSE of $\hat{\boldsymbol{\alpha}}$ (true value times 100)

		SIR	FIB(T)			FIB(S)		
			$h = 0.054$	$h = 0.109$	$h = 0.164$	$h = 0.054$	$h = 0.109$	$h = 0.164$
	α_1	0.881	0.099	0.126	0.128	0.287	0.130	0.147
$n = 200$	α_2	0.829	0.113	0.144	0.124	0.324	0.144	0.137
	α_3	1.066	0.110	0.152	0.137	0.388	0.154	0.167
			$h = 0.045$	$h = 0.100$	$h = 0.149$	$h = 0.045$	$h = 0.100$	$h = 0.149$
	α_1	0.435	0.066	0.046	0.046	0.125	0.050	0.045
$n = 400$	α_2	0.447	0.063	0.054	0.051	0.121	0.055	0.052
	α_3	0.411	0.062	0.052	0.052	0.123	0.053	0.052
			$h = 0.037$	$h = 0.091$	$h = 0.137$	$h = 0.037$	$h = 0.091$	$h = 0.137$
	α_1	0.215	0.047	0.022	0.029	0.063	0.035	0.024
$n = 800$	α_2	0.256	0.034	0.035	0.040	0.044	0.029	0.027
	α_3	0.226	0.065	0.031	0.058	0.062	0.050	0.030

$$m_1(\boldsymbol{x}) = 1 + 3x_2 \text{ and } m_2(\boldsymbol{x}) = -1 + 2x_1 + 3x_3,$$
$$\sigma_1^2 = 0.7 \text{ and } \sigma_2^2 = 0.6,$$

where $m_1(\boldsymbol{x})$ and $m_2(\boldsymbol{x})$ are the regression functions for the first and second components, respectively. Therefore, $\boldsymbol{\beta}_1 = (1, 0, 3, 0)$ and $\boldsymbol{\beta}_2 = (-1, 2, 0, 3)$. $\boldsymbol{x}_i$ are trivariate with independent uniform (0,1) components, and the direction parameter is $\boldsymbol{\alpha} = (1, 1, 1)/\sqrt{3}$. MRSIP with true value (T) and SIR (S) as initial values are used to fit the data, and the results are compared to a two-component mixture of linear regression models (MixLinReg).

Table 3 reports the MSEs of parameter estimates, and Table 4 contains the MSEs of $\hat{\boldsymbol{\alpha}}$ and the average of $RASE_\pi$. From both tables, we can see that MRSIP works comparable to MixLinReg when the sample size is small, and outperforms MixLinReg when sample size is big. We further notice that MRSIP(S) provides similar results to MRSIP(T), implying that SIR provides good initial values for MRSIP.

Table 2: Mean and Standard Deviation of RASEs

n		OS	FIB(T)			FIB(S)		
		$h = 0.125$	$h = 0.054$	$h = 0.109$	$h = 0.164$	$h = 0.054$	$h = 0.109$	$h = 0.164$
	π	0.044(0.017)	0.057(0.015)	0.043(0.016)	0.049(0.017)	0.058(0.015)	0.043(0.016)	0.049(0.017)
200	μ	0.227(0.063)	0.181(0.098)	0.176(0.046)	0.287(0.056)	0.178(0.086)	0.177(0.051)	0.288(0.059)
	σ^2	0.197(0.084)	0.175(0.169)	0.163(0.081)	0.246(0.071)	0.162(0.131)	0.164(0.095)	0.247(0.080)
		$h = 0.108$	$h = 0.045$	$h = 0.100$	$h = 0.149$	$h = 0.045$	$h = 0.100$	$h = 0.149$
	π	0.023(0.008)	0.032(0.008)	0.023(0.008)	0.027(0.009)	0.032(0.008)	0.023(0.008)	0.027(0.009)
400	μ	0.118(0.022)	0.093(0.045)	0.100(0.022)	0.169(0.020)	0.094(0.046)	0.100(0.022)	0.169(0.020)
	σ^2	0.104(0.035)	0.089(0.077)	0.093(0.045)	0.143(0.028)	0.089(0.077)	0.093(0.045)	0.143(0.028)
		$h = 0.094$	$h = 0.037$	$h = 0.091$	$h = 0.137$	$h = 0.037$	$h = 0.091$	$h = 0.137$
	π	0.013(0.004)	0.017(0.003)	0.012(0.004)	0.016(0.004)	0.017(0.003)	0.012(0.004)	0.016(0.004)
800	μ	0.062(0.010)	0.050(0.023)	0.056(0.010)	0.102(0.011)	0.050(0.023)	0.056(0.010)	0.101(0.010)
	σ^2	0.055(0.015)	0.049(0.046)	0.052(0.015)	0.086(0.010)	0.049(0.046)	0.051(0.012)	0.085(0.010)

5 Real Data Example

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 National Basket Association (NBA). Of interest is how the height of the player (Height), minutes per game (MPG) and free throw percentage (FTP) affects 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. We see that Michael Jordan is an outlier in terms of PPM, so we will also omit him from the data (Chatterjee et al., 1995). These excludes 10 players. We divide each variable by its corresponding standard deviation, so that they have comparable numerical scale. An optimal bandwidth is selected at 0.344 by

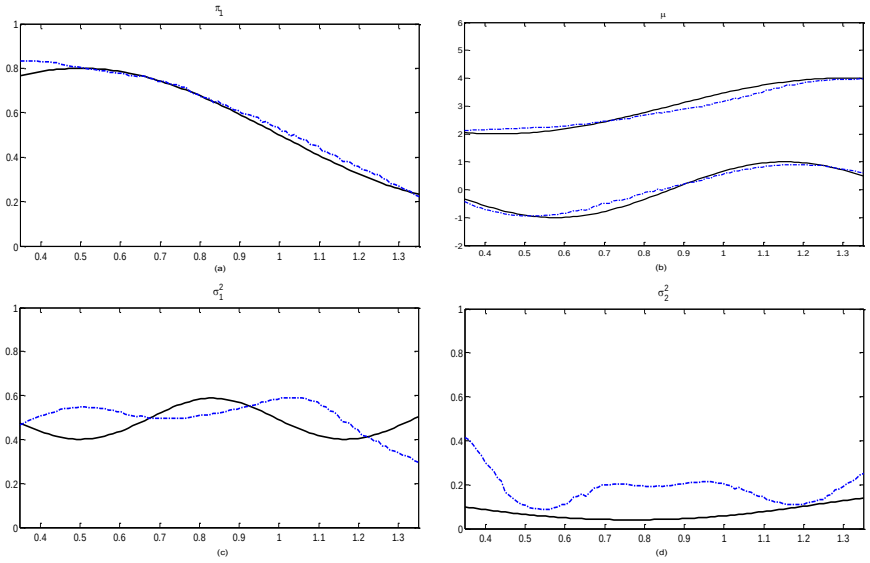

Figure 1: Simulation results of for $n = 800$ and $h = 0.091$. The FIB(S) (dashed line) and true function (solid line) of: (a) π_1; (b) μ_1 and μ_2; (c) σ_1^2; and (d) σ_2^2.

CV procedure. Figure 2(a) contains the estimated mean functions and hard-clustering results, denoted by dots and squares, respectively. The 95% confidence interval for $\hat{\boldsymbol{\alpha}}$ based on MSIM are (0.134,0.541), (0.715,0.949) and (0.202,0.679), indicating that MPG is the most influential factor on PPM.

To evaluate the prediction performance of the proposed model and compared it to linear regression model and mixture of linear regression models, we used d-fold cross-validation with d=5,10, and also Monte-Carlo cross-validation (MCCV) (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 over 500 replications are reported as boxplots in Figure 2(b). Apparently, the MSIM and the MRSIP have superior prediction power than the linear regression model or the mixture of linear regression models, and MSIM is more favorable then the MRSIP for this data set.

Table 3: The MSEs of parameters (true value times 100)

		β_{10}	β_{11}	β_{12}	β_{13}	β_{20}	β_{21}	β_{22}	β_{23}	σ_1^2	σ_2^2
$n = 200$	MRSIP(S)	46.37	32.78	34.73	37.61	11.19	16.55	15.05	16.36	4.649	1.754
	MRSIP(T)	51.91	33.62	39.01	37.25	11.10	16.56	15.07	16.04	4.584	1.649
$h = 0.131$	MixLinReg	50.87	33.67	42.53	34.68	12.03	12.66	18.84	12.30	4.250	1.265
$n = 400$	MRSIP(S)	13.83	11.89	14.19	11.47	5.541	6.332	6.767	7.165	1.631	0.721
	MRSIP(T)	14.79	12.49	14.84	11.59	5.513	6.254	6.632	6.926	1.672	0.675
$h = 0.103$	MixLinReg	29.03	14.97	29.46	15.72	8.045	5.967	12.46	6.269	1.864	0.626
$n = 800$	MRSIP(S)	6.324	4.491	6.150	4.736	2.365	2.973	2.773	3.584	0.669	0.334
	MRSIP(T)	6.788	4.614	6.820	4.922	2.301	2.829	2.718	3.348	0.691	0.307
$h = 0.080$	MixLinReg	21.89	6.866	21.84	8.223	5.413	3.163	8.775	3.640	0.848	0.352

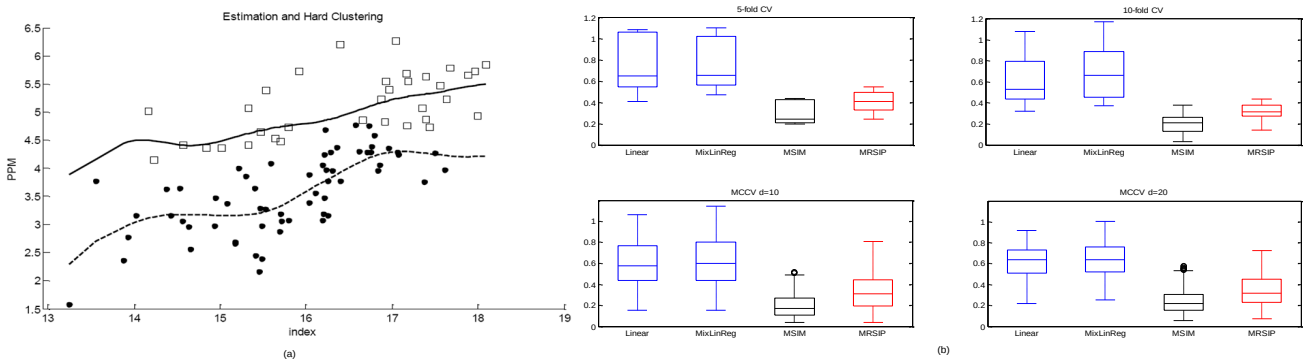

Figure 2: NBA data: (a) Estimated mean functions and a hard-clustering result; (b) Prediction accuracy: 5-fold CV; 10-fold CV; MCCV d=10; MCCV d=20.

6 Discussion

In this paper we proposed two finite semiparametric mixture of regression models and the corresponding backfitting estimates. We showed that the nonparametric functions can be estimated with the same rate as if the parameters were known and the parameters can be estimated with root-n convergence rate. In this 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. In addition, it

Table 4: The MSEs of direction parameter and the average of RASE_π (true value times 100)

		α_1	α_2	α_3	RASE_π
$n = 200$	MRSIP(S)	5.709	19.30	5.996	18.87
	MRSIP(T)	4.984	9.449	4.896	17.86
$h = 0.131$	MixLinReg	-	-	-	28.98
$n = 400$	MRSIP(S)	2.682	6.968	3.029	13.74
	MRSIP(T)	2.113	3.019	1.902	12.98
$h = 0.103$	MixLinReg	-	-	-	28.23
$n = 800$	MRSIP(S)	0.980	2.527	1.585	10.35
	MRSIP(T)	0.892	0.979	0.969	9.960
$h = 0.080$	MixLinReg	-	-	-	28.04

is also interesting to build some formal test to compare the proposed two semiparametric mixture models. One way is to apply generalized likelihood ratio statistic proposed by Fan et al., (2001).

7 Proofs

. **Technical Conditions:**

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

(C2) The marginal density of $\boldsymbol{\alpha}^T\boldsymbol{x}$, 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 $|\partial^3 \ell(\boldsymbol{\theta}, y)/\partial\theta_i\partial\theta_j\partial\theta_k| \leq M(y)$ 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 $j = 1, ..., k$, $\sigma_j^2(z) > 0$, and $\pi_j(z) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$.

(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 $|\partial^3 \ell(\boldsymbol{\pi}, y)/\partial\pi_i\partial\pi_j\partial\pi_k| \leq M(y)$ 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 $j = 1, ..., k$, $\pi_j(z) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$.

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

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

(C12) $\boldsymbol{\pi}''(\cdot)$ is continuous at the point z.

Proof of Theorem 2.1.

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 2.2.

Let

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

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

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

If $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{\sigma}}^2)^T$ maximizes (2.3), then $\hat{\boldsymbol{\theta}}^*$ maximizes

$$\ell_n^*(\boldsymbol{\theta}^*) = h\sum_{i=1}^{n}[\ell(\boldsymbol{\theta}(z) + a_n\boldsymbol{\theta}^*, \hat{\boldsymbol{\alpha}}, \boldsymbol{x}_i, Y_i) - \ell(\boldsymbol{\theta}(z), \hat{\boldsymbol{\alpha}}, \boldsymbol{x}_i, Y_i)]K_h(\hat{Z}_i - z) \tag{7.1}$$

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

$$\ell_n^*(\boldsymbol{\theta}^*) = \boldsymbol{W}_{1n}^T\boldsymbol{\theta}^* + \frac{1}{2}\boldsymbol{\theta}^{*T}\boldsymbol{A}_{1n}\boldsymbol{\theta}^* + o_p(1), \tag{7.2}$$

where

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

and

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

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

$$\ell_n^*(\boldsymbol{\theta}^*) = \boldsymbol{W}_{1n}^T\boldsymbol{\theta}^* - \frac{1}{2}f(z)\boldsymbol{\theta}^{*T}\mathcal{I}_\theta^{(1)}(z)\boldsymbol{\theta}^* + o_p(1). \tag{7.3}$$

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

$$\hat{\boldsymbol{\theta}}^* = f(z)^{-1}\mathcal{I}_\theta^{(1)}(z)^{-1}\boldsymbol{W}_{1n} + o_p(1). \tag{7.4}$$

Note that

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

where

$$D_{1n} = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\left\{\frac{\partial^2\ell(\boldsymbol{\theta}(z),\boldsymbol{\alpha},\boldsymbol{x}_i,Y_i)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}[\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)]^T K_h(Z_i - z)\right\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}).$$

Since $\sqrt{n}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}) = O_p(1)$, it can be shown that $D_{1n} = -\sqrt{h}f(z)E[\frac{\partial^2\ell(\boldsymbol{\theta}(z),\boldsymbol{\alpha},\boldsymbol{x},Y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^T] = o_p(1)$, and $O_p(\sqrt{\frac{h}{n}}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|^2) = o_p(1)$. 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

$$\begin{aligned} E(\boldsymbol{W}_{1n}) &= \sqrt{nh}E\left[E[\frac{\partial\ell(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{x}_i,Y_i)}{\partial\boldsymbol{\theta}}K_h(Z_i - z)|Z = z_0]\right] \\ &= \sqrt{nh}[\frac{1}{2}f(z)\Lambda_1''(z|z) + f'(z)\Lambda_1'(z|z)]\kappa_2 h^2. \end{aligned} \tag{7.5}$$

Similarly, we can show that $\mathrm{Cov}(\boldsymbol{W}_{1n}) = f(z)\mathcal{I}_\theta^{(1)}(z)\nu_0 + o_p(1)$, 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 2.3.

Denote $Z = \boldsymbol{\alpha}^T \boldsymbol{x}$ and $\hat{Z} = \hat{\boldsymbol{\alpha}}^T \boldsymbol{x}$. 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 (2.3), then it solves

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

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

$$\begin{aligned}\mathbf{0} = {} & n^{-1} \sum_{i=1}^n q_{1i}(Z_i) K_h(Z_i - z_0) + n^{-1} \sum_{i=1}^n \left[q_{2i}(Z_i) K_h(Z_i - z_0)\right] \left(\hat{\boldsymbol{\theta}}(z_0; \hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(z_0)\right) \\ & + n^{-1} \sum_{i=1}^n q_{2i}(Z_i) [\boldsymbol{x}_i \boldsymbol{\theta}'(Z_i)]^T K_h(Z_i - z_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + o_p(n^{-1/2}) + O_p(h^2)\end{aligned}$$

By similar argument as in the previous proof,

$$\begin{aligned}\hat{\boldsymbol{\theta}}(z_0; \hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(z_0) = {} & n^{-1} f^{-1}(z_0) \mathcal{I}_\theta^{(1)-1}(z_0) \sum_{i=1}^n q_{1i}(Z_i) K_h(Z_i - z_0) \\ & - \mathcal{I}_\theta^{(1)-1}(z_0) E\{q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^T | Z = z_0\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + o_p(n^{-1/2}) \qquad (7.6)\end{aligned}$$

Note that

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^T \boldsymbol{x}_i) &= \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^T \boldsymbol{x}_i) \\ &= (\hat{\boldsymbol{\theta}}'(\boldsymbol{\alpha}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}))^T (\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}^T) \boldsymbol{x}_i + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}_0^T \boldsymbol{x}_i) + o_p(n^{-1/2}) \\ &= (\boldsymbol{\theta}'(\boldsymbol{\alpha}^T \boldsymbol{x}_i))^T (\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}^T) \boldsymbol{x}_i + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^T \boldsymbol{x}_i) + o_p(n^{-1/2}) \qquad (7.7)\end{aligned}$$

where the second part is handled by (7.6).

Since $\hat{\boldsymbol{\alpha}}$ maximizes (2.4), it is the solution to

$$\mathbf{0} = \lambda \hat{\boldsymbol{\alpha}} + n^{-1/2} \sum_{i=1}^n \boldsymbol{x}_i \hat{\boldsymbol{\theta}}'(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \frac{\partial \ell(\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}), X_i, Y_i)}{\partial \boldsymbol{\theta}},$$

where λ is the Lagrange multiplier. By the Taylor expansion and using (7.7), we have that

$$\begin{aligned}
\mathbf{0} &= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{1i}(Z_i) + n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)[\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_i) - \boldsymbol{\theta}(\boldsymbol{\alpha}^T\boldsymbol{x}_i)] + o_p(1) \\
&= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{1i}(Z_i) + n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)(\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i))^T(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\
&+ n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)[\hat{\boldsymbol{\theta}}(Z_i) - \boldsymbol{\theta}(Z_i)]) + o_p(1).
\end{aligned}$$

Define

$$A_\alpha = E\{[\boldsymbol{x}\boldsymbol{\theta}'(Z)]q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^T\},$$

and apply (7.6),

$$\begin{aligned}
\mathbf{0} &= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{1i}(Z_i) + n^{1/2}A_\beta(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\
&- n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)\mathcal{I}_\theta^{-1}(Z_i)E\{q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^T|Z = Z_i\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\
&+ n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)n^{-1}f^{-1}(Z_i)\mathcal{I}_\theta^{-1}(Z_i)\sum_{t=1}^{n}q_{1t}(Z_t)K_h(Z_t - Z_i) + o_p(1) \\
&= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{1i}(Z_i) + \boldsymbol{Q}_1 n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\
&+ n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)n^{-1}f^{-1}(Z_i)\mathcal{I}_\theta^{(1)-1}(Z_i)\sum_{t=1}^{n}q_{1t}(Z_t)K_h(Z_t - Z_i) + o_p(1).
\end{aligned} \tag{7.8}$$

Interchanging the summations in the last term, we get

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\left[n^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{\theta}'(Z_t)q_{2t}(Z_t)K_h(Z_t - Z_i)f^{-1}(Z_t)\mathcal{I}_\theta^{-1}(Z_t)q_{1i}(Z_i)\right] \\
&= n^{-1/2}\sum_{i=1}^{n}E[\boldsymbol{x}\boldsymbol{\theta}'(Z)q_2(Z)|Z_i]\mathcal{I}_\theta^{(1)-1}(Z_i)q_{1i}(Z_i) + o_p(1)
\end{aligned} \tag{7.9}$$

Let $\Gamma_\alpha = I - \boldsymbol{\alpha}\boldsymbol{\alpha}^T + o_p(1)$. Combining (7.8) and (7.9), and multiply by Γ_α, we have

$$\Gamma_\alpha \boldsymbol{Q}_1 n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = n^{-1/2} \sum_{i=1}^{n} \Gamma_\alpha \{\boldsymbol{x}_i \boldsymbol{\theta}'(Z_i) + E[\boldsymbol{x}\boldsymbol{\theta}'(Z) q_2(Z)|Z_i] \mathcal{I}_\theta^{(1)-1}(Z_i)\} q_{1i}(Z_i) + o_p(1) \tag{7.10}$$

It can be shown that the right-hand side of (7.10) has the covariance matrix $\Gamma_\alpha \boldsymbol{Q}_1 \Gamma_\alpha$, and therefore, completes the proof. □

Proof of Theorem 3.1

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 3.2

This proof is similar to the proof of 2.2

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

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

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

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

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

Proof of Theorem 3.3

The proof is similar to the proof of Theorem 2.3. It can be shown that

$$\hat{\boldsymbol{\pi}}(z_0;\hat{\boldsymbol{\lambda}}) - \boldsymbol{\pi}(z_0) = n^{-1}f^{-1}(z_0)\mathcal{I}_{\pi}^{(2)-1}(z_0)\sum_{i=1}^{n} q_{\pi i}(Z_i)K_h(Z_i - z_0)$$

$$-\mathcal{I}_{\pi}^{(2)-1}(z_0)E\{q_{\pi\pi}(Z)[\boldsymbol{x}\boldsymbol{\pi}'(Z)]^T|Z = z_0\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - \mathcal{I}_{\pi}^{(2)-1}(z_0)E\{q_{\pi\eta}(Z)|Z = z_0\}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + o_p(n^{-1/2}),$$

and therefore,

$$\hat{\boldsymbol{\pi}}(\hat{Z}_i;\hat{\boldsymbol{\lambda}}) - \boldsymbol{\pi}(Z_i) = \{\boldsymbol{x}_i\boldsymbol{\pi}'(Z_i)\}^T(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \hat{\boldsymbol{\pi}}(Z_i;\hat{\boldsymbol{\lambda}}) - \boldsymbol{\pi}(Z_i) + o_p(n^{-\frac{1}{2}}). \tag{7.11}$$

Since $\hat{\boldsymbol{\lambda}}$ maximizes (3.4), it is the solution to

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

where γ is the Lagrange multiplier. By Taylor series and (7.11)

$$\begin{aligned}\mathbf{0} =&\gamma\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ \mathbf{0}\end{pmatrix} + n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\Lambda}_{1i}q_{\pi i}(Z_i) + n^{\frac{1}{2}}\boldsymbol{Q}_2\begin{pmatrix}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\end{pmatrix}\\ &+n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\Lambda}_{1i}q_{\pi\pi i}(Z_i)n^{-1}f^{-1}(Z_i)\mathcal{I}_{\pi}^{(2)-1}(Z_i)\sum_{j=1}^{n}q_{\pi j}(Z_j)K_h(Z_j - Z_i) + o_p(1)\\ =&\gamma\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ \mathbf{0}\end{pmatrix} + n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\Lambda}_{1i}q_{\pi i}(Z_i) + n^{\frac{1}{2}}\boldsymbol{Q}_2\begin{pmatrix}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\end{pmatrix}\\ &+n^{-\frac{1}{2}}\sum_{i=1}^{n}E[\boldsymbol{\Lambda}_{1i}q_{\pi\pi}(Z_i)]\mathcal{I}_{\pi}^{(2)-1}(Z_i)q_{\pi i}(Z_i) + o_p(1).\end{aligned} \tag{7.12}$$

where $\mathbf{\Lambda}_{1i} = \begin{pmatrix} \boldsymbol{x}_i \boldsymbol{\pi}'(Z_i) \\ \mathbf{I} \end{pmatrix}$, and the last equation is the result of interchanging the summations. Let $\Gamma_\alpha = \begin{pmatrix} \boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\alpha}^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} + o_p(1)$. By (7.12), and multiply by Γ_α, we have

$$n^{\frac{1}{2}} \Gamma_\alpha \boldsymbol{Q}_2 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \Gamma_\alpha \left\{ \mathbf{\Lambda}_{1i} - \mathcal{I}_\pi^{(2)-1}(Z_i) E[\mathbf{\Lambda}_{1i}(Z_i) q_{\pi\pi}(Z_i) | Z_i] \right\} q_{\pi i}(Z_i) + o_p(1). \tag{7.13}$$

It can be shown that the right-hand side of (7.13) has the covariance matrix $\Gamma_\alpha \boldsymbol{Q}_2 \Gamma_\alpha$, and thus, completes the proof. □

References

Böhning, D. (1999), *Computer-Assisted Analysis of Mixtures and Applications*, Boca Raton, FL: Chapman and Hall/CRC.

Chatterjee, S., Handcock, M.S. and Simmonoff, J.S. (1995). *A casebook for a first course in statistics and data analysis*. John Wiley & Sons, Inc.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, 30, 455-474.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, 29, 153-193.

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

Goldfeld, S.M. and Quandt, R.E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, 1, 3-6.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21, 157-178.

Henning, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17, 273-296.

Huang, M., Li, R. and Wang, S. (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108, 929-941.

Huang, M. and Yao, W. (2012). Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association*, 107, 711-724.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58, 71-120.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*. 6, 181214.

Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414), 316-327.

Lindsay, B. G., (1995), *Mixture Models: Theory, Geometry, and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics v 5, Hayward, CA: Institure of Mathematical Statistics.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Shao, J. (1993). Linear models selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.

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

Xiang, S. and Yao, W. (2015). Semiparametric mixtures of nonparametric regressions. *Computational Statistics & Data Analysis*, submitted for publication.

Young, D.S. and Hunter, D.R. (2010). Mixtures of regressions with predictors dependent mixing proportions. *Computational Statistics and Data Analysis*, 54, 2253-2266.

Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of American Statistical Association*, 104, 758-767.