SEMIPARAMETRIC MIXTURE MODELS

by

SIJIA XIANG

M.S., Kansas State University, 2012

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2014

Abstract

This dissertation consists of three parts that are related to semiparametric mixture models.

In Part I, we construct the minimum profile Hellinger distance (MPHD) estimator for a class of semiparametric mixture models where one component has known distribution with possibly unknown parameters while the other component density and the mixing proportion are unknown. Such semiparametric mixture models have been often used in biology and the sequential clustering algorithm.

In Part II, we propose a new class of semiparametric mixture of regression models, where the mixing proportions and variances are constants, but the component regression functions are smooth functions of a covariate. A one-step backfitting estimate and two EM-type algorithms have been proposed to achieve the optimal convergence rate for both the global parameters and nonparametric regression functions. We derive the asymptotic property of the proposed estimates and show that both proposed EM-type algorithms preserve the asymptotic ascent property.

In Part III, we apply the idea of single-index model to the mixture of regression models and propose three new classes of models: the mixture of single-index models (MSIM), the mixture of regression models with varying single-index proportions (MRSIP), and the mixture of regression models with varying single-index proportions and variances (MRSIPV). Backfitting estimates and the corresponding algorithms have been proposed for the new models to achieve the optimal convergence rate for both the parameters and the nonparametric functions. We show that the nonparametric functions can be estimated 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.

SEMIPARAMETRIC MIXTURE MODELS

by

SIJIA XIANG

M.S., Kansas State University, 2012

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2014

Approved by:

Major Professor
Weixin Yao

Copyright

Abstract

This dissertation consists of three parts that are related to semiparametric mixture models.

In Part I, we construct the minimum profile Hellinger distance (MPHD) estimator for a class of semiparametric mixture models where one component has known distribution with possibly unknown parameters while the other component density and the mixing proportion are unknown. Such semiparametric mixture models have been often used in biology and the sequential clustering algorithm.

In Part II, we propose a new class of semiparametric mixture of regression models, where the mixing proportions and variances are constants, but the component regression functions are smooth functions of a covariate. A one-step backfitting estimate and two EM-type algorithms have been proposed to achieve the optimal convergence rate for both the global parameters and nonparametric regression functions. We derive the asymptotic property of the proposed estimates and show that both proposed EM-type algorithms preserve the asymptotic ascent property.

In Part III, we apply the idea of single-index model to the mixture of regression models and propose three new classes of models: the mixture of single-index models (MSIM), the mixture of regression models with varying single-index proportions (MRSIP), and the mixture of regression models with varying single-index proportions and variances (MRSIPV). Backfitting estimates and the corresponding algorithms have been proposed for the new models to achieve the optimal convergence rate for both the parameters and the nonparametric functions. We show that the nonparametric functions can be estimated 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.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my appreciation to Dr. Weixin Yao, my major professor, for all his knowledge, guidance and suggestions. My sincere appreciation also goes to Dr. Gabriel Nagy, for his willingness to serve as the chairperson of the examining committee for my doctoral degree. I would also like to thank Dr. Dong Li, Dr. Weixing Song and Dr. Christopher I. Vahl for their willingness to serve on my committee and for their valuable insight.

I would like to thank all my friends for their help and support. I would also like to thank everyone in the department for their kindness. It is my pleasure to study in this department.

Finally, I would also like to thank my family for their endless love, support, understanding and encouragement. Thanks to my parents, who have always supported whatever I wanted to do in my life. Thank you also to my grandparents, who always check to see how I am doing and offer their support.

Chapter 1

Minimum Profile Hellinger Distance Estimation For A Semiparametric Mixture Model

Abstract

In this chapter, we propose a new effective estimator for a class of semiparametric mixture models where one component has known distribution with possibly unknown parameters while the other component density and the mixing proportion are unknown. Such semiparametric mixture models have been often used in multiple hypothesis testing and the sequential clustering algorithm. The proposed estimator is based on the minimum profile Hellinger distance (MPHD), and its theoretical properties are investigated. In addition, we use simulation studies to illustrate the finite sample performance of the MPHD estimator and compare it with some other existing approaches. The empirical studies demonstrate that the new method outperforms existing estimators when data are generated under contamination and works comparably to existing estimators when data are not contaminated. Applications to two real data sets are also provided to illustrate the effectiveness of the new methodology.

1.1 Introduction

The two-component mixture model considered in this chapter is defined by

$$h(x) = \pi f_0(x;\xi) + (1-\pi) f(x-\mu), \quad x \in \mathbb{R}, \tag{1.1}$$

where $f_0(x;\xi)$ is a known probability density function (pdf) with possibly unknown parameter ξ, f is an unknown pdf with non-null location parameter $\mu \in \mathbb{R}$, and π is the unknown mixing proportion.

Bordes et al. (2006) studied a special case when ξ is assumed to be known, i.e., the first component density is completely known and model (1.1) becomes

$$h(x) = \pi f_0(x) + (1-\pi) f(x-\mu), \quad x \in \mathbb{R}. \tag{1.2}$$

Model (1.2) is motivated by multiple hypothesis testing to detect differentially expressed genes under two or more conditions in microarray data. Please see Bordes et al. (2006) for more detail about the application of model (1.2) to microarray data analysis. For this purpose, we build a test statistic for each gene. The test statistics can be considered as coming from a mixture of two distributions: the known distribution f_0 under null hypothesis, and the other distribution $f(\cdot - \mu)$, the unknown distribution of the test statistics under the alternative hypothesis. Please see Section 1.5 for such an application on multiple hypothesis testing.

Song et al. (2010) studied another special case of model (1.1),

$$h(x) = \pi \phi_\sigma(x) + (1-\pi) f(x), \quad x \in \mathbb{R}, \tag{1.3}$$

where ϕ_σ is a normal density with mean 0 and unknown standard deviation σ and $f(x)$ is an unknown density. Model (1.3) was motivated by a sequential clustering algorithm (Song and Nicolae, 2009), which works by finding a local center of a cluster first, and then

identifying whether an object belongs to that cluster or not. If we assume that the objects belonging to the cluster come from a normal distribution with known mean (such as zero) and unknown variance σ^2 and that the objects not belonging to the cluster come from an unknown distribution f, then identifying the points in the cluster is equivalent to estimating the mixing proportion in model (1.3).

Note that the semiparametric mixture model (1.1) is not generally identifiable without any assumption for f. Specifically, Bordes et al. (2006) showed that the model (1.2) is not generally identifiable if we do not put any restriction on the unknown density f, but identifiability can be achieved under some sufficient conditions. One of these conditions is that $f(\cdot)$ is symmetric about 0. Under these conditions, Bordes et al. (2006) proposed an elegant estimation procedure based on the symmetry of f. Song et al. (2010) also addressed the unidentifiability problem and noticed that model (1.3) is not generally identifiable. However, due to the additional unknown parameter σ in the first component, Song et al. (2010) mentioned that it is hard to find the conditions to avoid unidentifiability of model (1.3) and proposed to use simulation studies to check the performance of proposed estimators. Please refer to Bordes et al. (2006) and Song et al. (2010) for detailed discussions on the identifiability of model (1.1).

Bordes et al. (2006) proposed to estimate model (1.2) based on symmetrization of the unknown distribution f and proved the consistency of their estimator. However, the asymptotic distribution of their estimator has not been provided. Song et al. (2010) also proposed an EM-type estimator and a maximizing $\pi-$type estimator (inspired by the constraints imposed to achieve identifiability of the parameters and Swanepoel's approach (Swanepoel, 1999)) to estimate model (1.3) without providing any asymptotic properties.

In this chapter, we propose a new estimation procedure for the unified model (1.1) based on minimum profile Hellinger distance (MPHD) (Wu et al., 2011). We will investigate the theoretical properties of the proposed MPHD estimator for the semiparametric mixture model, such as existence, consistency, and asymptotic normality. A simple and effective

algorithm is also given to compute the proposed estimator. Using simulation studies, we illustrate the effectiveness of the MPHD estimator and compare it with the estimators suggested by Bordes et al. (2006) and Song et al. (2010). Compared to the existing methods (Bordes et al., 2006; Song et al. 2010), the new method can be applied to the more general model (1.1). In addition, the MPHD estimator works competitively under semiparametric model assumptions, while it is more robust than the existing methods when data are contaminated.

Donoho and Liu (1988) have shown that the class of minimum distance estimators has automatic robustness properties over neighborhoods of the true model based on the distance functional defining the estimator. However, minimum distance estimators typically obtain this robustness at the expense of not being optimal at the true model. Beran (1977) has suggested the use of the minimum Hellinger distance (MHD) estimator which has certain robustness properties and is asymptotically efficient at the true model. For a comparison between MHD estimators, MLEs, and other minimum distance type estimators, and the balance between robustness and efficiency of estimators, see Lindsay (1994).

There are other well-known robust approaches within the mixture model-based clustering literature. García-Escudero et al. (2003) proposed exploratory graphical tools based on trimming for detecting main clusters in a given dataset, where the trimming is obtained by resorting to trimmed k-means methodology. García-Escudero et al. (2008) introduced a new method for performing clustering with the aim of fitting clusters with different scatters and weights. García-Escudero et al. (2010) reviewed different robust clustering approaches in the literature, emphasizing on methods based on trimming which try to discard most outlying data when carrying out the clustering process. A more recent work by Punzo and McNicholas (2013) introduced a family of fourteen parsimonious mixtures of contaminated Gaussian distributions models within the general model-based classification framework.

The rest of the chapter is organized as follows. In Section 1.2, we review models (1.2) and (1.3) and the existing estimation methods suggested by Bordes et al. (2006) and Song

et al. (2010), respectively. In Section 1.3, we introduce the proposed MPHD estimator and discuss its asymptotic properties. Section 1.4 presents simulation results for comparing the new estimation with some existing methods. Applications to two real data sets are also provided in Section 1.5 to illustrate the effectiveness of the proposed methodology. A discussion section ends the chapter.

1.2 Review of Some Existing Methods

1.2.1 Estimation by symmetrization

Bordes et al. (2006) proposed an inference procedure based on the symmetry of the unknown component of model (1.2). Let $X_1, ..., X_n$ be random variables from model (1.2) and H be the cumulative distribution function (cdf) of model (1.2), i.e.

$$H(x) = \pi F_0(x) + (1-\pi)F(x-\mu), \quad x \in \mathbb{R}, \tag{1.4}$$

where H, F_0 and F are the corresponding cdf's of h, f_0 and f. Assume that H in (1.4) is identifiable, then

$$F(x) = \frac{1}{1-\pi}\left[H(x+\mu) - \pi F_0(x+\mu)\right], \quad x \in \mathbb{R}.$$

Let

$$D_1(x;\pi,\mu,H) = \frac{1}{1-\pi}H(x+\mu) + (1-\frac{1}{1-\pi})F_0(x+\mu),$$
$$D_2(x;\pi,\mu,H) = 1 - \frac{1}{1-\pi}H(\mu-x) + (\frac{1}{1-\pi}-1)F_0(\mu-x).$$

Since f is assumed to be symmetric, $F(x) = 1-F(-x)$ for all $x \in \mathbb{R}$. Thus $D_1(\cdot;\pi_0,\mu_0,H) = D_2(\cdot;\pi_0,\mu_0,H)$, where π_0 and μ_0 are the unknown true values of π and μ. Consequently,

with d a distance measure, say, L_2-norm, we have $d(D_1(\cdot;\pi_0,\mu_0,H), D_2(\cdot;\pi_0,\mu_0,H)) = 0$, where

$$d(\pi,\mu) = \|D_1 - D_2\|_2 = \left(\int |D_1(x;\pi,\mu,H) - D_2(x;\pi,\mu,H)|^2 dx\right)^{1/2}.$$

Since H is unknown, it can be estimated by

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x), \quad x \in \mathbb{R},$$

where $I(\cdot)$ is the indicator function. With H replaced by H_n, we get an empirical version d_n of d defined by $d_n(\pi,\mu) = d(D_1(\cdot;\pi,\mu,H_n), D_2(\cdot;\pi,\mu,H_n))$. Bordes et al. (2006) proposed to estimate π and μ in model (1.2) by minimizing $d_n(\pi,\mu)$.

1.2.2 EM-type estimator

Let

$$Z_i = \begin{cases} 1, & \text{if } X_i \text{ is from the first component;} \\ 0, & \text{otherwise.} \end{cases}.$$

Song et al. (2010) proposed the following EM-type estimator for model (1.3).

E-step: In the $(k+1)^{\text{th}}$ iteration, compute the conditional expectation of Z_i, given the data and the parameters from the k^{th} iteration, as

$$Z_i^{(k+1)} = E(Z_i|\pi^{(k)},\sigma^{(k)},X_i) = \frac{\pi^{(k)}\phi_{\sigma^{(k)}}(X_i)}{\widehat{h(X_i)}},$$

where

$$\widehat{h(x)} = \frac{1}{nc}\sum_{i=1}^{n} K\left(\frac{x - X_i}{c}\right)$$

with K being a kernel function, such as Gaussian kernel, and c the bandwidth.

M-step: The values of the parameters are updated as follows:

$$\pi^{(k+1)} = \frac{\sum_{i=1}^{n} Z_i^{(k+1)}}{n},$$
$$\sigma^{(k+1)} = \sqrt{\frac{\sum_{i=1}^{n} Z_i^{(k+1)} X_i^2}{\sum_{i=1}^{n} Z_i^{(k+1)}}}.$$

In addition, Song et al. (2010) recommended, in the E-step, to use

$$Z_i^{(k+1)} = \frac{2\pi^{(k)}\phi_{\sigma^{(k)}}(X_i)}{\pi^{(k)}\phi_{\sigma^{(k)}}(X_i) + \widehat{h(X_i)}}$$

truncated to 1 when it is greater than 1 to stabilize the Z_i values.

1.2.3 Maximizing π-type estimator

Song et al. (2010) demonstrated that, based on their simulation studies, the EM-type estimator introduced in Section 1.2.2 is biased when the two component densities overlap significantly. Therefore, they proposed an alternative estimator, by finding the maximum mixing proportion π that satisfies

$$\pi\phi_\sigma(X_i) \leq \widehat{h(X_i)}, \quad i = 1, ..., n.$$

Therefore, we can estimate π by

$$\hat{\pi} = \max_{\sigma} \min_{X_i} \frac{\widehat{h(X_i)}}{\phi_\sigma(X_i)}$$

with $\widehat{h(X_i)}$ defined in (1.2.2), and estimate σ by

$$\hat{\sigma} = \arg\max_{\sigma} \min_{X_i} \frac{\widehat{h(X_i)}}{\phi_\sigma(X_i)}.$$

More detailed explanation of this estimator is presented in Song et al. (2010) and therefore omitted here.

1.3 MPHD Estimation

1.3.1 Introduction of MPHD estimator

In this section, we develop a MPHD estimator for model (1.1). Let

$$\mathscr{H} = \{h_{\boldsymbol{\theta},f}(x) = \pi f_0(x;\xi) + (1-\pi)f(x-\mu) : \boldsymbol{\theta} \in \Theta, f \in \mathscr{F}\},$$

where

$$\Theta = \{\boldsymbol{\theta} = (\pi, \xi, \mu) : \pi \in (0,1), \xi \in \mathbb{R}, \mu \in \mathbb{R}\},$$
$$\mathscr{F} = \{f : f \geq 0, \int f(x)dx = 1\}$$

be the functional space for the semiparametric model (1.1). In practice, the parameter space of ξ depends on its interpretation. For example, if ξ is the standard deviation of f_0, then the parameter space of ξ will be $\mathbb{R}^+$. For model (1.2), ξ is known and thus the parameter space of ξ is a singleton and, as a result, $\boldsymbol{\theta} = (\pi, \mu)$.

Let $\|\cdot\|$ denote the $L_2(v)$-norm. For any $g_1, g_2 \in L_2(v)$, the Hellinger distance between them is defined as

$$d_H(g_1, g_2) = \left\| g_1^{1/2} - g_2^{1/2} \right\|.$$

Suppose a sample $X_1, X_2, ..., X_n$ is from a population with density function $h_{\boldsymbol{\theta},f} \in \mathscr{H}$. We propose to estimate $\boldsymbol{\theta}$ and f by minimizing the Hellinger distance

$$\left\| h_{t,l}^{1/2} - \hat{h}_n^{1/2} \right\| \tag{1.5}$$

over all $t \in \Theta$ and $l \in \mathscr{F}$, where $\hat{h}_n$ is an appropriate nonparametric density estimator of $h_{\boldsymbol{\theta},f}$. Note that the above objective function (1.5) contains both the parametric component t and the nonparametric component l. Here, we propose to use the profile idea to implement the calculation.

For any density function g and t, define functional $f(t,g)$ as

$$f(t,g) = \arg\min_{l\in\mathscr{F}} \left\| h_{t,l}^{1/2} - g^{1/2} \right\|$$

and then define the profile Hellinger distance as

$$d_{PH}(t,g) = \| h_{t,f(t,g)}^{1/2} - g^{1/2} \|.$$

Now the MPHD functional $T(g)$ is defined as

$$T(g) \;=\; \arg\min_{t\in\Theta} d_{PH}(t,g) \;=\; \arg\min_{t\in\Theta} \left\| h_{t,f(t,g)}^{1/2} - g^{1/2} \right\|. \tag{1.6}$$

Given the sample $X_1, X_2, ..., X_n$, one can construct an appropriate nonparametric density estimator of $h_{\boldsymbol{\theta},f}$, say $\hat{h}_n$, and then the proposed MPHD estimator of $\boldsymbol{\theta}$ is given by $T(\hat{h}_n)$. In the examples of Section 1.4 and Section 1.5, we use kernel density estimator for $\hat{h}_n$ and the bandwidth h is chosen based on Botev et al.(2010).

1.3.2 Algorithm

In this section, we propose the following two-step algorithm to calculate the proposed MPHD estimator. Suppose the initial estimates of $\boldsymbol{\theta} = (\pi, \xi, \mu)$ and f are $\boldsymbol{\theta}^{(0)} = (\pi^{(0)}, \xi^{(0)}, \mu^{(0)})$ and $f^{(0)}$.

Step 1: Given $\pi^{(k)}, \xi^{(k)}$ and $\mu^{(k)}$, find $f^{(k+1)}$ which minimizes

$$\left\| [\pi^{(k)} f_0(\cdot;\xi^{(k)}) + (1-\pi^{(k)}) f^{(k+1)}(\cdot - \mu^{(k)})]^{1/2} - \hat{h}_n^{1/2}(\cdot) \right\|.$$

Similar to Wu et al. (2011), we obtain that

$$f^{(k+1)}(x-\mu^{(k)}) = \begin{cases} \dfrac{\alpha}{1-\pi^{(k)}}\hat{h}_n(x) - \dfrac{\pi^{(k)}}{1-\pi^{(k)}} f_0(x;\xi^{(k)}), & \text{if } x \in M, \\ 0, & \text{if } x \in M^C, \end{cases}$$

where $M = \{x : \alpha\hat{h}_n(x) \geq \pi^{(k)} f_0(x;\xi^{(k)})\}$ and $\alpha = \sup_{0<\alpha\leq 1} \{\pi^{(k)} \int_M f_0(x;\xi^{(k)})dx + (1-\pi^{(k)}) \geq \alpha \int_M \hat{h}_n(x)dx\}$.

Step 2: Given fixed $f^{(k+1)}$, find $\pi^{(k+1)}, \xi^{(k+1)}$, and $\mu^{(k+1)}$ which minimize

$$\left\| [\pi^{(k+1)} f_0(\cdot;\xi^{(k+1)}) + (1-\pi^{(k+1)}) f^{(k+1)}(\cdot - \mu^{(k+1)})]^{1/2} - \hat{h}_n^{1/2}(\cdot) \right\| \tag{1.7}$$

Then go back to **Step 1**.

Each of the above two steps monotonically decreases the objective function (1.5) until convergence. In Step 1, if $f(\cdot)$ is assumed to be symmetric, then we can further symmetrize $f^{(k+1)}(\cdot)$ as

$$\tilde{f}^{(k+1)}(x) = \frac{f^{(k+1)}(x) + f^{(k+1)}(-x)}{2}.$$

Note that there is no closed form for (1.7) in Step 2 and thus some numerical algorithms, such as Newton-Raphson algorithm, is needed to mimimize (1.7). In our examples, we used the "fminsearch" function in matlab to find the minimizer numerically. "fminsearch" function uses the Nelder-Mead simplex algorithm as described in Lagarias et al. (1998).

1.3.3 Asymptotic results

Note that $\boldsymbol{\theta}$ and f in the semiparametric mixture model (1.1) are not generally identifiable without any assumptions for f. Bordes et al. (2006) showed that model (1.2) is not generally identifiable if we do not put any restrictions on the unknown density f, but identifiability

can be achieved under some sufficient conditions. One of these conditions is that $f(\cdot)$ is symmetric about 0. Under these conditions, Bordes et al. (2006) proposed an elegant estimation procedure based on the symmetry of f. Song et al. (2010) also addressed the non-identifiability problem and noticed that model (1.3) is not generally identifiable. However, due to the additional unknown parameter σ in the first component, Song et al. (2010) mentioned that it is hard to find the conditions to avoid unidentifiability of model (1.3) and proposed using simulation studies to check the performance of the proposed estimators. Please refer to Bordes et al. (2006) and Song et al. (2010) for detailed discussions on the identifiability of model (1.1).

Next, we discuss some asymptotic properties of the proposed MPHD estimator. Here, for simplicity of explanation, we will only consider model (1.2) for which Bordes et al. (2006) proved identifiability. However, we conjecture that all the results presented in this section also apply to the unified model (1.1) when it is identifiable. But this is beyond the scope of the article and requires more research to find the identifiable conditions for the general model (1.1).

The next theorem gives results on the existence and uniqueness of the proposed estimator, and the continuity of the functional defined in (1.6), which is in line with Theorem 1 of Beran (1977).

Theorem 1.3.1. *With T defined by (1.6), if model (1.2) is identifiable, then we have*

(i) For every $h_{\boldsymbol{\theta},f} \in \mathscr{H}$, there exists $T(h_{\boldsymbol{\theta},f}) \in \Theta$ satisfying (1.6);

(ii) $T(h_{\boldsymbol{\theta},f}) = \boldsymbol{\theta}$ uniquely for any $\boldsymbol{\theta} \in \Theta$;

(iii) $T(h_n) \to T(h_{\boldsymbol{\theta},f})$ for any sequences $\{h_n\}_{n\in\mathbb{N}}$ such that $\left\| h_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} \right\| \to 0$ and

$$\sup_{t\in\Theta} \left\| h_{t,f(t,h_n)} - h_{t,f(t,h_{\boldsymbol{\theta},f})} \right\| \to 0$$

as $n \to \infty$.

Remark. Bordes et al. (2006) provided sufficient conditions for the identifiability of model (1.2). For example, model (1.2) is identifiable if $f > 0$ has a first-order moment, and there exists a real number $a > 0$ such that for all $|x| > a$ we have $f_0(x) = 0$ and $f(x) = f(-x)$. Readers are referred to Bordes et al. (2006) for a detailed discussion of the identifiability of model (1.2) and other sufficient conditions for identifiability. Without the global identifiability of model (1.2), the local identifiability of model (1.2) proved by Bordes et al. (2006) tells that there exists one solution that has the asymptotic properties presented in Theorem 1.3.1.

Define a kernel density estimator based on $X_1, X_2, ..., X_n$ as:

$$\hat{h}_n(x) = \frac{1}{nc_n s_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{c_n s_n}\right), \tag{1.8}$$

where $\{c_n\}$ is a sequence of constants (bandwidths) converging to zero at an appropriate rate, and s_n is a robust scale parameter.

Under further conditions on the kernel density estimator defined in (1.8), the consistency of the MPHD estimator is established in the next theorem.

Theorem 1.3.2. *Suppose that*

(i) The kernel function $K(\cdot)$ is absolutely continuous and bounded with compact support.

(ii) $\lim_{n\to\infty} c_n = 0$, $\lim_{n\to\infty} n^{1/2} c_n = \infty$.

(iii) The model (1.2) is identifiable and $h_{\boldsymbol{\theta},f}$ is uniformly continuous.

Then $\|\hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2}\| \xrightarrow{p} 0$ *as* $n \to \infty$, *and therefore* $T(\hat{h}_n) \xrightarrow{p} T(h_{\boldsymbol{\theta},f})$ *as* $n \to \infty$.

Define the map $\boldsymbol{\theta} \mapsto s_{\boldsymbol{\theta},g}$ as $s_{\boldsymbol{\theta},g} = h_{\boldsymbol{\theta},f(\boldsymbol{\theta},g)}^{1/2}$, and suppose that for $\boldsymbol{\theta} \in \Theta$, there exists a 2×1 vector $\dot{s}_{\boldsymbol{\theta}}(x)$ with components in L_2 and a 2×2 matrix $\ddot{s}_{\boldsymbol{\theta}}$ with components in L_2 such that for every 2×1 real vector e of unit Euclidean length and for every scalar α in a

neighborhood of zero,

$$s_{\boldsymbol{\theta}+\alpha e}(x) = s_{\boldsymbol{\theta}}(x) + \alpha e^T \dot{s}_{\boldsymbol{\theta}}(x) + \alpha e^T u_\alpha(x), \tag{1.9}$$

$$\dot{s}_{\boldsymbol{\theta}+\alpha e}(x) = \dot{s}_{\boldsymbol{\theta}}(x) + \alpha \ddot{s}_{\boldsymbol{\theta}}(x)e + \alpha v_\alpha(x)e, \tag{1.10}$$

where $u_\alpha(x)$ is 2×1, $v_\alpha(x)$ is 2×2, and the components of u_α and v_α tend to zero in L_2 as $\alpha \to 0$.

The next theorem shows that the MPHD estimator has an asymptotic normal distribution.

Theorem 1.3.3. *Suppose that*

(i) Model (1.2) is identifiable.

(ii) The conditions in Theorem 1.3.2 hold.

(iii) The map $\boldsymbol{\theta} \mapsto s_{\boldsymbol{\theta},g}$ satisfies (1.9) and (1.10) with continuous gradient vector $\dot{s}_{\boldsymbol{\theta},g}$ and continuous Hessian matrix $\ddot{s}_{\boldsymbol{\theta},g}$ in the sense that $\|\dot{s}_{\boldsymbol{\theta}_n,g_n} - \dot{s}_{\boldsymbol{\theta},g}\| \to 0$ and $\|\ddot{s}_{\boldsymbol{\theta}_n,g_n} - \ddot{s}_{\boldsymbol{\theta},g}\| \to 0$ whenever $\boldsymbol{\theta}_n \to \boldsymbol{\theta}$ and $\|g_n^{1/2} - g^{1/2}\| \to 0$ as $n \to \infty$.

(iv) $< \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h^{1/2}_{\boldsymbol{\theta},f} >$ is invertible.

Then, with T defined in (1.6) for model (1.2), the asymptotic distribution of $n^{1/2}(T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f}))$ is $N(0, \Sigma)$ with variance matrix Σ defined by

$$\Sigma = < \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h^{1/2}_{\boldsymbol{\theta},f} >^{-1} < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \dot{s}^T_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}} > < \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h^{1/2}_{\boldsymbol{\theta},f} >^{-1} .$$

1.4 Simulation Studies

In this section, we investigate the finite sample performance of the proposed MPHD estimator and compare it to Maximizing-π type estimator (Song et al., 2010), EM-type estimator (Song et al., 2010), and the Symmetrization estimator (Bordes et al., 2006) under both model (1.2) and model (1.3).

Model (1.3) that Song et al. (2010) considered does not have a location parameter in the second component. However, we can equivalently replace $f(x)$ with $f(x-\mu)$, where $\mu \in \mathbb{R}$ is a location parameter. Throughout this section, we will consider this equivalent form of (1.3). Under this model, after we have $\hat{\pi}$ and $\hat{\sigma}$, we can simply estimate μ by

$$\hat{\mu} = \frac{\sum_{i=1}^{n}(1-\hat{Z}_i)X_i}{\sum_{i=1}^{n}(1-\hat{Z}_i)},$$

where $\hat{Z}_i$ is

$$\hat{Z}_i = \frac{2\hat{\pi}\phi_{\hat{\sigma}}(X_i)}{\hat{\pi}\phi_{\hat{\sigma}}(X_i) + \hat{h}(X_i)}.$$

We first compare the performance of different estimators under model (1.2). Suppose $(X_1, \ldots, X_n)$ are generated from one of the following five cases:

Case I: $X \sim 0.3N(0,1) + 0.7N(1.5,1) \Rightarrow (\pi, \mu) = (0.3, 1.5)$,

Case II: $X \sim 0.3N(0,1) + 0.7N(3,1) \Rightarrow (\pi, \mu) = (0.3, 3)$,

Case III: $X \sim 0.3N(0,1) + 0.7U(2,4) \Rightarrow (\pi, \mu) = (0.3, 3)$,

Case IV: $X \sim 0.7N(0,4) + 0.3N(3,1) \Rightarrow (\pi, \mu) = (0.7, 3)$,

Case V: $X \sim 0.85N(0,4) + 0.15N(3,1) \Rightarrow (\pi, \mu) = (0.85, 3)$.

Figure 1.1 shows the density plots of the five cases. Cases I, II, and III are the models used by Song et al. (2010) to show the performance of their Maximizing-π type and EM-type estimators. Case I represents the situation when two components are close and Case

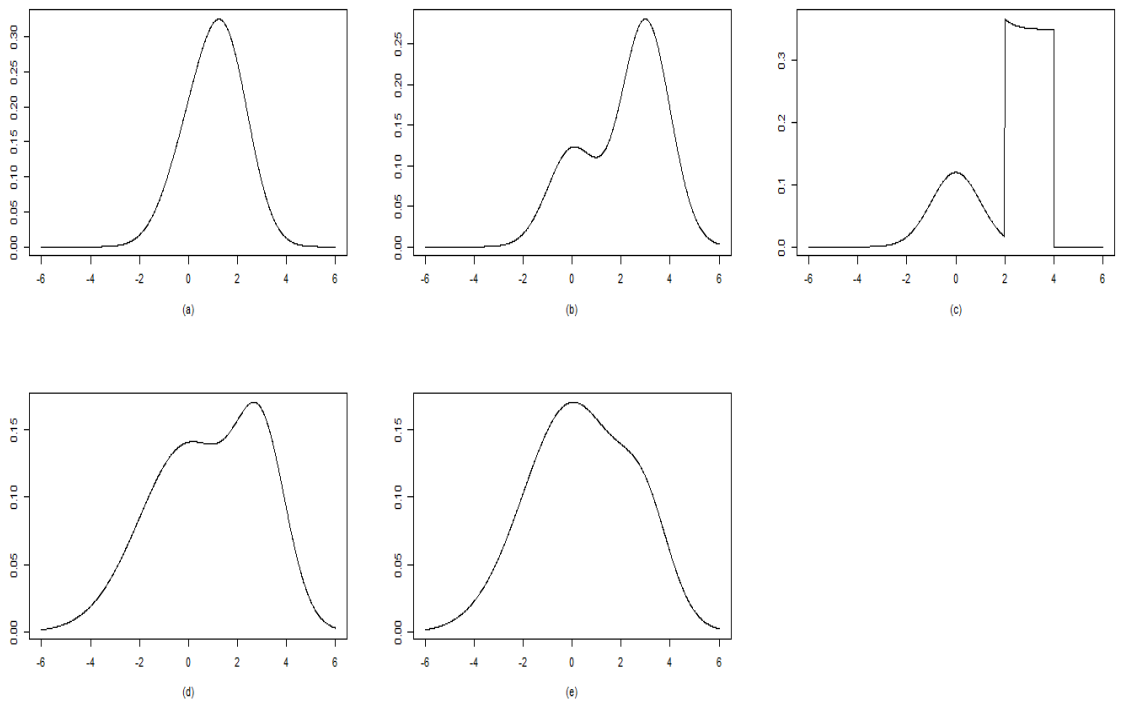

Figure 1.1: *Density plots of: (a) Case I; (b) Case II; (c) Case III; (d) Case IV and (e) Case V.*

II represents the situation when two components are apart. Cases IV and V are suggested by Bordes et al. (2006) to show the performance of their semiparametric EM algorithm. In addition, we also consider the corresponding contaminated models by adding 2% outliers from $U(10, 20)$ to the above five models.

Tables 1.1, 1.2 and 1.3 report the bias and MSE of the parameter estimates of (π, μ) for the four methods when $n = 100$, $n = 250$ and $n = 1000$, respectively, based on 200 repetitions. Tables 1.4, 1.5 and 1.6 report the respective results for $n = 100$, $n = 250$ and for $n = 1000$ when the data are under 2% contamination from $U(10, 20)$. The best values are highlighted in bold. From the six tables, we can see that the MPHD estimator has better overall performance than the Maximizing-π type, the EM-type, and the Symmetrization estimators, especially when sample size is large. When the sample is not contaminated by outliers, the MPHD estimator and the Symmetrization estimator are very competitive

and perform better than other estimators. When the sample is contaminated by outliers, the MPHD estimator performs much better and therefore is more robust than the other three methods. We also observe that when the sample is contaminated by outliers, among the Maximizing-π type, the EM-type, and the Symmetrization estimators, the EM-type estimator tends to give better mixing proportion estimates than the other two.

Table 1.1: *Bias (MSE) of point estimates for model (1.2) over 200 repetitions with* $n = 100$

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.092(0.030)	0.057(0.011)	0.271(0.078)	**0.003(0.009)**
	$\mu : 1.5$	-0.113(0.118)	0.196(0.070)	0.465(0.239)	**0.020(0.026)**
II	$\pi : 0.3$	**-0.014(0.003)**	-0.052(0.005)	**0.027(0.003)**	**-0.002(0.003)**
	$\mu : 3$	**-0.000(0.021)**	-0.123(0.038)	**0.020(0.017)**	-0.009(0.025)
III	$\pi : 0.3$	-0.046(0.005)	-0.108(0.014)	-0.045(0.005)	**0.001(0.003)**
	$\mu : 3$	**-0.008(0.004)**	-0.341(0.138)	-0.212(0.058)	**-0.002(0.006)**
IV	$\pi : 0.7$	**-0.044(0.015)**	-0.131(0.025)	**0.086(0.010)**	-0.089(0.028)
	$\mu : 3$	0.173(0.247)	-0.697(0.659)	**-0.053(0.177)**	-0.326(0.465)
V	$\pi : 0.85$	-0.094(0.041)	-0.147(0.030)	**0.039(0.003)**	-0.106(0.024)
	$\mu : 3$	**0.109(1.145)**	-1.375(2.298)	**-0.697(1.136)**	-0.742(1.184)

Next, we also evaluate how the MPHD estimator performs under model (1.3), *where the variance* σ^2 *is assumed to be unknown*, and compare it with other methods using the same five cases as in Tables 1.1-1.6. Tables 1.7, 1.8, and 1.9 report the bias and MSE of the parameter estimates for $n = 100$, $n = 250$ and $n = 1000$, respectively, when there are no contaminations. Based on these three tables, we can see that when there are no contaminations, the MPHD estimator and the Symmetrization estimator perform better than the Maximizing-π type estimator and the EM-type estimator. Tables 1.10, 1.11, and 1.12 report the results when models are under 2% contamination from $U(10, 20)$ for $n = 100$, $n = 250$, and $n = 1000$, respectively. From these three tables, we can see that the MPHD estimator performs much better again than the other three methods.

Table 1.2: *Bias (MSE) of point estimates for model (1.2) over 200 repetitions with* $n = 250$

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.090(0.028)	**0.028(0.005)**	0.269(0.074)	-0.080(0.021)
	$\mu : 1.5$	-0.110(0.084)	**0.162(0.041)**	0.472(0.231)	**-0.107(0.060)**
II	$\pi : 0.3$	**-0.009(0.001)**	-0.058(0.005)	0.034(0.002)	**-0.001(0.001)**
	$\mu : 3$	**0.007(0.007)**	-0.118(0.027)	0.057(0.009)	**-0.004(0.009)**
III	$\pi : 0.3$	-0.041(0.003)	-0.071(0.006)	**-0.016(0.001)**	**-0.001(0.001)**
	$\mu : 3$	**-0.001(0.001)**	-0.188(0.043)	-0.082(0.010)	**-0.001(0.002)**
IV	$\pi : 0.7$	**-0.009(0.003)**	-0.108(0.018)	0.102(0.012)	-0.017(0.009)
	$\mu : 3$	**0.131(0.067)**	-0.618(0.501)	**0.063(0.069)**	-0.095(0.159)
V	$\pi : 0.85$	**-0.040(0.014)**	-0.121(0.021)	**0.052(0.003)**	-0.041(0.011)
	$\mu : 3$	**0.217(0.444)**	-1.134(1.503)	**-0.323(0.349)**	-0.345(0.625)

To see the comparison and difference better, we also plot in Figures 1.2-1.4 the results reported in Tables 1.6 and 1.9. Figure 1.2 contains the MSE of point estimates of μ that are presented in Table 1.9 for model (1.3) (σ unknown), and Figures 1.3 and 1.4 contain the MSEs of point estimates of μ and π, respectively, that are presented in Table 1.6 for model (1.2) (σ known) under 2% contamination from $U(10, 20)$. From the plots, we can see that all four estimators perform well in cases II and III. The EM-type estimator performs poorly in case I, and is the worst estimate of μ in cases IV and V when data are contaminated. The Symmetrization estimator is sensitive to contamination, especially in cases IV and V, no matter σ is known or not. Comparatively, the Maximizing-π type estimator is more robust, but it doesn't perform well in cases IV and V when data are not under contamination. However, the MPHD estimator performs well in all cases.

Table 1.3: *Bias (MSE) of point estimates for model (1.2) over 200 repetitions with* $n = 1000$

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	**-0.009(0.005)**	**-0.020(0.003)**	0.263(0.069)	-0.024(0.005)
	$\mu : 1.5$	**0.003(0.016)**	0.083(0.017)	0.459(0.213)	**-0.031(0.015)**
II	$\pi : 0.3$	**-0.006(0.001)**	-0.055(0.004)	0.039(0.002)	**-0.003(0.001)**
	$\mu : 3$	**0.006(0.002)**	-0.083(0.016)	0.093(0.010)	**-0.002(0.002)**
III	$\pi : 0.3$	**-0.028(0.001)**	-0.061(0.005)	**-0.004(0.001)**	**0.000(0.001)**
	$\mu : 3$	**-0.003(0.001)**	-0.153(0.029)	-0.044(0.002)	**-0.002(0.001)**
IV	$\pi : 0.7$	**-0.008(0.001)**	-0.115(0.020)	0.104(0.011)	**-0.007(0.001)**
	$\mu : 3$	**0.045(0.013)**	-0.554(0.400)	0.174(0.039)	**-0.030(0.017)**
V	$\pi : 0.85$	**-0.007(0.001)**	-0.101(0.016)	0.061(0.004)	**-0.007(0.002)**
	$\mu : 3$	**0.172(0.063)**	-0.929(1.043)	**0.019(0.067)**	-0.066(0.104)

1.5 Real Data Application

Example 1(Iris data). We illustrate the application of the new estimation procedure to the sequential clustering algorithm using the Iris data, which is perhaps one of the best known data sets in pattern recognition literature. Iris data was first introduced by Fisher (1936) and is referenced frequently to this day. This data contains four attributes: sepal length (in cm), sepal width (in cm), petal length (in cm), and petal width (in cm), and there are 3 classes of 50 instances each, where each class refers to a type of Iris plant. One class is linearly separable from the other two and the latter are not linearly separable from each other.

Assuming the class indicators are unknown, we want to recover the three clusters in the data. After applying the search algorithm for centers of clusters by Song et al. (2010), observation 8 is selected as the center of the first cluster. We adjust all observations by subtracting observation 8 from each observation. As discussed by Song et al. (2010), the proportion of observations that belong to a cluster can be considered as the mixing proportion in the two-component semiparametric mixture model (1.3).

Table 1.4: *Bias (MSE) of point estimates for model (1.2), under* 2% *contamination from* $U(10, 20)$*, over 200 repetitions with* $n = 100$

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.124(0.036)	**0.060(0.010)**	0.267(0.075)	-0.063(0.014)
	$\mu : 1.5$	-0.163(0.128)	0.692(0.629)	1.079(1.348)	**-0.031(0.015)**
II	$\pi : 0.3$	-0.029(0.005)	-0.055(0.006)	**0.018(0.004)**	-0.300(0.090)
	$\mu : 3$	**-0.011(0.046)**	0.252(0.136)	0.398(0.228)	-3.000(9.000)
III	$\pi : 0.3$	**-0.034(0.003)**	-0.108(0.015)	-0.048(0.005)	**-0.032(0.004)**
	$\mu : 3$	**-0.011(0.004)**	-0.034(0.080)	0.104(0.091)	-0.014(0.009)
IV	$\pi : 0.7$	**-0.054(0.020)**	-0.133(0.027)	**0.081(0.009)**	-0.200(0.083)
	$\mu : 3$	**0.152(0.389)**	0.172(0.668)	1.141(2.123)	-0.582(0.867)
V	$\pi : 0.85$	-0.125(0.071)	-0.158(0.033)	**0.024(0.002)**	-0.217(0.080)
	$\mu : 3$	0.048(1.364)	**-0.007(1.314)**	1.373(4.337)	-0.910(1.444)

Principal component analysis shows that the first principal component accounts for 92.46% of the total variability, so it would seem that the Iris data tend to fall within a one-dimensional subspace of the 4-dimensional sample space. Figure 1.5 is a histogram of the first principal component. From the histogram, we can see that the first cluster is separated from the rest of the data, with observation 8 (first principal component score equals -2.63) being the center of it. The first principal component loading vector is $(0.36, -0.08, 0.86, 0.35)$, which implies that the petal length contains most of the information. We apply each of the four estimation methods discussed above to the first principal component. Note however that the leading principal components are not necessary to have better clustering information than other components. Some cautious are needed when using principal components in clustering applications.

Similar to Song et al. (2010), in Table 1.13, we report the estimates of proportion based on the first principal component. Noting that the true proportion is 1/3, we can see that the MPHD and the Symmetrization estimators perform better than the other two estimators.

Example 2 (Breast cancer data). Next, we illustrate the application of the new estima-

Table 1.5: *Bias (MSE) of point estimates for model (1.2), under* 2% *contamination from* $U(10, 20)$*, over 200 repetitions with* $n = 250$

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.090(0.026)	**0.032(0.006)**	0.263(0.071)	-0.180(0.043)
	$\mu : 1.5$	**-0.102(0.085)**	0.613(0.434)	1.043(1.146)	**-0.224(0.081)**
II	$\pi : 0.3$	**-0.019(0.001)**	-0.065(0.006)	0.027(0.002)	-0.044(0.003)
	$\mu : 3$	**-0.009(0.007)**	0.213(0.076)	0.415(0.202)	-0.044(0.012)
III	$\pi : 0.3$	**-0.021(0.001)**	-0.073(0.007)	**-0.015(0.001)**	-0.028(0.002)
	$\mu : 3$	**-0.004(0.001)**	0.119(0.043)	0.245(0.086)	-0.011(0.003)
IV	$\pi : 0.7$	**-0.020(0.005)**	-0.122(0.021)	0.086(0.009)	-0.302(0.164)
	$\mu : 3$	**0.149(0.096)**	0.162(0.296)	1.149(1.594)	-0.746(1.137)
V	$\pi : 0.85$	-0.053(0.025)	-0.131(0.023)	**0.034(0.002)**	-0.311(0.140)
	$\mu : 3$	**0.220(0.513)**	0.358(1.000)	1.859(4.597)	-1.093(1.785)

tion procedure to multiple hypothesis testing using the breast cancer data from Hedenfalk et al. (2001), who examined gene expressions in breast cancer tissues from women who were carriers of the hereditary BRCA1 or BRCA2 gene mutations, predisposing to breast cancer. The breast cancer data was downloaded from "http://research.nhgri.nih.gov/microarray/NEJM_Supplement/" and contains gene expression ratios derived from the fluorescent intensity (proportional to the gene expression level) from a tumor sample divided by the fluorescent intensity from a common reference sample (MCF-10A cell line). The ratios were normalized (or calibrated) such that the majority of the gene expression ratios from a pre-selected internal control gene set was around 1.0, but no log-transformation was used. The data set consists of 3,226 genes on $n_1 = 7$ BRCA1 arrays and $n_2 = 8$ BRCA2 arrays. If any gene had one or more measurement exceeding 20, then this gene was eliminated (Storey and Tibshirani, 2003). This left 3,170 genes. The p-values were calculated based on permutation tests (Storey and Tibshirani, 2003). We then transform the p-values via the probit transformation to z-score, given by $z_i = \Phi^{-1}(1 - p_i)$ (McLachlan and Wockner, 2010). Figure 1.6 displays the fitted densities, and Table 1.14

Table 1.6: *Bias (MSE) of point estimates for model (1.2), under* 2% *contamination from* $U(10, 20)$*, over 200 repetitions with* $n = 1000$

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.460(0.007)	**-0.024(0.003)**	0.255(0.065)	-0.240(0.059)
	$\mu : 1.5$	**-0.056(0.019)**	0.509(0.284)	1.048(1.119)	-0.313(0.103)
II	$\pi : 0.3$	**-0.014(0.001)**	-0.057(0.004)	**0.032(0.001)**	-0.043(0.002)
	$\mu : 3$	**0.001(0.002)**	0.257(0.081)	0.444(0.204)	-0.034(0.005)
III	$\pi : 0.3$	**-0.019(0.001)**	-0.066(0.005)	**-0.011(0.001)**	-0.035(0.002)
	$\mu : 3$	**-0.001(0.001)**	0.179(0.044)	0.299(0.096)	**-0.011(0.001)**
IV	$\pi : 0.7$	**-0.019(0.001)**	-0.128(0.023)	0.089(0.008)	-0.311(0.149)
	$\mu : 3$	**0.067(0.013)**	0.203(0.257)	1.252(1.628)	-0.829(1.165)
V	$\pi : 0.85$	**-0.019(0.001)**	-0.112(0.018)	0.045(0.002)	-0.347(0.134)
	$\mu : 3$	**0.177(0.067)**	0.574(0.836)	2.275(5.478)	-1.466(2.329)

lists the parameter estimates of the four methods discussed in the article. MPHD estimator shows that among the 3170 genes examined, around 29% genes are differentially expressed between those tumour types, which is close to the 33% from Storey and Tibshirani (2003) and 32.5% from Langaas et al. (2005).

Let

$$\hat{\tau}_0(z_i) = \hat{\pi}\phi_{\hat{\sigma}}(z_i)/[\hat{\pi}\phi_{\hat{\sigma}}(z_i) + (1 - \hat{\pi})\hat{f}(z_i - \hat{\mu})]$$

be the classification probability that the ith gene is not differentially expressed. Then we select all genes with $\hat{\tau}_0(z_i) \leq c$ to be differentially expressed. The threshold c can be selected by controlling the false discovery rate (FDR, Benjamini and Hochberg, 1995). Based on McLachlan et al. (2006), the FDR can be estimated by

$$\widehat{FDR} = \frac{1}{N_r} \sum_i \hat{\tau}_0(z_i) I_{[0,c_0]}(\hat{\tau}_0(z_i)),$$

where $N_r = \sum_i I_{[0,c_0]}(\hat{\tau}_0(z_i))$ is the total number of found differentially expressed genes

Table 1.7: *Bias (MSE) of point estimates for model (1.3) over 200 repetitions with* $n = 100$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.058(0.021)	0.110(0.021)	0.302(0.097)	**-0.047(0.015)**
	$\sigma : 1$	0.052(0.045)	0.758(2.207)	**0.143(0.042)**	**-0.047(0.071)**
	$\mu : 1.5$	-0.057(0.082)	0.098(0.095)	0.463(0.242)	**-0.055(0.061)**
II	$\pi : 0.3$	**-0.008(0.004)**	0.062(0.017)	0.082(0.014)	**-0.006(0.004)**
	$\sigma : 1$	**0.095(0.041)**	1.821(5.180)	0.331(0.252)	**0.012(0.056)**
	$\mu : 3$	**-0.014(0.025)**	-0.341(0.216)	0.081(0.031)	-0.032(0.030)
III	$\pi : 0.3$	-0.051(0.005)	0.024(0.011)	-0.042(0.006)	**-0.009(0.003)**
	$\sigma : 1$	**-0.101(0.030)**	2.258(6.708)	**-0.028(0.105)**	-0.031(0.045)
	$\mu : 3$	**-0.021(0.005)**	-0.436(0.223)	-0.187(0.049)	**-0.008(0.008)**
IV	$\pi : 0.7$	**-0.014(0.011)**	-0.060(0.012)	0.114(0.016)	-0.054(0.018)
	$\sigma : 2$	0.101(0.047)	0.195(0.161)	**0.120(0.034)**	**0.039(0.065)**
	$\mu : 3$	0.100(0.201)	-0.537(0.504)	**0.019(0.175)**	-0.320(0.511)
V	$\pi : 0.85$	**-0.028(0.009)**	-0.076(0.014)	**0.042(0.003)**	-0.159(0.078)
	$\sigma : 2$	0.098(0.043)	0.179(0.100)	**-0.006(0.021)**	-0.118(0.247)
	$\mu : 3$	**0.275(0.432)**	-1.080(1.719)	-0.622(1.088)	-0.845(1.717)

and $I_A(x)$ is the indicator function, which is one if $x \in A$ and is zero otherwise. Table 1.15 reports the number of selected differentially expressed genes (N_r) and the estimated false discovery rate (FDR) for different threshold c values based on MPHD estimate. For comparison, we also include the results of McLachlan and Wockner (2010), which assumes a two-component mixture of heterogeneous normals (MLE) for z_is.

1.6 Summary and Future Work

In this chapter, we proposed a minimum profile Hellinger distance estimator for a class of semiparametric mixture models and investigated its existence, consistency, and asymptotic normality. Simulation study shows that the MPHD estimator outperforms existing estima-

Table 1.8: *Bias (MSE) of point estimates for model (1.3) over 200 repetitions with* $n = 250$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	**-0.043(0.014)**	**0.064(0.006)**	0.302(0.093)	-0.048(0.015)
	$\sigma : 1$	**0.058(0.021)**	-0.101(0.075)	0.157(0.032)	**0.020(0.033)**
	$\mu : 1.5$	**-0.064(0.051)**	0.220(0.059)	0.421(0.186)	**-0.079(0.049)**
II	$\pi : 0.3$	**-0.005(0.001)**	-0.028(0.003)	0.093(0.011)	**-0.002(0.001)**
	$\sigma : 1$	**0.046(0.013)**	0.330(0.912)	0.377(0.191)	**-0.001(0.021)**
	$\mu : 3$	**-0.005(0.010)**	-0.129(0.054)	0.121(0.022)	-0.017(0.011)
III	$\pi : 0.3$	-0.037(0.002)	-0.043(0.004)	0.005(0.002)	**0.002(0.001)**
	$\sigma : 1$	**-0.061(0.013)**	0.609(1.741)	0.163(0.100)	**0.013(0.022)**
	$\mu : 3$	**-0.006(0.001)**	-0.233(0.085)	-0.069(0.009)	**0.001(0.002)**
IV	$\pi : 0.7$	**-0.008(0.003)**	-0.068(0.009)	0.121(0.016)	-0.014(0.007)
	$\sigma : 2$	**0.036(0.023)**	0.023(0.035)	0.142(0.028)	**0.009(0.032)**
	$\mu : 3$	**0.108(0.054)**	-0.437(0.269)	0.153(0.067)	**-0.070(0.140)**
V	$\pi : 0.85$	**-0.014(0.003)**	-0.076(0.010)	0.060(0.004)	-0.076(0.028)
	$\sigma : 2$	0.093(0.027)	0.069(0.035)	**0.046(0.011)**	**0.027(0.048)**
	$\mu : 3$	**0.115(0.205)**	-0.912(1.024)	-0.222(0.266)	-0.573(0.981)

tors when data are under contamination, while performs competitively to other estimators when there is no contamination.

We indicate two fields of application of the model. The first is microarray data analysis, which is the initial motivation of introducing model (1.2) (see Bordes et al., 2006). The second is sequential clustering algorithm, which is the initial motivation of introducing model (1.3) (see Song et al., 2010). A real data application involving sequential clustering algorithm is also provided to illustrate the effectiveness of the proposed methodology.

In this chapter, we only considered the asymptotic results for model (1.2), since its identifiability property has been established by Bordes et al. (2006). When the first component of the general model (1.1) has normal distribution, the empirical studies demonstrated the success of proposed MPHD estimator. We conjecture that the asymptotic results of MPHD

Table 1.9: *Bias (MSE) of point estimates for model (1.3) over 200 repetitions with* $n = 1000$.

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	**-0.019(0.005)**	**0.053(0.004)**	0.301(0.091)	-0.020(0.005)
	$\sigma : 1$	**0.040(0.008)**	-0.147(0.028)	0.177(0.034)	**0.025(0.011)**
	$\mu : 1.5$	**-0.019(0.017)**	0.236(0.059)	0.423(0.181)	-0.024(0.018)
II	$\pi : 0.3$	**-0.001(0.001)**	-0.037(0.002)	0.099(0.010)	**0.000(0.001)**
	$\sigma : 1$	**0.017(0.003)**	-0.044(0.007)	0.407(0.176)	**-0.002(0.005)**
	$\mu : 3$	**0.009(0.002)**	-0.042(0.005)	0.151(0.025)	**0.003(0.002)**
III	$\pi : 0.3$	**-0.029(0.001)**	-0.047(0.003)	**0.011(0.001)**	**0.001(0.001)**
	$\sigma : 1$	-0.051(0.005)	-0.029(0.007)	0.177(0.044)	**0.005(0.004)**
	$\mu : 3$	**-0.003(0.001)**	-0.122(0.017)	-0.031(0.002)	**-0.001(0.001)**
IV	$\pi : 0.7$	**-0.008(0.001)**	-0.069(0.006)	0.125(0.016)	**-0.004(0.001)**
	$\sigma : 2$	**0.002(0.006)**	-0.051(0.013)	0.172(0.032)	**-0.001(0.006)**
	$\mu : 3$	0.058(0.017)	-0.346(0.153)	0.161(0.035)	**-0.018(0.015)**
V	$\pi : 0.85$	**-0.003(0.001)**	-0.067(0.006)	0.072(0.005)	-0.025(0.010)
	$\sigma : 2$	0.053(0.009)	**-0.005(0.008)**	0.087(0.010)	0.008(0.031)
	$\mu : 3$	**0.099(0.042)**	-0.745(0.633)	0.135(0.060)	-0.180(0.293)

also apply to the more general model (1.1) when it is identifiable. However, it requires further research to find sufficient conditions for the identifiability of model (1.1). In addition, more work remains to be done on the application of MPHD estimator to the regression setting such as mixture of regression models.

1.7 Proofs

The proofs of Theorems 1.3.1, 1.3.2, and 1.3.3 are presented in this section.

Proof of Theorem 1.3.1.

The method of proof is similar to that of Theorem 2.1 of Beran (1977).

Table 1.10: *Bias (MSE) of point estimates for model (1.3), under* 2% *contamination from* $U(10, 20)$*, over 200 repetitions with* $n = 100$*.*

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.104(0.025)	**0.102(0.018)**	0.295(0.093)	-0.132(0.031)
	$\sigma : 1$	**0.132(0.090)**	0.680(1.919)	**0.133(0.046)**	-0.213(0.150)
	$\mu : 1.5$	-0.148(0.088)	0.591(0.560)	1.115(1.507)	**-0.137(0.068)**
II	$\pi : 0.3$	**-0.022(0.005)**	0.051(0.016)	0.067(0.011)	-0.062(0.010)
	$\sigma : 1$	**0.081(0.034)**	1.755(5.036)	0.301(0.235)	-0.244(0.121)
	$\mu : 3$	**-0.025(0.036)**	0.053(0.180)	0.467(0.323)	-0.079(0.051)
III	$\pi : 0.3$	**-0.036(0.003)**	**0.019(0.012)**	-0.036(0.005)	-0.046(0.006)
	$\sigma : 1$	**-0.061(0.019)**	2.229(6.635)	**0.025(0.102)**	-0.201(0.076)
	$\mu : 3$	**-0.022(0.004)**	-0.116(0.114)	0.144(0.085)	-0.034(0.009)
IV	$\pi : 0.7$	**-0.033(0.017)**	**-0.066(0.013)**	**0.099(0.013)**	-0.110(0.033)
	$\sigma : 2$	**0.088(0.058)**	0.184(0.147)	**0.104(0.032)**	-0.152(0.110)
	$\mu : 3$	**0.103(0.262)**	0.449(0.928)	1.209(2.263)	-0.226(0.354)
V	$\pi : 0.85$	-0.045(0.023)	-0.084(0.014)	**0.024(0.002)**	-0.198(0.106)
	$\sigma : 2$	0.145(0.082)	0.222(0.135)	**-0.013(0.027)**	-0.172(0.199)
	$\mu : 3$	**0.379(2.637)**	0.646(2.505)	1.235(3.351)	**-0.501(1.258)**

(i) Let $d(t) = \left\| h^{1/2}_{t,f(t,h\boldsymbol{\theta}_{,f})} - h^{1/2}_{\boldsymbol{\theta},f} \right\|$. For any sequence $\{t_n : t_n \in \Theta, t_n \to t \text{ as } n \to \infty\}$,

$$
\begin{aligned}
|d^2(t_n) - d^2(t)| &= |\int (h^{1/2}_{t_n,f(t_n,h\boldsymbol{\theta}_{,f})}(x) - h^{1/2}_{\boldsymbol{\theta},f}(x))^2 dx - \int (h^{1/2}_{t,f(t,h\boldsymbol{\theta}_{,f})}(x) - h^{1/2}_{\boldsymbol{\theta},f}(x))^2 dx| \\
&= 2|\int (h^{1/2}_{t_n,f(t_n,h\boldsymbol{\theta}_{,f})}(x) - h^{1/2}_{t,f(t,h\boldsymbol{\theta}_{,f})}(x)) h^{1/2}_{\boldsymbol{\theta},f}(x) dx| \\
&\leq 2\|h^{1/2}_{t_n,f(t_n,h\boldsymbol{\theta}_{,f})} - h^{1/2}_{t,f(t,h\boldsymbol{\theta}_{,f})}\|
\end{aligned}
$$

Table 1.11: *Bias (MSE) of point estimates for model (1.3), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 250$.*

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.108(0.024)	**0.060(0.006)**	0.292(0.087)	-0.164(0.038)
	$\sigma : 1$	0.103(0.056)	**-0.015(0.184)**	**0.155(0.031)**	-0.216(0.116)
	$\mu : 1.5$	**-0.145(0.070)**	1.697(0.550)	1.085(1.277)	**-0.177(0.067)**
II	$\pi : 0.3$	**-0.011(0.001)**	-0.033(0.003)	0.087(0.009)	-0.049(0.005)
	$\sigma : 1$	**0.056(0.014)**	0.306(0.843)	0.400(0.204)	-0.195(0.062)
	$\mu : 3$	**-0.011(0.012)**	0.245(0.115)	0.525(0.316)	-0.047(0.016)
III	$\pi : 0.3$	**-0.025(0.001)**	-0.073(0.008)	-0.723(0.002)	-0.042(0.003)
	$\sigma : 1$	**-0.057(0.012)**	1.125(3.379)	0.081(0.055)	-0.203(0.056)
	$\mu : 3$	**-0.008(0.001)**	-0.068(0.060)	0.207(0.073)	-0.029(0.004)
IV	$\pi : 0.7$	**-0.024(0.004)**	-0.089(0.012)	0.102(0.011)	-0.077(0.013)
	$\sigma : 2$	**0.010(0.018)**	0.035(0.041)	0.138(0.028)	-0.213(0.078)
	$\mu : 3$	**0.118(0.064)**	0.406(0.435)	1.339(2.125)	**-0.032(0.084)**
V	$\pi : 0.85$	**-0.027(0.006)**	-0.098(0.014)	**0.037(0.002)**	-0.114(0.038)
	$\sigma : 2$	0.052(0.029)	0.069(0.034)	**0.041(0.010)**	-0.193(0.099)
	$\mu : 3$	**0.215(0.228)**	0.715(1.406)	1.963(4.889)	**-0.130(0.460)**

Since $\int h_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x)dx = \int h_{t,f(t,h_{\boldsymbol{\theta},f})}(x)dx = 1$,

$$\|h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})} - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}\|^2 = \int (h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x))^2 dx$$
$$\leq \int (h_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) - h_{t,f(t,h_{\boldsymbol{\theta},f})}(x))dx = 2\int [h_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) - h_{t,f(t,h_{\boldsymbol{\theta},f})}(x)]^+ dx$$

Also, $[h_{t_n,f(t_n,h_{\boldsymbol{\theta},f})}(x) - h_{t,f(t,h_{\boldsymbol{\theta},f})}(x)]^+ \leq h_{t,f(t,h_{\boldsymbol{\theta},f})}(x)$, and for every x, $h_{t,f(t,h_{\boldsymbol{\theta},f})}(x)$ is continuous at t. Thus, by the Dominated Convergence Theorem, $\|h^{1/2}_{t_n,f(t_n,h_{\boldsymbol{\theta},f})} - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}\| \to 0$ as $n \to \infty$. So, $d(t_n) \to d(t)$ as $n \to \infty$, i.e., d is continuous on Θ and achieves a minimum for $t \in \Theta$.

(ii) By assumption, $h_{\boldsymbol{\theta},f}$ is identifiable. Immediately, we have $T(h_{\boldsymbol{\theta},f}) = \boldsymbol{\theta}$ uniquely.

Table 1.12: *Bias (MSE) of point estimates for model (1.3), under* 2% *contamination from* $U(10, 20)$*, over 200 repetitions with* $n = 1000$*.*

Case	TRUE	MPHD	Maximizing π-type	EM-type	Symmetrization
I	$\pi : 0.3$	-0.083(0.015)	**0.049(0.003)**	0.291(0.085)	-0.211(0.051)
	$\sigma : 1$	0.099(0.026)	**-0.128(0.022)**	0.178(0.033)	**-0.096(0.050)**
	$\mu : 1.5$	**-0.116(0.039)**	0.706(0.515)	1.068(1.162)	-0.258(0.085)
II	$\pi : 0.3$	**-0.012(0.001)**	-0.042(0.002)	0.092(0.009)	-0.05(0.003)
	$\sigma : 1$	**0.025(0.003)**	-0.031(0.007)	0.422(0.189)	-0.199(0.045)
	$\mu : 3$	**-0.008(0.002)**	0.299(0.099)	0.537(0.297)	-0.047(0.005)
III	$\pi : 0.3$	**-0.021(0.001)**	-0.053(0.003)	**0.004(0.001)**	-0.042(0.002)
	$\sigma : 1$	**-0.040(0.004)**	**-0.033(0.006)**	0.185(0.050)	-0.194(0.042)
	$\mu : 3$	**-0.004(0.001)**	0.208(0.049)	0.302(0.099)	**-0.02(0.001)**
IV	$\pi : 0.7$	**-0.017(0.001)**	-0.079(0.008)	0.110(0.012)	-0.059(0.004)
	$\sigma : 2$	**-0.019(0.004)**	-0.045(0.013)	0.178(0.034)	-0.187(0.042)
	$\mu : 3$	0.094(0.020)	0.493(0.324)	1.386(2.005)	**0.024(0.012)**
V	$\pi : 0.85$	**-0.019(0.001)**	-0.081(0.008)	0.053(0.003)	-0.070(0.008)
	$\sigma : 2$	**0.013(0.004)**	**-0.008(0.007)**	0.083(0.009)	-0.167(0.034)
	$\mu : 3$	**0.193(0.064)**	0.909(1.093)	2.559(6.866)	**0.038(0.068)**

(iii) Let $d_n(t) = \|h^{1/2}_{t,f(t,h_n)} - h^{1/2}_n\|$ and $d(t) = \|h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})} - h^{1/2}_{\boldsymbol{\theta},f}\|$. By Minkowski's inequality,

$$
\begin{aligned}
|d_n(t) - d(t)| &= |[\int (h^{1/2}_{t,f(t,h_n)}(x) - h^{1/2}_n(x))^2 dx]^{1/2} - [\int (h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x) - h^{1/2}_{\boldsymbol{\theta},f}(x))^2 dx]^{1/2}| \\
&\leq \{\int (h^{1/2}_{t,f(t,h_n)}(x) - h^{1/2}_n(x) - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x) + h^{1/2}_{\boldsymbol{\theta},f}(x))^2 dx\}^{1/2} \\
&\leq \{2\int [h^{1/2}_{t,f(t,h_n)}(x) - h^{1/2}_{t,f(t,h_{\boldsymbol{\theta},f})}(x)]^2 dx + 2\int [h^{1/2}_n(x) - h^{1/2}_{\boldsymbol{\theta},f}(x)]^2 dx\}^{1/2}
\end{aligned}
$$

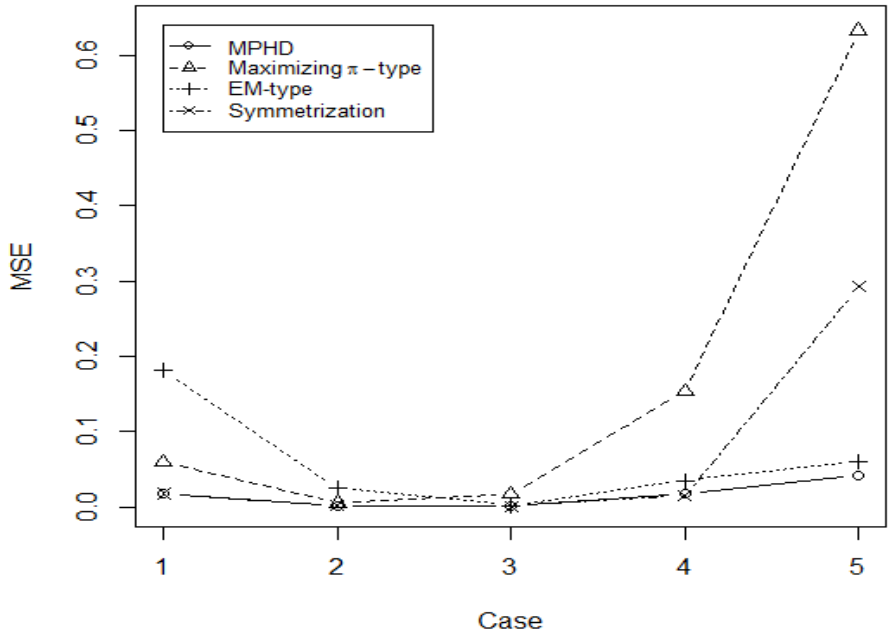

Figure 1.2: *MSE of point estimates of μ of model (1.3) over 200 repetitions with $n = 1000$.*

Table 1.13: *Estimates of first principal component in Iris data.*

Variable	True Value	MPHD	Maximizing π-type	EM-type	Symmetrization
π	0.3000	0.3195	0.3986	0.2896	0.3266
σ	0.2208	0.2457	4.0000	0.1629	0.2055
μ	3.9469	3.9526	2.6240	3.6979	3.9077

Consequently,

$$\sup_{t\in\Theta}|d_n(t)-d(t)| \leq \{2\sup_{t\in\Theta}\int[h^{1/2}_{t,f(t,h_n)}(x)-h^{1/2}_{t,f(t,h_{\boldsymbol{\theta}_{,f}})}(x)]^2dx+2\int[h_n^{1/2}(x)-h^{1/2}_{\boldsymbol{\theta},f}(x)]^2dx\}^{1/2}, \tag{1.11}$$

and the right hand side of (1.11) goes to zero as $n \to \infty$ by assumptions. Write $\boldsymbol{\theta}_0 = T(h_{\boldsymbol{\theta}_{,f}})$ and $\boldsymbol{\theta}_n = T(h_n)$, then we have, as $n \to \infty$, $d_n(\boldsymbol{\theta}_0) \to d(\boldsymbol{\theta}_0)$ and $d_n(\boldsymbol{\theta}_n) \to d(\boldsymbol{\theta}_n)$.

If $\boldsymbol{\theta}_n \nrightarrow \boldsymbol{\theta}_0$, then there exists a subsequence $\{\boldsymbol{\theta}_m\} \subseteq \{\boldsymbol{\theta}_n\}$ such that $\boldsymbol{\theta}_m \to \boldsymbol{\theta}' \neq \boldsymbol{\theta}_0$,

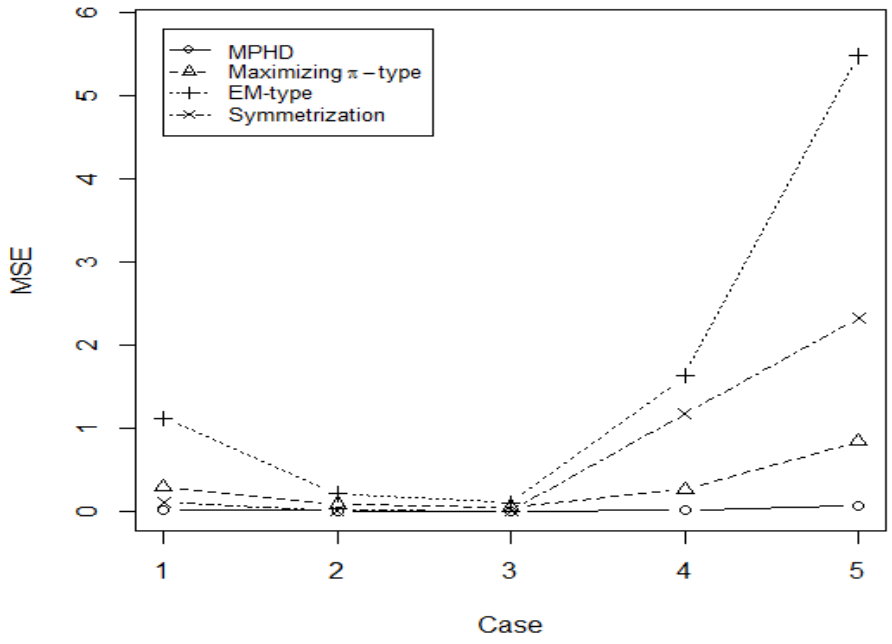

Figure 1.3: *MSE of point estimates of μ of model (1.2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1000$.*

Table 1.14: *Parameter estimates for the Breast Cancer data.*

Variable	MPHD	Maximizing π-type	EM-type	Symmetrization
π	0.7109	0.6456	0.8365	0.5027
σ	1.0272	1	1.1441	1.0773
μ	1.8027	1.6756	1.9366	1.0765

implying that $\boldsymbol{\theta}' \in \Theta$ and $d(\boldsymbol{\theta}_m) \to d(\boldsymbol{\theta}')$ by the continuity of d. From the above result, we have $d_m(\boldsymbol{\theta}_m) - d_m(\boldsymbol{\theta}_0) \to d(\boldsymbol{\theta}') - d(\boldsymbol{\theta}_0)$. By the definition of $\boldsymbol{\theta}_m$, $d_m(\boldsymbol{\theta}_m) - d_m(\boldsymbol{\theta}_0) \leq 0$, and therefore, $d(\boldsymbol{\theta}') - d(\boldsymbol{\theta}_0) \leq 0$. However, by the definition of $\boldsymbol{\theta}_0$ and the uniqueness of it, $d(\boldsymbol{\theta}') > d(\boldsymbol{\theta}_0)$. This is a contradiction, and therefore, $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$. □

Proof of Theorem 1.3.2.

Let H_n denote the empirical cdf of $X_1, X_2, ..., X_n$, which are assumed i.i.d. with density

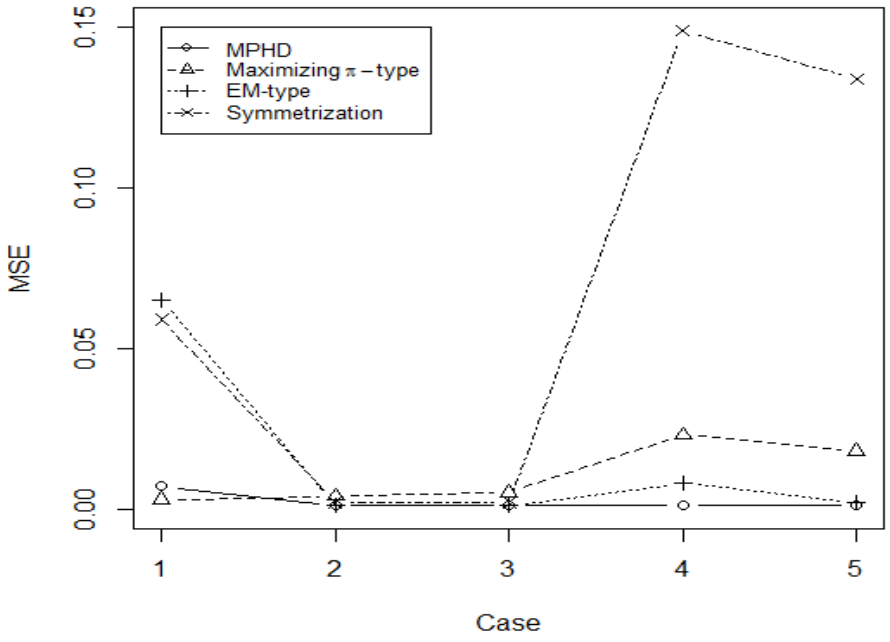

Figure 1.4: *MSE of point estimates of* π *of model (1.2), under* 2% *contamination from* $U(10, 20)$*, over 200 repetitions with* $n = 1000$*.*

$h_{\boldsymbol{\theta},f}$ and cdf H. Let

$$\tilde{h}_n(x) = (c_n s_n)^{-1} \int K((c_n s_n)^{-1}(x - y))dH(y). \tag{1.12}$$

Let $B_n(x) = n^{1/2}[H_n(x) - H(x)]$, then

$$\begin{aligned} \sup_x |\hat{h}_n(x) - \tilde{h}_n(x)| &= \sup_x n^{-1/2}(c_n s_n)^{-1} \int K((c_n s_n)^{-1}(x - y))dB_n(y) \\ &\leq n^{-1/2}(c_n s_n)^{-1} \sup_x B_n(x) \int |K'(x)|dx \xrightarrow{p} 0. \end{aligned} \tag{1.13}$$

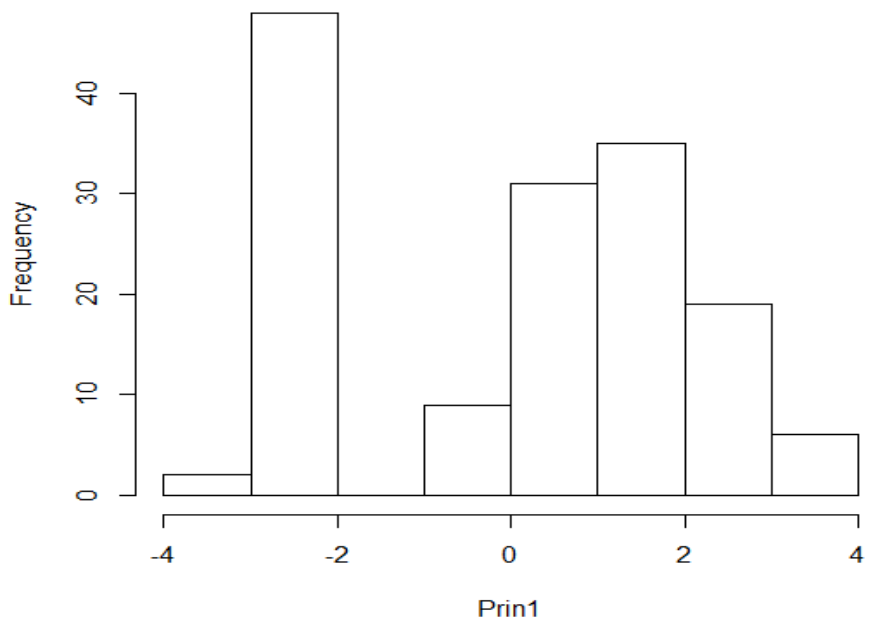

Figure 1.5: *Histogram of the first principal component in the Iris data.*

Suppose $[a, b]$ is an interval that contains the support of $K(\cdot)$, then

$$\begin{aligned}
\sup_x |\tilde{h}_n(x) - h_{\boldsymbol{\theta},f}(x)| &= \sup_x |\int K(t) h_{\boldsymbol{\theta},f}(x - c_n s_n t) dt - h_{\boldsymbol{\theta},f}(x)| \\
&= \sup_x |h_{\boldsymbol{\theta},f}(x - c_n s_n \xi) \int K(t) dt - h_{\boldsymbol{\theta},f}(x)|, \text{ with } \xi \in [a, b] \\
&\leq \sup_x \sup_{t \in [a,b]} |h_{\boldsymbol{\theta},f}(x - c_n s_n t) - h_{\boldsymbol{\theta},f}(x)| \xrightarrow{p} 0 \qquad (1.14)
\end{aligned}$$

From (1.13) and (1.14), we have

$$\sup_x |\hat{h}_n(x) - h_{\boldsymbol{\theta},f}(x)| \xrightarrow{p} 0.$$

From an argument similar to the proof of Theorem 1.3.1, $\|\hat{h}_n^{1/2}(x) - h_{\boldsymbol{\theta},f}^{1/2}(x)\| \xrightarrow{p} 0$ as $n \to \infty$. By Theorem 1.3.1, $T(\hat{h}_n) \xrightarrow{p} T(h_{\boldsymbol{\theta},f})$ as $n \to \infty$. □

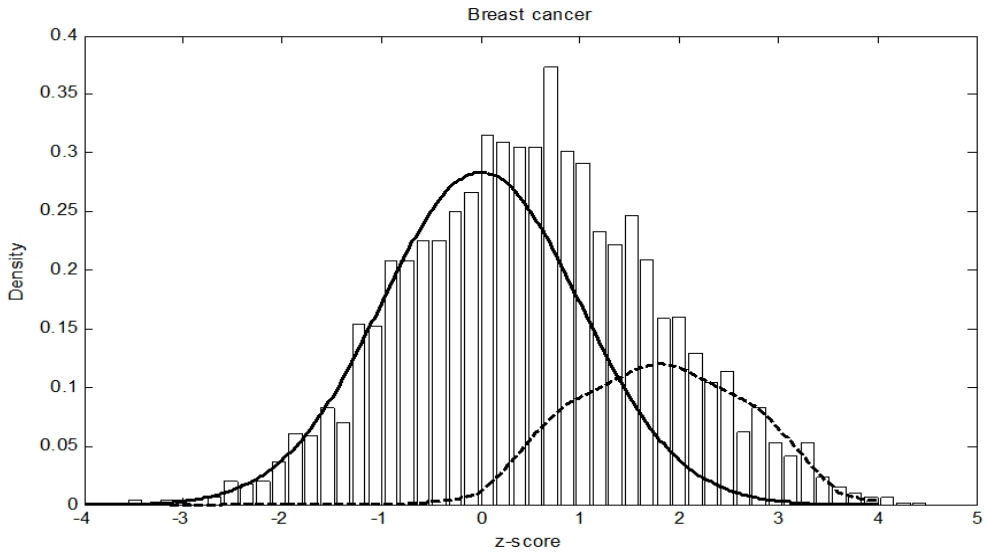

Figure 1.6: *Breast cancer data: plot of fitted two-component mixture model with theoretical* $N(0,1)$ *null and non-null component (weighted respectively by* $\hat{\pi}$ *and* $(1-\hat{\pi})$*) imposed on histogram of z-score.*

Table 1.15: *Estimated FDR for various levels of the threshold c applied to the posterior probability of nondifferentially expression for the breast cancer data.*

	MLE		MPHD	
c	N_r	$\widehat{FDR}$	N_r	$\widehat{FDR}$
0.1	143	0.06	179	0.052
0.2	338	0.11	320	0.093
0.3	539	0.16	477	0.144
0.4	743	0.21	624	0.193
0.5	976	0.27	780	0.244

Proof of Theorem 1.3.3.

Let

$$D(\boldsymbol{\theta}, g) = \int \dot{s}_{\boldsymbol{\theta},g}(x) g^{1/2}(x) dx = < \dot{s}_{\boldsymbol{\theta},g}, g^{1/2} >, \tag{1.15}$$

and it follows that $D(T(h_{\boldsymbol{\theta},f}), h_{\boldsymbol{\theta},f}) = 0$, $D(T(\hat{h}_n), \hat{h}_n) = 0$, and therefore

$$\begin{aligned} 0 &= D(T(\hat{h}_n), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), h_{\boldsymbol{\theta},f}) \\ &= [D(T(\hat{h}_n), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), \hat{h}_n)] + [D(T(h_{\boldsymbol{\theta},f}), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), h_{\boldsymbol{\theta},f})]. \end{aligned}$$

Since the map $\boldsymbol{\theta} \mapsto s_{\boldsymbol{\theta},g}$ satisfies (1.9) and (1.10), $D(\boldsymbol{\theta}, g)$ is differentiable in $\boldsymbol{\theta}$ with derivative

$$\dot{D}(\boldsymbol{\theta}, g) = < \ddot{s}_{\boldsymbol{\theta},g}, g^{1/2} >$$

that is continuous in $\boldsymbol{\theta}$. Then,

$$D(T(\hat{h}_n), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), \hat{h}_n) = (T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f}))\dot{D}(T(h_{\boldsymbol{\theta},f}), \hat{h}_n) + o_p(T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f})).$$

With $\boldsymbol{\theta} = T(h_{\boldsymbol{\theta},f})$,

$$\begin{aligned} &D(T(h_{\boldsymbol{\theta},f}), \hat{h}_n) - D(T(h_{\boldsymbol{\theta},f}), h_{\boldsymbol{\theta},f}) = < \dot{s}_{\boldsymbol{\theta},\hat{h}_n}, \hat{h}_n^{1/2} > - < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} > \\ =&2 < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > + < \dot{s}_{\boldsymbol{\theta},\hat{h}_n} - \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > + < \dot{s}_{\boldsymbol{\theta},\hat{h}_n}, h_{\boldsymbol{\theta},f}^{1/2} > - < \hat{h}_n^{1/2}, \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}} > \\ =&2 < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > + [< \dot{s}_{\boldsymbol{\theta},\hat{h}_n}, h_{\boldsymbol{\theta},f}^{1/2} > - < \hat{h}_n^{1/2}, \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}} >] \\ &+ O(\|\dot{s}_{\boldsymbol{\theta},\hat{h}_n} - \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}\| \cdot \|\hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2}\|) \\ =&2 < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > + o_p(\|\hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2}\|). \end{aligned} \tag{1.16}$$

Applying the algebraic identity

$$b^{1/2} - a^{1/2} = (b-a)/(2a^{1/2}) - (b-a)^2/[2a^{1/2}(b^{1/2} + a^{1/2})^2],$$

we have that

$$
\begin{aligned}
n^{1/2} < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \hat{h}_n^{1/2} - h_{\boldsymbol{\theta},f}^{1/2} > &= n^{1/2} \int \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(x) \frac{\hat{h}_n(x) - h_{\boldsymbol{\theta},f}(x)}{2h_{\boldsymbol{\theta},f}^{1/2}(x)} dx + R_n \\
&= n^{1/2} \int \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(x) \frac{\hat{h}_n(x)}{2h_{\boldsymbol{\theta},f}^{1/2}(x)} dx + R_n \\
&= n^{1/2} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(X_i)}{2h_{\boldsymbol{\theta},f}^{1/2}(X_i)} + o_p(1) + R_n
\end{aligned}
$$

with $|R_n| \leq n^{1/2} \int \frac{|\dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(x)|}{2h_{\boldsymbol{\theta},f}^{3/2}(x)} [\hat{h}_n(x) - h_{\boldsymbol{\theta},f}(x)]^2 dx \xrightarrow{p} 0$. Since $< \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} >$ is assumed to be invertible, then

$$
T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f}) = -\big[< \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} >^{-1} + o_p(1) \big] \frac{1}{n} \sum_{i=1}^{n} \frac{\dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}(X_i)}{h_{\boldsymbol{\theta},f}^{1/2}(X_i)} + o_p(n^{-1/2})
$$

and therefore, the asymptotic distribution of $n^{1/2}(T(\hat{h}_n) - T(h_{\boldsymbol{\theta},f}))$ is $N(0, \Sigma)$ with variance matrix Σ defined by

$$
\Sigma = < \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} >^{-1} < \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, \dot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}^{T} > < \ddot{s}_{\boldsymbol{\theta},h_{\boldsymbol{\theta},f}}, h_{\boldsymbol{\theta},f}^{1/2} >^{-1} .
$$

□

Chapter 2

Mixtures of Nonparametric Regression Models

Abstract

In this chapter, we propose and study a new class of semiparametric mixture of regression models, where the mixing proportions and variances are constants, but the component regression functions are smooth functions of a covariate. A one-step backfitting estimate and two EM-type algorithms have been proposed to achieve the optimal convergence rate for both the global parameters and the nonparametric regression functions. We derive the asymptotic property of the proposed estimates and show that both proposed EM-type algorithms preserve the asymptotic ascent property. A generalized likelihood ratio test is proposed for semiparametric inferences. We prove that the test follows an asymptotic χ^2-distribution under the null hypothesis, which is independent of the nuisance parameters. A simulation study and two real data examples have been conducted to demonstrate the finite sample performance of the proposed model.

2.1 Introduction

Finite mixture of regression models, also known as switching regression models in econometrics, has been widely applied in various fields, see, for example, in econometrics (Wedel and DeSarbo, 1993; Frühwirth-Schnatter, 2001), and in epidemiology (Green and Richardson, 2002). Since Goldfeld and Quandt (1973) first introduced the mixture regression model, many efforts have been made to extend the traditional parametric mixture of linear regression models. For example, Young and Hunter (2010), and Huang and Yao (2012) studied models which allow the mixing proportions to depend on the covariates nonparametrically; Huang et al. (2013) proposed a fully nonparametric mixture of regression models by assuming the mixing proportions, the regression functions, and the variance functions to be nonparametric functions of a covariate; Cao and Yao (2012) suggested a semiparametric mixture of binomial regression models for binary data.

In this article, we propose a new semiparametric mixture of regression models, where the mixing proportions and variances are constants, but the component regression functions are nonparametric functions of a covariate. Compared to traditional finite mixture of linear regression models, the newly proposed model relaxes the parametric assumption on the regression functions, and allows the regression function in each component to be an unknown but smooth function of covariates.

Our new model is motivated by a US house price index data. The data set contains the monthly change of S&P/Case-Shiller House Price Index (HPI) and monthly growth rate of United States Gross Domestic Product (GDP) from January 1990 to December 2002, see Figure 2.3(a) for a scatter plot. Based on the plot, it can be seen that there are two homogeneous groups and the relationship between HPI and GDP are different in different groups. In addition, it can be seen that the relationship in each group is not linear. Therefore, the traditional mixture of linear regression models can not be applied. In Figure 2.3(b), we added the two fitted component regression curves based on our new model, and it is clear that the new model successfully recovered the two component regression curves.

In addition, the observations were classified into two groups corresponding to two different macroeconomic cycles, which possibly explains that the impact of GDP growth rate on HPI change may be different in different macroeconomic cycles.

We will show the identifiability of the proposed model under some regularity conditions. To estimate the unknown smoothing functions, we propose both a regression spline based estimator and a local likelihood estimator using the kernel regression technique. In order to achieve the optimal convergence rate for both the global parameters and the nonparametric functions, we propose a one-step backfitting estimation procedure. The asymptotic properties of the one-step backfitting estimate are investigated. In addition, we propose two EM-type algorithms to compute the proposed estimates and prove their asymptotic ascent properties. A generalized likelihood ratio test is proposed for testing whether the mixing proportions and variances are indeed constants. We investigate the asymptotic behavior of the test and prove that its limiting null distribution follows a χ^2-distribution independent of the nuisance parameters. A simulation study and two real data applications are used to demonstrate the effectiveness of the new model.

The rest of the chapter is organized as follows. In Section 2.2, we introduce the new semiparametric mixture of regression models and the estimation procedure. In particular, we propose a regression spline estimate and a one-step backfitting estimate. A generalized likelihood ratio test is also introduced for some semiparametric inferences. In Section 2.3, we use a Monte Carlo study and two real data examples to demonstrate the finite sample performance of the proposed model and estimates. We conclude the chapter with a brief discussion in Section 2.4 and defer the proofs to Section 2.5.

2.2 Estimation Procedure and Asymptotic Properties

2.2.1 The semiparametric mixture of regression models

Assume $\{(X_i, Y_i), i = 1, ..., n\}$ are a random sample from the population (X, Y). Let Z be a latent variable with $P(Z = j) = \pi_j$ for $j = 1, ..., k$. Suppose $E(Y|X = x, Z = j) = m_j(x)$ and conditioning on $Z = j$ and $X = x$, Y follows a normal distribution with mean $m_j(x)$ and variance σ_j^2. Then, without observing Z, the conditional distribution of Y given $X = x$ can be written as

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

where $\phi(y|\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. In this chapter, we only considered the case when X is univariate. The estimation methodology and theoretical results discussed can be readily extended to multivariate X, but due to the "curse of dimensionality", the extension is less applicable and thus omitted here. Throughout the chapter, we assume that k is fixed, and therefore, refer to (2.1) as a finite semiparametric mixture of regression models, since $m_j(x)$ is a nonparametric function of x, while π_j and σ_j are global parameters. If $m_j(x)$ is indeed linear in x, model (2.1) boils down to a regular finite mixture of linear regression models. When $k = 1$, then model (2.1) is a nonparametric regression model. Therefore, model (2.1) is a natural extension of the finite mixture of linear regression models and the nonparametric regression model.

Huang et al. (2013) studied a nonparametric mixture of regression models (NMR),

$$Y|_{X=x} \sim \sum_{j=1}^{k} \pi_j(x) \phi(Y|m_j(x), \sigma_j^2(x)), \tag{2.2}$$

where $\pi_j(\cdot)$, $m_j(\cdot)$, and $\sigma_j^2(\cdot)$ are unknown but smooth functions. Compared to model (2.2), model (2.1) improves the efficiency of the estimates of π_j, σ_j and $m_j(x)$ by assuming the mixing proportions and variances to be constants, which are also presumed by the

traditional mixture of linear regressions. We will demonstrate such improvement in Section 2.3. However, the new model (2.1) is more challenging to estimate than model (2.2) due to the existence of both global parameters and local parameters. In fact, we will demonstrate later that the model estimate of (2.2) is an intermediate result of the proposed one-step backfitting estimate. In this chapter, we will also develop a generalized likelihood ratio test to compare the proposed model with model (2.2) and illustrate its use in Section 2.3.

Identifiability is a critical issue in many mixture models. Some well known results of identifiability of finite mixture models include: mixture of univariate normals is identifiable (Titterington et al., 1985), and finite mixture of linear regression models is identifiable provided that covariates have a certain level of variability (Hennig, 2000). The following theorem gives a result on the identifiability of model (2.1) and its proof is given in Section 2.5.

Theorem 2.2.1. *Assume that*

(1) $m_j(x)$ *are differentiable functions,* $j = 1, ..., k$.

(2) One of the following conditions holds:

- *(a) For any* $i \neq j$, $\sigma_i \neq \sigma_j$;
- *(b) If there exists* $i \neq j$ *such that* $\sigma_i = \sigma_j$, *then* $\|m_i(x) - m_j(x)\| + \|m_i'(x) - m_j'(x)\| \neq 0$ *for any* x.

(3) The domain $\mathscr{X}$ *of* x *is an interval in* $\mathbb{R}$.

Then, model (2.1) is identifiable.

2.2.2 Estimation procedure and asymptotic properties

Regression spline based estimator

We first introduce a regression spline based estimator, which uses the regression spline (Hastie, et al., 2003; de Boor, 2001) to transfer the semiparametric mixture model to a

parametric mixture model. A cubic spline approximation for $m_j(x)$ can be expressed as

$$m_j(x) \approx \sum_{q=1}^{Q+4} \beta_{jq} B_q(x), j = 1, ..., k, \tag{2.3}$$

where $B_1(x), ..., B_{Q+4}(x)$ is a cubic spline basis and Q is the number of internal knots. Many spline bases can be used here, such as a truncated power spline basis or a B-spline basis. In this chapter, we mainly focus on the B-spline basis.

Based on the approximation (2.3), model (2.1) becomes

$$Y|_{X=x} \sim \sum_{j=1}^{k} \pi_j \phi(Y| \sum_{q=1}^{Q+4} \beta_{jq} B_q(x), \sigma_j^2).$$

The log likelihood of the collected data $\{(X_i, Y_i), i = 1, .., n\}$ is

$$\ell(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j \phi(Y_i| \sum_{q=1}^{Q+4} \beta_{jq} B_q(X_i), \sigma_j^2) \right\},$$

where $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_{k-1}\}^T$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\}^T$, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j,Q+4})^T$, and $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \ldots, \sigma_k^2\}^T$. The parameters $(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ can be estimated by the traditional EM algorithm for mixtures of linear regression models.

The estimation method based on the regression spline approximation is easy to implement, and therefore will be used as an initial value for our other estimation procedures.

One-step backfitting estimation procedure

In this section, we propose a one-step backfitting estimation procedure to achieve the optimal convergence rates for both the global parameters and the nonparametric component regression functions.

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

is,

$$\ell^*(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j \phi(Y_i | m_j(X_i), \sigma_j^2)\}, \tag{2.4}$$

where $\boldsymbol{\pi} = \{\pi_1, ..., \pi_{k-1}\}^T$, $\boldsymbol{m}(\cdot) = \{m_1(\cdot), ..., m_k(\cdot)\}^T$, and $\boldsymbol{\sigma}^2 = \{\sigma_1^2, ..., \sigma_k^2\}^T$. Since $\boldsymbol{m}(\cdot)$ consists of nonparametric functions, (2.4) is not ready for maximization. Next, we propose a one-step backfitting procedure. First, we estimate $\boldsymbol{\pi}$, $\boldsymbol{m}$ and $\boldsymbol{\sigma}^2$ locally by maximizing the following local log-likelihood function:

$$\ell_1(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j \phi(Y_i | m_j, \sigma_j^2)\} K_h(X_i - x), \tag{2.5}$$

where $K_h(t) = h^{-1}K(t/h)$, $K(\cdot)$ is a kernel density function, and h is a tuning parameter.

Let $\tilde{\boldsymbol{\pi}}(x)$, $\tilde{\boldsymbol{m}}(x)$, and $\tilde{\boldsymbol{\sigma}}^2(x)$ be the maximizer of (2.5), which are in fact the model estimates of (2.2) proposed by Huang et al. (2013). Note that, in (2.5), the global parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}^2$ are estimated locally. To improve the efficiency, we propose to update the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}^2$ by maximizing the following log-likelihood function:

$$\ell_2(\boldsymbol{\pi}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j \phi(Y_i | \tilde{m}_j(X_i), \sigma_j^2)\}, \tag{2.6}$$

which, compared to (2.4), replaces $m_j(\cdot)$ by $\tilde{m}_j(\cdot)$.

Denote by $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}^2$ the solution of maximizing (2.6). We can then further improve the estimate of $\boldsymbol{m}(\cdot)$ by maximizing the following local log-likelihood function:

$$\ell_3(\boldsymbol{m}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \hat{\pi}_j \phi(Y_i | m_j, \hat{\sigma}_j^2)\} K_h(X_i - x). \tag{2.7}$$

which, compared to (2.5), replaces π_j and σ_j^2 by $\hat{\pi}_j$ and $\hat{\sigma}_j^2$, respectively.

Let $\hat{\boldsymbol{m}}(x)$ be the solution of (2.7), and we refer to $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{m}}(x)$, and $\hat{\boldsymbol{\sigma}}^2$ as the one-step backfitting estimates. In Section 2.2.2, we show that the one-step backfitting estimates

achieve the optimal convergence rate for both the global parameters, and the nonparametric mean functions. In (2.7), since $\hat{\pi}_j$ and $\hat{\sigma}_j^2$ have root n convergence rate, unlike $\tilde{\boldsymbol{m}}(x)$, $\hat{\boldsymbol{m}}(x)$ does not need to adjust the uncertainty of estimating π_j and σ_j^2. Therefore, $\hat{\boldsymbol{m}}(x)$ can have better estimation accuracy than $\tilde{\boldsymbol{m}}(x)$ proposed by Huang et al. (2013).

Computing algorithms

In this section, we propose a local EM-type algorithm (LEM) and a global EM-type algorithm (GEM) to perform the one-step backfitting.

Local EM-type algorithm (LEM)

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 an EM algorithm separately for different grid points, the labels in the found estimators may change at different grid points, and we may not be able to get smoothed estimated curves (Huang and Yao, 2012). Next, we propose a 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.

Step 1: Modified EM-type algorithm to maximize ℓ_1 in (2.5)

In Step 1, we propose a modified EM-type algorithm to maximize ℓ_1 and obtain the estimates $\tilde{\boldsymbol{\pi}}(\cdot)$, $\tilde{\boldsymbol{m}}(\cdot)$, and $\tilde{\boldsymbol{\sigma}}^2(\cdot)$. At the $(l+1)^{th}$ iteration,

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

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)}(X_i)\phi(Y_i|m_j^{(l)}(X_i), \sigma_j^{2(l)}(X_i))}{\sum_{j=1}^{k} \pi_j^{(l)}(X_i)\phi(Y_i|m_j^{(l)}(X_i), \sigma_j^{2(l)}(X_i))}, i = 1, \ldots, n, j = 1, \ldots, k.$$

M-step: Update the estimates

$$\pi_j^{(l+1)}(x) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}, \tag{2.8}$$

$$m_j^{(l+1)}(x) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} Y_i K_h(X_i - x)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)}, \tag{2.9}$$

$$\sigma_j^{2(l+1)}(x) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - m_j^{(l+1)}(x))^2 K_h(X_i - x)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)}, \tag{2.10}$$

for $x \in \{u_t, t = 1, ..., N\}$. We then update $\pi_j^{(l+1)}(X_i)$, $m_j^{(l+1)}(X_i)$, and $\sigma_j^{2(l+1)}(X_i)$, $i = 1, ..., n$, 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).

Step 2: EM algorithm to maximize ℓ_2 in (2.6)

In Step 2, given $\tilde{m}_j(x)$ from Step 1, a regular EM algorithm can be used to maximize ℓ_2 and update the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}^2$ as $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}^2$. At the $(l+1)^{th}$ iteration,

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

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)} \phi(Y_i | \tilde{m}_j(X_i), \sigma_j^{2(l)})}{\sum_{j=1}^k \pi_j^{(l)} \phi(Y_i | \tilde{m}_j(X_i), \sigma_j^{2(l)})}, i = 1, \ldots, n, j = 1, \ldots, k.$$

M-step: Update the estimates

$$\pi_j^{(l+1)} = \frac{\sum_{i=1}^n p_{ij}^{(l+1)}}{n},$$

$$\sigma_j^{2(l+1)} = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - \tilde{m}_j(X_i))^2}{\sum_{i=1}^n p_{ij}^{(l+1)}}.$$

The ascent property of the above algorithm follows from the theory of the ordinary EM algorithm.

Step 3: Modified EM-type algorithm to maximize ℓ_3 in (2.7)

In Step 3, given $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}^2$ from Step 2, we would then maximize ℓ_3 to find the estimates $\hat{\boldsymbol{m}}(x)$. At the $(l+1)^{th}$ iteration,

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

$$p_{ij}^{(l+1)} = \frac{\hat{\pi}_j \phi(Y_i | m_j^{(l)}(X_i), \hat{\sigma}_j^2)}{\sum_{j=1}^k \hat{\pi}_j \phi(Y_i | m_j^{(l)}(X_i), \hat{\sigma}_j^2)}, i = 1, \ldots, n, j = 1, \ldots, k. \tag{2.11}$$

M-step: Update the estimate

$$m_j^{(l+1)}(x) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} Y_i K_h(X_i - x)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)},$$

for $x \in \{u_t, t = 1, ..., N\}$. Similar to Step 1, we update the estimates at a set of grid points first, and then update $m_j^{(l+1)}(X_i)$, $i = 1, ..., n$, by linear interpolating $m_j^{(l+1)}(u_t)$, $t = 1, ..., N$.

Global EM-type algorithm (GEM)

To improve the estimation efficiency, one might further iterate Step 1 to Step 3 until convergence. Next, we propose a global EM-type algorithm (GEM) to approximate such iteration, but with much less computation. At the $(l+1)^{th}$ iteration,

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

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)} \phi(Y_i | m_j^{(l)}(X_i), \sigma_j^{2(l)})}{\sum_{j=1}^k \pi_j^{(l)} \phi(Y_i | m_j^{(l)}(X_i), \sigma_j^{2(l)})}, i = 1, \ldots, n, j = 1, \ldots, k.$$

M-step: Simultaneously update the estimates

$$\begin{aligned}
\pi_j^{(l+1)} &= \frac{\sum_{i=1}^n p_{ij}^{(l+1)}}{n}, \\
m_j^{(l+1)}(x) &= \frac{\sum_{i=1}^n p_{ij}^{(l+1)} Y_i K_h(X_i - x)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)}, \\
\sigma_j^{2(l+1)} &= \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - m_j^{(l+1)}(X_i))^2}{\sum_{i=1}^n p_{ij}^{(l+1)}},
\end{aligned}$$

for $x \in \{u_t, j = 1, ..., N\}$. We then update $m_j^{(l+1)}(X_i)$, $i = 1, ..., n$ by linear interpolating $m_j^{(l+1)}(u_t)$, $t = 1, ..., N$.

Asymptotic properties

Next, we investigate the asymptotic properties of the proposed one-step backfitting estimates and the asymptotic ascent properties of the two proposed EM type algorithms.

Let $\boldsymbol{\theta} = (\boldsymbol{m}^T, \boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T)^T$, then $\boldsymbol{\theta} = (\boldsymbol{m}^T, \boldsymbol{\beta}^T)^T$. Define

$$\ell(\boldsymbol{\theta}, y) = \log \sum_{j=1}^{k} \pi_j \phi(y|m_j, \sigma_j^2), \tag{2.12}$$

and let

$$I_\theta(x) = -E[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}|X = x], I_\beta(x) = -E[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}|X = x], I_m(x) = -E[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial \boldsymbol{m} \partial \boldsymbol{m}^T}|X = x],$$
$$I_{\beta m}(x) = -E[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial \boldsymbol{\beta} \partial \boldsymbol{m}^T}|X = x], \Lambda(u|x) = E[\frac{\partial \ell(\boldsymbol{\theta}(x), y)}{\partial \boldsymbol{m}}|X = u].$$

Define

$$\kappa_l = \int t^l K(t) dt, \qquad \nu_l = \int t^l K^2(t) dt.$$

Under further conditions defined in Section 2.5, the consistency and asymptotic normality of $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}^2$ are established in the next theorem.

Theorem 2.2.2. *Suppose that conditions (C1) and (C3)—(C10) in Section 2.5 are satisfied, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(0, B^{-1} \Sigma B^{-1}),$$

where $B = E\{I_\beta(X)\}$, $\Sigma = Var\{\partial \ell(\boldsymbol{\theta}(X), Y)/\partial \boldsymbol{\beta} - \varpi(X, Y)\}$, $\varpi(x, y) = I_{\beta m} \varphi(x, y)$, *and* $\varphi(x, y)$ *is a* $k \times 1$ *vector consisting of the first* k *elements of* $I_\theta^{-1}(x) \partial \ell(\boldsymbol{\theta}(x), y)/\partial \boldsymbol{\theta}$.

Based on the above theorem, we can see that the proposed one-step backfitting estimator of the global parameters have achieved the optimal root n convergence rate.

The next theorem gives the asymptotic property of $\hat{\boldsymbol{m}}(\cdot)$.

Theorem 2.2.3. *Suppose that conditions (C2)—(C10) in Section 2.5 are satisfied, then*

$$\sqrt{nh}(\hat{\boldsymbol{m}}(x) - \boldsymbol{m}(x) - \Delta_m(x) + o_p(h^2)) \xrightarrow{D} N(0, f^{-1}(x) I_m^{-1}(x) \nu_0),$$

where $f(\cdot)$ is the density of X, $\Delta_m(x)$ is a $k \times 1$ vector consisting of the first k elements of $\Delta(x)$ with

$$\Delta(x) = I_m^{-1}(x)\{\frac{1}{2}\Lambda''(x|x) + f^{-1}(x) f'(x) \Lambda'(x|x)\}\kappa_2 h^2.$$

Based on the above theorem, we can see that $\hat{\boldsymbol{m}}(x)$ has the same asymptotic properties as if $\boldsymbol{\beta}$ were known, since $\hat{\boldsymbol{\beta}}$ has faster convergence rate than $\hat{\boldsymbol{m}}(x)$.

The asymptotic ascent properties of proposed EM-type algorithms are provided in the following theorem.

Theorem 2.2.4. *(i) For the modified EM-type algorithm (Step 1) to maximize ℓ_1, given condition (C2),*

$$\liminf_{n\to\infty} n^{-1}[\ell_1(\boldsymbol{\theta}^{(l+1)}(x)) - \ell_1(\boldsymbol{\theta}^{(l)}(x))] \geq 0$$

in probability, for any given point $x \in \mathscr{X}$, where $\ell_1(\cdot)$ is defined in (2.5).

(ii) For the modified EM-type algorithm (Step 3) to maximize ℓ_3, given condition (C2),

$$\liminf_{n\to\infty} n^{-1}[\ell_3(\boldsymbol{m}^{(l+1)}(x)) - \ell_3(\boldsymbol{m}^{(l)}(x))] \geq 0$$

in probability, for any given point $x \in \mathscr{X}$, where $\ell_3(\cdot)$ is defined in (2.7).

(iii) For the GEM algorithm, we have

$$\liminf_{n\to\infty} n^{-1}[\ell^*(\boldsymbol{m}^{(l+1)}(\cdot), \boldsymbol{\pi}^{(l+1)}, \boldsymbol{\sigma}^{2(l+1)}) - \ell^*(\boldsymbol{m}^{(l)}(\cdot), \boldsymbol{\pi}^{(l)}, \boldsymbol{\sigma}^{2(l)})] \geq 0$$

in probability, for any given point $x \in \mathscr{X}$, where $\ell^*(\cdot)$ is defined in (2.4).

2.2.3 Hypothesis testing

Huang et al. (2013) proposed a nonparametric mixture of regression models where mixing proportions, means, and variances are all unknown but smooth functions of a covariate. Compared to Huang et al. (2013), our model can be more efficient by assuming the mixing proportions and variances to be constants. Then, a natural question to ask is whether or not the mixing proportions and variances indeed depend on the covariate. This amounts to testing the following hypothesis:

$$H_0 : \pi_j(x) \equiv \pi_j, j = 1, ..., k-1;$$
$$\sigma_j^2(x) \equiv \sigma_j^2, j = 1, ..., k.$$

Next, we propose to use the idea of the generalized likelihood ratio test (Fan et al., 2001) to compare model (2.1) with model (2.2).

Let $\ell_n(H_0)$ and $\ell_n(H_1)$ be the log-likelihood functions computed under the null and alternative hypothesis, respectively. Then, we can construct a likelihood ratio test statistic

$$T = \ell_n(H_1) - \ell_n(H_0). \tag{2.13}$$

Note that this likelihood ratio statistic is different from the parametric likelihood ratio statistics, since the null and alternative are both semiparametric models, and the number of parameters under H_0 or H_1 are undefined. The following theorem establishes the Wilks types of results for (2.13), that is, the asymptotic null distribution is independent of the nuisance parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$, and the nuisance nonparametric mean functions $\boldsymbol{m}(x)$.

Theorem 2.2.5. *Suppose that conditions (C9)-(C13) in Section 2.5 hold and that* $nh^{9/2} \to 0$

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

$$r_K T \overset{a}{\sim} \chi^2_\delta,$$

where $r_K = [K(0) - 0.5 \int K^2(t)dt] / \int [K(t) - 0.5K * K(t)]^2 dt$, $\delta = r_K(2k-1)|\mathscr{X}|[K(0) - 0.5 \int K^2(t)dt]/h$, $|\mathscr{X}|$ *denotes the length of the support of* X, *and* $K*K$ *is the* 2^{nd} *convolution of* $K(\cdot)$.

Theorem 2.2.5 unveils a new Wilks type of phenomenon, and provides a simple and useful method for semiparametric inferences. We will demonstrate its application in Section 2.3.

2.3 Examples

2.3.1 Simulation study

In this section, we use a simulation study to investigate the finite sample performance of the proposed regression spline estimate (Spline), the one-step backfitting estimate using local EM-type algorithm (LEM), and the global EM-type algorithm (GEM), and compare them with the traditional mixture of linear regressionss estimate (MLR), and the nonparametric mixture of regression models (NMR, Huang et al., 2013). For the regression spline, we use $Q = 5$, where Q is the number of internal knots. For LEM, GEM and NMR, we use both the true value and the regression spline estimate as initial values, denoted by (T) and (S), respectively.

We conduct a simulation study for a two-component semiparametric mixture of regression models:

$$\pi_1 = 0.5 \text{ or } \pi_1 = 0.7,$$
$$m_1(x) = 4 - \sin(2\pi x) \text{ and } m_2(x) = 1.5 + \cos(3\pi x),$$
$$\sigma_1^2 = 0.09 \text{ and } \sigma_2^2 = 0.16.$$

The covariate X is generated from the one-dimensional uniform distribution in $[0, 1]$, and the Gaussian kernel is used in the simulation. The sample sizes $n = 200$ and $n = 400$ are conducted over 500 repetitions.

The performance of the estimates of the mean functions $\boldsymbol{m}(x)$ is measured by the square root of the average squared errors (RASE),

$$\text{RASE}_m^2 = N^{-1} \sum_{j=1}^{2} \sum_{t=1}^{N} [\hat{m}_j(u_t) - m_j(u_t)]^2,$$

where $\{u_t, t = 1, ..., N\}$ are a set of grid points at which the unknown functions are evaluated. In our simulation, we set $N = 100$. In order to compare between model (2.1) and the nonparametric mixture of regression models proposed by Huang et al. (2013), we also report the RASE of π and σ^2, denoted by RASE_π and RASE_{σ^2}, respectively.

Bandwidth plays an important role in the estimation of $\boldsymbol{m}(\cdot)$. There are ways to calculate the 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 $\{\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{m}}(\cdot), \hat{\boldsymbol{\sigma}}^2\}$, then consider a likelihood version CV, which is defined by

$$CV(h) = \sum_{l=1}^{L} \sum_{t \in \mathscr{T}_l} \log\{\sum_{j=1}^{k} \hat{\pi}_j \phi(y_t | \hat{m}_j(x_t), \hat{\sigma}_j^2)\}.$$

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

Table 2.1 and Table 2.2 report the average of RASE_π, RASE_m, and RASE_{σ^2}, for $\pi_1 = 0.5$ and $\pi_1 = 0.7$, respectively. All the values are multiplied by 100. From Table 2.1 and Table 2.2, we can see that LEM, GEM, and the regression spline estimates give better results

than the mixture of linear regressions estimate. Compared to NMR, model (2.1) improves the efficiency of the estimation of mixing proportions and variances, and provides slightly better estimates for the mean functions. In addition, both LEM and GEM provide better results for the mean functions than the regression spline estimate when the sample size is small. We further notice that LEM(S) and GEM(S) provide similar results to LEM(T) and GEM(T). Therefore, the spline estimate provides good initial values for other estimates.

Table 2.1: *The average of $RASE_\pi$, $RASE_{\sigma^2}$ & $RASE_m$ when $\pi_1 = 0.5$ (true values times 100)*

n	h		LEM(T)	GEM(T)	LEM(S)	GEM(S)	Spline	MLR	NMR(T)	NMR(S)
200	US	$RASE_\pi$	2.82	2.84	2.83	2.84	2.85	4.40	13.38	13.37
		$RASE_{\sigma^2}$	4.34	4.39	4.35	4.40	2.72	65.62	9.88	9.94
		$RASE_m$	20.81	20.84	20.98	21.02	39.98	87.32	20.48	21.35
	AS	$RASE_\pi$	2.83	2.81	2.84	2.81	2.83	4.37	9.55	9.53
		$RASE_{\sigma^2}$	2.69	2.73	2.70	2.73	2.78	63.29	12.62	12.66
		$RASE_m$	17.72	17.67	17.73	17.67	45.60	87.13	18.77	19.52
	OS	$RASE_\pi$	2.79	2.69	2.78	2.69	2.76	4.57	8.39	8.38
		$RASE_{\sigma^2}$	2.73	2.42	2.73	2.42	2.74	64.52	20.77	20.81
		$RASE_m$	23.12	22.99	23.14	22.99	32.33	87.48	25.30	25.39
400	US	$RASE_\pi$	2.02	2.00	2.03	2.00	1.98	3.39	10.56	10.54
		$RASE_{\sigma^2}$	2.88	2.91	2.89	2.91	1.80	66.15	7.99	7.98
		$RASE_m$	15.76	15.78	15.77	15.78	15.10	85.88	15.82	15.85
	AS	$RASE_\pi$	2.03	2.02	2.04	2.02	2.03	3.41	7.35	7.35
		$RASE_{\sigma^2}$	1.87	1.88	1.87	1.88	1.77	65.54	9.87	9.89
		$RASE_m$	13.20	13.19	13.20	13.19	17.65	85.77	14.11	14.15
	OS	$RASE_\pi$	2.19	2.15	2.20	2.15	2.14	3.38	6.54	6.54
		$RASE_{\sigma^2}$	1.92	1.76	1.92	1.76	1.85	65.46	16.21	16.22
		$RASE_m$	16.86	16.78	16.86	16.78	15.85	85.73	18.56	18.56

Next, we test the accuracy of the standard error estimation and the confidence inter-

Table 2.2: *The average of $RASE_\pi$, $RASE_{\sigma^2}$ & $RASE_m$ when $\pi_1 = 0.7$ (true values times 100)*

n	h		LEM(T)	GEM(T)	LEM(S)	GEM(S)	Spline	MLR	NMR(T)	NMR(S)
200	US	$RASE_\pi$	2.66	2.68	2.66	2.68	2.66	4.07	11.54	11.53
		$RASE_{\sigma^2}$	5.45	5.58	5.48	5.58	3.50	62.56	11.25	11.33
		$RASE_m$	23.57	23.63	23.75	24.43	48.12	90.04	23.09	23.49
	AS	$RASE_\pi$	2.56	2.54	2.55	2.54	2.58	4.21	8.35	8.36
		$RASE_{\sigma^2}$	3.27	3.35	3.29	3.35	3.84	64.35	14.40	14.53
		$RASE_m$	20.10	20.09	20.11	20.09	47.52	90.16	21.38	21.41
	OS	$RASE_\pi$	2.74	2.64	2.88	2.77	2.73	4.18	7.30	7.42
		$RASE_{\sigma^2}$	3.10	2.81	3.13	2.83	3.60	64.13	22.09	22.19
		$RASE_m$	26.18	25.99	27.02	26.80	48.17	90.15	28.82	29.74
400	US	$RASE_\pi$	1.79	1.80	1.80	1.80	1.78	3.16	9.24	9.24
		$RASE_{\sigma^2}$	3.74	3.81	3.74	3.81	2.16	66.98	9.23	9.24
		$RASE_m$	18.00	18.03	18.00	18.03	18.91	87.49	17.93	17.99
	AS	$RASE_\pi$	1.87	1.86	1.87	1.86	1.89	3.20	6.45	6.45
		$RASE_{\sigma^2}$	2.26	2.27	2.26	2.27	2.14	65.31	11.29	11.32
		$RASE_m$	14.92	14.90	14.92	14.90	19.67	87.57	16.02	16.00
	OS	$RASE_\pi$	1.94	1.89	1.94	1.89	1.87	2.95	5.59	5.59
		$RASE_{\sigma^2}$	2.27	2.09	2.27	2.09	2.21	65.44	18.02	18.03
		$RASE_m$	19.48	19.41	19.48	19.41	19.79	87.12	21.59	21.63

val construction for π_1, σ_1 and σ_2 via a conditional bootstrap procedure. Given the covariate $X = x$, the response Y^* can be generated from the estimated distribution $\sum_{j=1}^{k} \hat{\pi}_j \phi(Y|\hat{m}_j(x), \hat{\sigma}_j^2)$. For the simplicity of presentation, we only report the results for GEM(T). We apply the proposed estimation procedure to each of the 200 bootstrap samples, and further obtain the confidence intervals.

Table 2.3 reports the results from the bootstrap procedure. SD contains the standard deviation of 500 replicates, and can be considered as true standard errors. SE and STD

contain the mean and standard deviation of the 500 estimated standard errors based on the conditional bootstrap procedure. In addition, the coverage probability of the 95% confidence intervals based on the estimated standard errors are also reported. From Table 2.3 we can see that the bootstrap procedure estimates the true standard error quite well, since all the differences between the true value and the estimates are less than two standard errors of the estimates. The coverage probabilities are satisfactory for π_1, but a bit low for σ_1 and σ_2, especially for over-smoothing bandwidth.

Table 2.3: *Standard errors and coverage probabilities*

	h	SD	SE(STD)	95%	SD	SE(STD)	95%	SD	SE(STD)	95%
		π_1			σ_1			σ_2		
$n = 200$	US	0.037	0.036(0.002)	94.11	0.014	0.013(0.003)	88.82	0.024	0.023(0.004)	91.09
$(0.5, 0.5)$	AS	0.037	0.036(0.002)	93.40	0.014	0.013(0.002)	94.00	0.029	0.022(0.004)	91.20
	OS	0.038	0.035(0.002)	90.60	0.014	0.015(0.002)	96.20	0.022	0.025(0.004)	97.20
$n = 400$	US	0.027	0.025(0.001)	94.40	0.010	0.009(0.001)	94.80	0.018	0.017(0.003)	96.20
$(0.5, 0.5)$	AS	0.026	0.025(0.001)	93.80	0.009	0.009(0.001)	94.00	0.016	0.016(0.002)	96.40
	OS	0.026	0.025(0.001)	93.20	0.009	0.010(0.001)	93.80	0.016	0.018(0.002)	94.80
$n = 200$	US	0.031	0.032(0.002)	94.80	0.011	0.011(0.002)	90.20	0.035	0.029(0.009)	83.20
$(0.7, 0.3)$	AS	0.033	0.032(0.002)	94.60	0.011	0.011(0.001)	96.40	0.028	0.027(0.006)	85.60
	OS	0.033	0.032(0.002)	93.20	0.013	0.013(0.002)	89.60	0.032	0.033(0.008)	97.00
$n = 400$	US	0.023	0.023(0.001)	94.80	0.008	0.008(0.001)	94.20	0.023	0.023(0.004)	92.20
$(0.7, 0.3)$	AS	0.025	0.023(0.001)	93.40	0.008	0.008(0.001)	95.00	0.021	0.021(0.003)	93.40
	OS	0.023	0.023(0.001)	94.60	0.009	0.009(0.001)	83.20	0.021	0.023(0.004)	96.20

We also apply the bootstrap procedure to investigate the point-wise coverage probability of the mean functions, at a set of evenly distributed grid points. Table 2.4 shows the results of the 95% confidence interval for the two component mean functions. From the table, we can see that the mean function of the first component tends to have higher coverage probability than the second component, especially for over-smoothing bandwidth. In addition, the coverage probability is generally lower than the nominal level for over-smoothing bandwidth.

Next, we assess the performance of the testing procedure proposed in Section 2.2.3. Under the null hypothesis, the mixing proportion π_1 and variances σ_1^2 and σ_2^2 are constants. We compute the distribution of T with $n = 200$ and $n = 400$ via 500 repetitions, and compare it with the χ^2-approximation. The histogram of the null distribution is shown in Figure 2.1, where the solid line corresponds to a density of the χ^2-distribution with degrees of freedom δ defined in Theorem 2.2.5. Figure 2.2 shows the Q-Q plot for the two cases. From Figure 2.1 and 2.2, the finite sample null distribution is quite close to a χ^2-distribution with degrees of freedom δ, especially for the case of $n = 400$.

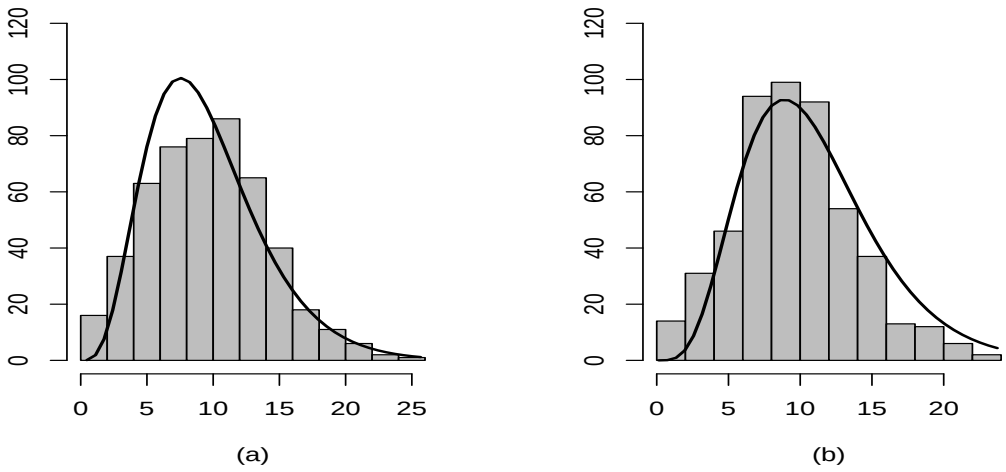

Figure 2.1: *Histogram of T_n and χ^2-approximation of T_n: (a) $n = 200$, (b) $n = 400$.*

2.3.2 Real data applications

Example 1 (The US house price index data). In this section, we illustrate the proposed methodologies with an empirical analysis of US house price index data that are introduced in Section 2.1. GDP is a well known measure of the size of a nation's economy, as it recognizes the total goods and services produced within a nation in a given period, and HPI is known as a measure of a nation's average housing price in repeat sales. It is believed that

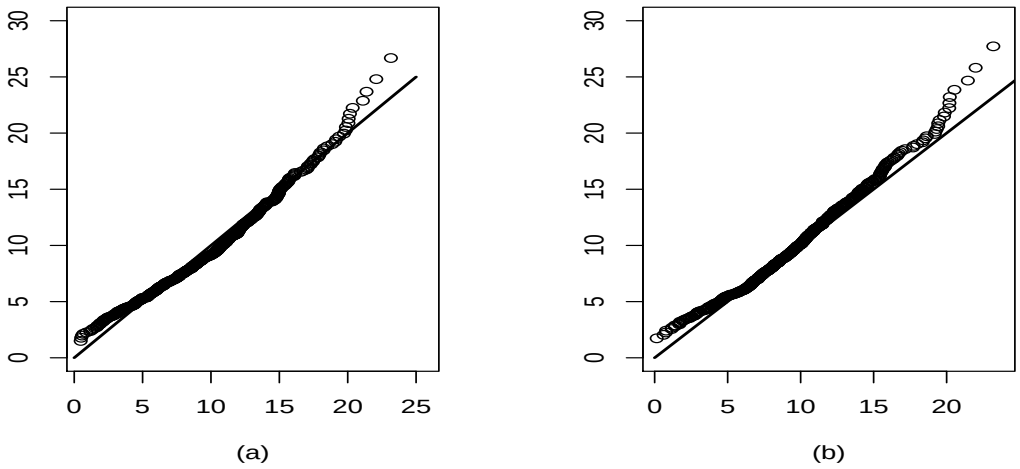

Figure 2.2: *Q-Q plot: (a)* $n = 200$*, (b)* $n = 400$*.*

the housing price and GDP are correlated, and so it is of interest to study how GDP growth rate helps to predict HPI change.

First, a two-component mixture of nonparametric regression models is fitted to the data. Figure 2.3(b) contains the estimated mean functions and their 95% point-wise confidence intervals through the conditional bootstrap procedure, and the 95% confidence interval for π_1, σ_1 and σ_2 are $(0.343, 0.522)$, $(0.090, 0.147)$ and $(0.061, 0.093)$, respectively. Figure 2.3(b) also reports the hard-clustering results, denoted by dots and squares, respectively, for the two components. The hard-clustering results are obtained by maximizing classification probabilities $\{p_{i1}, p_{i2}\}$ for all $i = 1, \ldots, n$. It can be checked that the dots in the lower cluster are mainly from Jan 1990 to Sep 1997, while the squares in the upper cluster are mainly from Oct 1997 to Dec 2002, when the economy experienced an internet boom and bust. In addition, it can be seen that in the first cycle of lower component, GDP growth has an overall positive impact on HPI change. However, in the second cycle of the upper component, GDP growth has a negative impact on HPI change, if GDP growth is smaller than 0.3; when GDP growth is larger than 0.3, it then has a similar positive impact on HPI change as the first cycle.

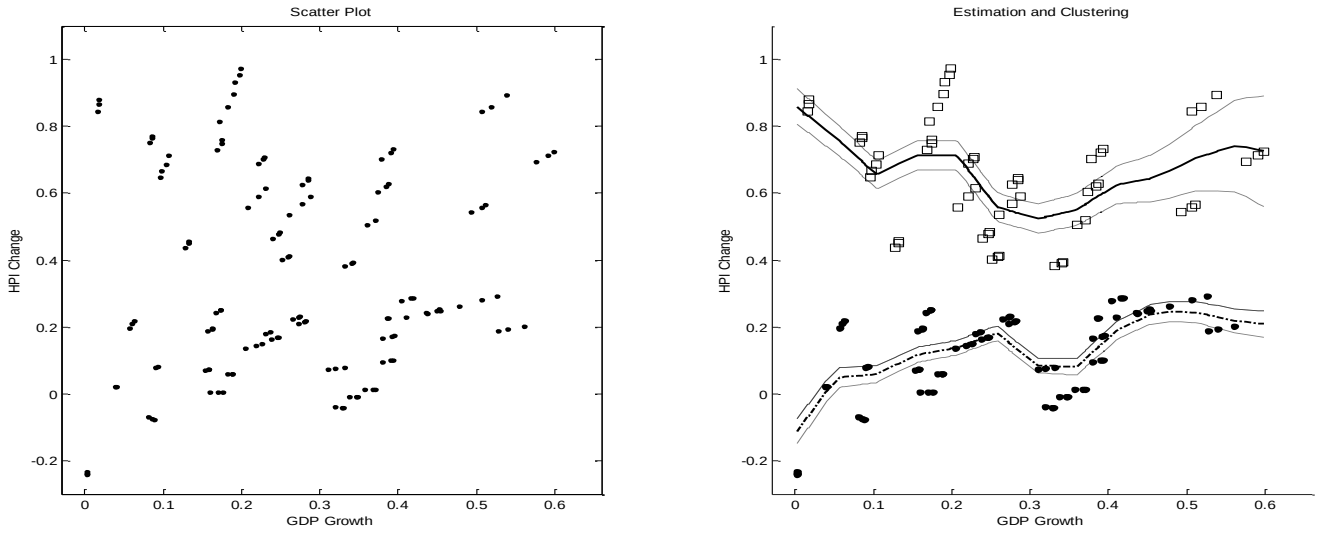

Figure 2.3: *(a) Scatterplot of US house price index data; (b) Estimated mean functions with 95% confidence intervals and a clustering result.*

To examine whether the mixing proportions and variances are indeed constant, we apply the generalized likelihood ratio test developed in Section 2.2.3. The p-value is 0.089, and shows that model (2.1) is more appropriate for the data. To evaluate the prediction performance of the proposed model and compared it to the NMR model proposed by Huang et al. (2013), we use 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 $n - s$) and test subsets (with size s). Table 2.5 reports the average of the mean squared prediction error (MSPE) evaluated at the testing sets, and shows that the prediction performance of model (2.1) is slightly better than that of the NMR model (Huang et al., 2013).

Example 2 (NO data). This data set gives the equivalence ratio, a measure of the richness of the air-ethanol mix in an engine, and peak nitrogen oxide emissions in a study using pure ethanol as a spark-ignition engine fuel. A two-component mixture of nonparametric regression models is fitted to the data, and Figure 2.4(b) contains the estimated mean functions and their 95% point-wise confidence intervals through the bootstrap procedure.

The p-value of the generalized likelihood ratio test is 0.219, indicating that model (2.1) is the preferred model. The result of cross validation in Table 2.5 also shows that the new model predicts the data better than the NMR model.

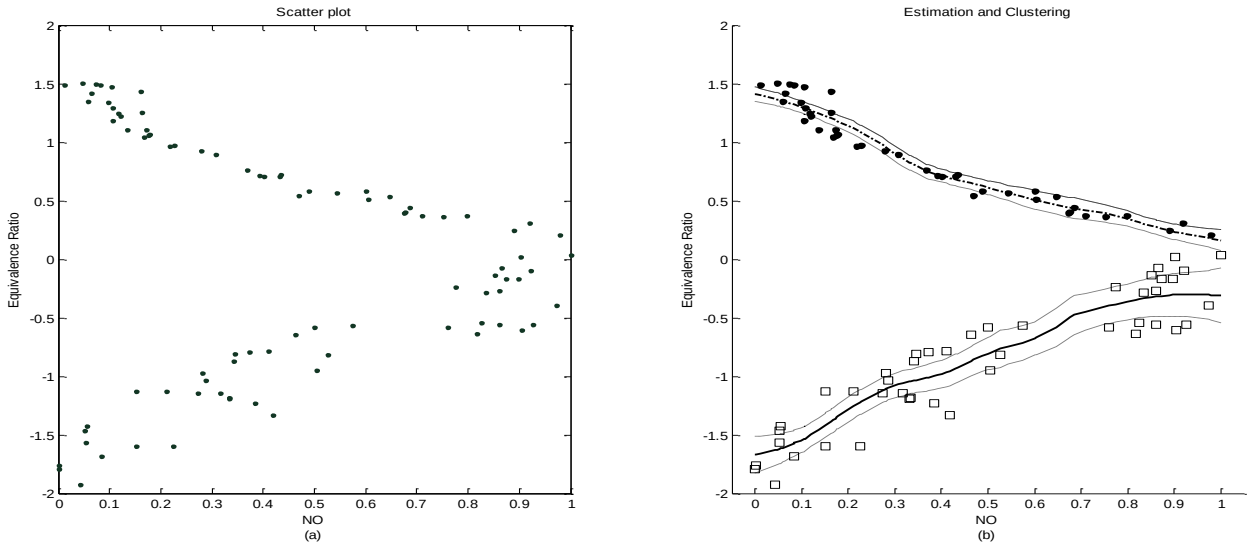

Figure 2.4: *(a) Scatterplot of NO data; (b) Estimated mean functions with 95% confidence intervals and a clustering result.*

2.4 Summary and Future Work

Motivated by a US house index data, in this chapter, we proposed a new class of semiparametric mixture of regression models, where mixing proportions and variances are constants, but the component regression functions are smooth functions of a covariate. The identifiability of the proposed model is established and a one-step backfitting estimation procedure is proposed to achieve the optimal convergence rate for both the global parameters and the nonparametric regression functions. The proposed regression spline estimate is simple to calculate and can be easily extended to some other semiparametric and nonparametric mixture of regression models (Young and Hunter, 2010; Huang et al., 2013; Huang and Yao, 2012). But it requires more research to derive the asymptotic results for such regression

spline based estimators for mixture models. A generalized likelihood ratio test has been proposed for semiparametric inferences.

When the dimension of the predictors is high, due to the curse of dimensionality, it is unpractical to estimate the component regression functions fully nonparametrically. Therefore, it is our interest to further extend the proposed mixture of nonparametric regression models to some other nonparametric or semiparametric models, such as mixture of partial linear regression models, mixture of additive models, and mixture of varying coefficient partial linear models.

In this chapter, we assume that the number of components is known. However, in some applications, it might be infeasible to assume a known number of components in advance. Therefore, more research are needed to select the number of components for the proposed semiparametric mixture model. It will also be interesting to know whether any selection methods for parametric mixture models can be used for the proposed semiparametric mixture model. For example, for information criteria based methods such as AIC and BIC methods (Leroux, 1992), it is not clear how to define the degree of freedom for a semiparametric mixture model.

2.5 Proofs

In this section, the conditions required by Theorems 2.2.2, 2.2.3, 2.2.4 and 2.2.5 are listed. They are not the weakest sufficient conditions, but could easily facilitate the proofs. The proofs of Theorems 2.2.1, 2.2.2, 2.2.3, 2.2.4 and 2.2.5 are also presented in this section.

Technical Conditions:

(C1) $nh^4 \to 0$ and $nh^2 \log(1/h) \to \infty$ as $n \to \infty$ and $h \to 0$.

(C2) $nh \to \infty$ as $n \to \infty$ and $h \to 0$.

(C3) The sample $\{(X_i, Y_i), i = 1, ..., n\}$ are independently and identically distributed from

$f(x, y)$ with finite sixth moments. The support for x, denoted by $\mathscr{X} \in \mathbb{R}$, is bounded and closed.

(C4) $f(x, y) > 0$ in its support and has continuous first derivative.

(C5) $|\partial^3 \ell(\boldsymbol{\theta}, x, y)/\partial\theta_i\partial\theta_j\partial\theta_k| \leq M_{ijk}(x, y)$, where $E(M_{ijk}(x, y))$ is bounded for all i, j, k and all X, Y.

(C6) The unknown functions $m_j(x)$, $j = 1, ..., k$, have continuous second derivative.

(C7) $\sigma_j^2 > 0$ and $\pi_j > 0$ for $j = 1, ..., k$ and $\sum_{j=1}^k \pi_j = 1$.

(C8) $E(X^{2r}) < \infty$ for some $\epsilon < 1 - r^{-1}$, $n^{2\epsilon-1}h \to \infty$.

(C9) $I_\theta(x)$ and $I_m(x)$ are positive definite.

(C10) The kernel function $K(\cdot)$ is symmetric, continuous with compact support.

(C11) The marginal density $f(x)$ of X is Lipschitz continuous and bounded away from 0. X has a bounded support $\mathscr{X}$.

(C12) $t^3K(t)$ and $t^3K'(t)$ are bounded and $\int t^4K(t)dt < \infty$.

(C13) $E|q_\theta|^4 < \infty$, $E|q_m|^4 < \infty$.

The following lemma is from Titterington el al. (1985) and will be used in the proof of Theorem 2.2.1.

Lemma 1. The finite mixture of normal distributions is identifiable. More precisely, if

$$\sum_{j=1}^{k} \pi_j N(\mu_j, \sigma_j^2) = \sum_{t=1}^{p} \lambda_t N(\nu_t, \delta_t^2),$$

where the parameters satisfy $\pi_j > 0$, $j = 1, ..., k$, $\sigma_1^2 \leq \cdots \leq \sigma_k^2$, and if $\sigma_c^2 = \sigma_d^2$ and $c < d$, then $\mu_c < \mu_d$; similarly $\lambda_t > 0$, $t = 1, ..., p$, $\delta_1^2 \leq \cdots \leq \delta_p^2$, and if $\delta_c^2 = \delta_d^2$ and $c < d$, then $\nu_c < \nu_d$. Then, $k = p$ and $(\pi_j, \mu_j, \sigma_j^2) = (\lambda_j, \nu_j, \sigma_j^2)$, $j = 1, ..., k$.

Proof of Theorem 2.2.1.

It is easy to see that when (1), (2a), and (3) hold, model (2.1) is identifiable. Let

$$T = \{x : m_i(x) = m_j(x), \sigma_i^2 = \sigma_j^2, i \neq j\}.$$

By (1), (2b) and (3), for any $x \in T$, $m_i'(x) \neq m_j'(x)$, then any $x \in T$ must be an isolated point, and therefore, T has no limit point, and contains at most countably infinite points. Assume $x_1, x_2, ...,$ are the points in T, and $x_s < x_{s+1}$, $(x_s, x_{s+1}) \bigcap T = \emptyset$.

Assume that model (2.1) has another representation

$$Y|_{X=x} \sim \sum_{t=1}^{p} \lambda_t \phi(Y_i|\nu_t(x), \delta_t^2).$$

Then, by Lemma 1, model (2.1) is identifiable for $x \notin T$. Thus, $k = p$, and there exists a permutation $\omega_x = \{\omega_x(1), ..., \omega_x(k)\}$ of $\{1, ..., k\}$ such that

$$\lambda_{\omega_x(j)} = \pi_j, \nu_{\omega_x(j)}(x) = m_j(x), \delta_{\omega_x(j)}^2 = \sigma_j^2, j = 1, ..., k. \tag{2.14}$$

Since all $m_j(x)$ are differentiable, and for any point $x_s \in T$, the two curves do not interact in the interval (x_s, x_{s+1}), therefore, the permutation ω_x stay the same on (x_s, x_{s+1}).

Next, consider a small neighborhood $(x_s - \epsilon, x_s + \epsilon)$ for any $x_s \in T$, such that $(x_s - \epsilon, x_s + \epsilon) \subset (x_{s-1}, x_{s+1})$. Since $x_s \in T$, $m_i(x_s) = m_j(x_s), \sigma_i^2 = \sigma_j^2$, for some $i \neq j$, but $m_i'(x_s) \neq m_j'(x_s)$. Equation (2.14) implies the identity of derivatives of the curves on either side of x_s, and so the permutation must stay constant on the neighborhood $(x_s - \epsilon, x_s + \epsilon)$. Thus, there exists a permutation $\omega = \{\omega(1), ..., \omega(k)\}$ of $\{1, ..., k\}$, which is free of x, such that

$$\lambda_{\omega(j)} = \pi_j, \nu_{\omega(j)}(x) = m_j(x), \delta_{\omega(j)}^2 = \sigma_j^2, j = 1, ..., k.$$

Therefore, model (2.1) is identifiable.

The next lemma is from Fan and Huang (2005), and will be used throughout the rest of the proofs.

Lemma 2. Let $\{(X_i, Y_i), i = 1, ..., n\}$ be i.i.d random vectors from (X, Y), where X is a random vector and Y is a scalar random variable. Let f be the joint density of (X, Y), and further assume that $E|Y|^r < \infty$ and $\sup_x \int |y|^r f(x, y)dy < \infty$. Let $K(\cdot)$ be a bounded positive function with bounded support, satisfying a Lipschitz condition. Then,

$$\sup_{x \in \mathscr{X}} \left| \frac{1}{n} \sum_{i=1}^{n} [K_h(X_i - x)Y_i - E\{K_h(X_i - x)Y_i\}] \right| = O_p(\gamma_n \log^{1/2}(1/h)),$$

given $n^{2\epsilon-1}h \to \infty$, for some $\epsilon < 1 - 1/r$, where $\gamma_n = (nh)^{-1/2}$.

In order to prove the asymptotic properties of $\{\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{\sigma}}^2\}$, we first need to study the asymptotic property of $\{\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{m}}, \tilde{\boldsymbol{\sigma}}^2\}$, which is the maximum local log-likelihood estimator of (2.5).

Define

$$\tilde{\pi}_j^* = \sqrt{nh}\{\tilde{\pi}_j - \pi_j\},\ \tilde{m}_j^* = \sqrt{nh}\{\tilde{m}_j - m_j\},\ \tilde{\sigma}_j^{2*} = \sqrt{nh}\{\tilde{\sigma}_j^2 - \sigma_j^2\}.$$

Let $\tilde{\boldsymbol{\pi}}^* = (\tilde{\pi}_1^*, ..., \tilde{\pi}_{k-1}^*)^T$, $\tilde{\boldsymbol{m}}^* = (\tilde{m}_1^*, ..., \tilde{m}_k^*)^T$, and $\tilde{\boldsymbol{\sigma}}^{2*} = (\tilde{\sigma}_1^{2*}, ..., \tilde{\sigma}_k^{2*})^T$. Furthermore, define $\tilde{\boldsymbol{\theta}}^* = ((\tilde{\boldsymbol{m}}^*)^T, (\tilde{\boldsymbol{\pi}}^*)^T, (\tilde{\boldsymbol{\sigma}}^{2*})^T)^T$, $\boldsymbol{\beta} = ((\tilde{\boldsymbol{\pi}})^T, (\tilde{\boldsymbol{\sigma}}^{2*})^T)^T$.

Lemma 3. Suppose that conditions (C2)-(C10) are satisfied, then,

$$\sup_{x \in \mathscr{X}} \left| \tilde{\boldsymbol{\theta}}^* - f^{-1}(x) I_\theta^{-1}(x) S_n \right| = O_p(h^2 + \gamma_n \log^{1/2}(1/h)),$$

where S_n is defined in (2.17).

Proof of Lemma 3.

Since $\{\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{m}}, \tilde{\boldsymbol{\sigma}}^2\}$ maximizes $\ell_1(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\sigma}^2)$ defined in (2.5), it is easy to see that $\tilde{\boldsymbol{\theta}}^*$ maximizes

$$\ell_n^*(\boldsymbol{\theta}^*) = h \sum_{i=1}^{n} \{\ell(\boldsymbol{\theta}(x) + \gamma_n \boldsymbol{\theta}^*, Y_i) - \ell(\boldsymbol{\theta}(x), Y_i)\} K_h(X_i - x), \tag{2.15}$$

where $\ell(\cdot)$ is defined in (2.12). Apply a Taylor's expansion to $\ell(\boldsymbol{\theta}(x)+\gamma_n\boldsymbol{\theta}^*, Y_i)$ at $\boldsymbol{\theta}(x)$, we have

$$\ell_n^*(\boldsymbol{\theta}^*) = S_n\boldsymbol{\theta}^* + \frac{1}{2}\boldsymbol{\theta}^{*T}W_n\boldsymbol{\theta}^* + o_p(\|\boldsymbol{\theta}^*\|^2), \tag{2.16}$$

where

$$\begin{aligned} S_n &= \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\boldsymbol{\theta}(x), Y_i)}{\partial\boldsymbol{\theta}}K_h(X_i - x), \\ W_n &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\boldsymbol{\theta}(x), Y_i)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}K_h(X_i - x). \end{aligned} \tag{2.17}$$

Noting that $E(W_n) = -f(x)I_\theta(x) + o_p(1)$, by the weak law of large numbers, $W_n = -f(x)I_\theta(x) + o_p(1)$, and therefore,

$$\ell_n^*(\boldsymbol{\theta}^*) = S_n\boldsymbol{\theta}^* - \frac{1}{2}f(x)\boldsymbol{\theta}^{*T}I_\theta(x)\boldsymbol{\theta}^* + o_p(\|\boldsymbol{\theta}^*\|^2). \tag{2.18}$$

By definition, each element of W_n is the sum of i.i.d. random variables, by Lemma 2 and assumption (C9), it can be shown that for all $x \in \mathscr{X}$, W_n converges to $-f(x)I_\theta(x)$ uniformly. From (2.18) and assumption (C7) and (C9), we know that $-\ell_n^*(\boldsymbol{\theta}^*)$ is convex function defined on a convex open set, when n is large enough. Therefore, by the convexity lemma (Pollard, 1991),

$$\sup_{x\in\mathscr{X}}\left|(S_n\boldsymbol{\theta}^* + \frac{1}{2}\boldsymbol{\theta}^{*T}W_n\boldsymbol{\theta}^*) - [S_n\boldsymbol{\theta}^* - \frac{1}{2}f(x)\boldsymbol{\theta}^{*T}I_\theta(x)\boldsymbol{\theta}^*]\right| \xrightarrow{P} 0$$

holds uniformly for all $x \in \mathscr{X}$ and $\boldsymbol{\theta}^*$ in any compact set. We know that $-f^{-1}(x)I_\theta^{-1}(x)S_n$ is a unique maximizer of (2.18), and by definition, $\tilde{\boldsymbol{\theta}}^*$ is a maximizer of (2.16), then, by Lemma A.1 of Carroll et al. (1997),

$$\sup_{x\in\mathscr{X}}\left|\tilde{\boldsymbol{\theta}}^* - f^{-1}(x)I_\theta^{-1}(x)S_n\right| \xrightarrow{P} 0,$$

which also implies that

$$\tilde{\boldsymbol{\theta}}^* = f^{-1}(x) I_\theta^{-1}(x) S_n + o_p(1). \tag{2.19}$$

Since $\tilde{\boldsymbol{\theta}}^*$ maximizes (2.15), then

$$\frac{\partial \ell_n^*(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*}\Big|_{\boldsymbol{\theta}^*=\tilde{\boldsymbol{\theta}}^*} = h\gamma_n \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}(x) + \gamma_n \tilde{\boldsymbol{\theta}}^*, Y_i)}{\partial \boldsymbol{\theta}} K_h(X_i - x) = 0. \tag{2.20}$$

Apply a Taylor's expansion to (2.20) at $\boldsymbol{\theta}(x)$,

$$0 = h\gamma_n \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}(x), Y_i)}{\partial \boldsymbol{\theta}} K_h(X_i - x) + h\gamma_n^2 \tilde{\boldsymbol{\theta}}^* \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{\theta}(x), Y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} K_h(X_i - x) + O_p(\gamma_n \|\tilde{\boldsymbol{\theta}}^*\|^2),$$

that is, $W_n \tilde{\boldsymbol{\theta}}^* + O_p(\gamma_n \|\tilde{\boldsymbol{\theta}}^*\|^2) = -S_n$. Therefore,

$$\{W_n - E(W_n)\}\tilde{\boldsymbol{\theta}}^* + O_p(\gamma_n \|\tilde{\boldsymbol{\theta}}^*\|^2) = -S_n - E(W_n)\tilde{\boldsymbol{\theta}}^* = -S_n + f(x) I_\theta(x) \tilde{\boldsymbol{\theta}}^*, \tag{2.21}$$

where the last equality is deduced by the fact that $E(W_n) = -f(x) I_\theta(x) + o_p(1)$. Notice the structure of W_n, from Lemma 2, we know that

$$\sup_{x \in \mathscr{X}} |W_n - E(W_n)| = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}.$$

From (2.19), (2.27) and (2.28), it is easy to show that $\sup_{x \in \mathscr{X}} |\tilde{\boldsymbol{\theta}}^*| = O_p(1)$, and thus

$$\{W_n - E(W_n)\}\tilde{\boldsymbol{\theta}}^* + O_p(\gamma_n \|\tilde{\boldsymbol{\theta}}^*\|^2) = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}.$$

Combined with (2.21), we have

$$\sup_{x \in \mathscr{X}} \left| -S_n + f(x) I_\theta(x) \tilde{\boldsymbol{\theta}}^* \right| = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}.$$

Since $f(x)$ and $I_\theta(x)$ are bounded and continuous functions in a closed set of $\mathscr{X}$ and $I_\theta(x)$

is positive definite,

$$\sup_{x\in\mathscr{X}}\left|\tilde{\boldsymbol{\theta}}^{*}-f^{-1}(x)I_{\theta}^{-1}S_{n}\right|=O_{p}\{h^{2}+\gamma_{n}\log^{1/2}(1/h)\}.$$

Proof of Theorem 2.2.2.

Define $\hat{\boldsymbol{\beta}}^{*}=\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$, where $\hat{\boldsymbol{\beta}}$ maximizes $\ell_2(\boldsymbol{\beta})$ in (2.6) and $\boldsymbol{\beta}=(\boldsymbol{\pi}^{T},(\boldsymbol{\sigma}^{2})^{T})^{T}$ is the true value. Let

$$\ell(\tilde{\boldsymbol{m}}(X_i),\boldsymbol{\beta},Y_i)=\log\{\sum_{j=1}^{k}\pi_j\phi(Y_i|\tilde{m}_j(X_i),\sigma_j^2\},$$

$$\ell(\tilde{\boldsymbol{m}}(X_i),\boldsymbol{\beta}+\boldsymbol{\beta}^{*}/\sqrt{n},Y_i)=\log\{\sum_{j=1}^{k}(\pi_j+\pi_j^{*}/\sqrt{n})\phi(Y_i|\tilde{m}_j(X_i),\sigma_j^2+\sigma_j^{2*}/\sqrt{n}\}.$$

Since $\hat{\boldsymbol{\beta}}$ maximizes ℓ_2, it is easy to see that $\hat{\boldsymbol{\beta}}^{*}$ maximizes

$$\ell_n(\boldsymbol{\beta}^{*})=\sum_{i=1}^{n}\{\ell(\tilde{\boldsymbol{m}}(X_i),\boldsymbol{\beta}+\boldsymbol{\beta}^{*}/\sqrt{n},Y_i)-\ell(\tilde{\boldsymbol{m}}(X_i),\boldsymbol{\beta},Y_i\}.$$

Similar to the proof of Lemma 2, apply a Taylor's expansion to $\ell(\tilde{\boldsymbol{m}}(X_i),\boldsymbol{\beta}+\boldsymbol{\beta}^{*}/\sqrt{n},Y_i)$ at $\boldsymbol{\beta}$ and after some calculation,

$$\ell_n(\boldsymbol{\beta}^{*})=A_n\boldsymbol{\beta}^{*}+\frac{1}{2}\boldsymbol{\beta}^{*T}B_n\boldsymbol{\beta}^{*}+o_p(\|\boldsymbol{\beta}^{*}\|^2),$$

where

$$A_n=\sqrt{\frac{1}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\tilde{\boldsymbol{m}}(X_i),\boldsymbol{\beta},Y_i)}{\partial\boldsymbol{\beta}},$$

$$B_n=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\tilde{\boldsymbol{m}}(X_i),\boldsymbol{\beta},Y_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{T}}.$$

Similar to the proof of Lemma 3, it is not difficult to show that $B_n=-B+o_p(1)$, where

$B = E\{I_\beta(X)\}$, and then

$$\ell_n(\boldsymbol{\beta}^*) = A_n\boldsymbol{\beta}^* - \frac{1}{2}\boldsymbol{\beta}^{*T}B\boldsymbol{\beta}^* + o_p(1). \tag{2.22}$$

By (2.22) and the quadratic approximation lemma,

$$\hat{\boldsymbol{\beta}}^* = B^{-1}A_n + o_p(1). \tag{2.23}$$

Next, apply a Taylor's expansion to A_n at $\boldsymbol{m}(X_i)$,

$$\begin{aligned} A_n &= \sqrt{\frac{1}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\tilde{\boldsymbol{m}}(X_i),\boldsymbol{\beta},Y_i)}{\partial\boldsymbol{\beta}} \\ &= \sqrt{\frac{1}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\boldsymbol{m}(X_i),\boldsymbol{\beta},Y_i)}{\partial\boldsymbol{\beta}} + R_{1n} + O_p(\sqrt{\frac{1}{n}}\|\tilde{\boldsymbol{m}}-\boldsymbol{m}\|_\infty^2), \end{aligned}$$

where

$$R_{1n} = \sqrt{\frac{1}{n}}\sum_{i=1}^{n}\frac{\partial^2\ell(\boldsymbol{m}(X_i),\boldsymbol{\beta},Y_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{m}^T}(\tilde{\boldsymbol{m}}(X_i)-\boldsymbol{m}(X_i)).$$

From Lemma 3,

$$\tilde{\boldsymbol{\theta}}^*(X_i) = f^{-1}(X_i)I_\theta^{-1}(X_i)\sqrt{\frac{h}{n}}\sum_{t=1}^{n}\frac{\partial\ell(\boldsymbol{\theta}(X_i),Y_t)}{\partial\boldsymbol{\theta}}K_h(X_t-X_i) + O_p\{h^2+\gamma_n\log^{1/2}(1/h)\}.$$

Since $\tilde{\boldsymbol{\theta}}^*(X_i) = \sqrt{nh}(\tilde{\boldsymbol{\theta}}(X_i)-\boldsymbol{\theta}(X_i))$,

$$\tilde{\boldsymbol{\theta}}(X_i)-\boldsymbol{\theta}(X_i) = \frac{1}{n}f^{-1}(X_i)I_\theta^{-1}(X_i)\sum_{t=1}^{n}\frac{\partial\ell(\boldsymbol{\theta}(X_i),Y_t)}{\partial\boldsymbol{\theta}}K_h(X_t-X_i)+O_p\{\gamma_nh^2+\gamma_n^2\log^{1/2}(1/h)\} \tag{2.24}$$

Let $\varphi(X_t,Y_t)$ be a $k\times 1$ vector whose elements are the first k entries of $I_\theta^{-1}(X_t)\frac{\partial\ell(\boldsymbol{\theta}(X_t),Y_t)}{\partial\boldsymbol{\theta}}$. From assumption (C1), we know that $O_p\{n^{1/2}[\gamma_nh^2+\gamma_n^2\log^{1/2}(1/h)]\} = o_p(1)$. It can be

shown that $\boldsymbol{m}(X_i) - \boldsymbol{m}(X_t) = O(X_i - X_t)$ and $K(\cdot)$ is symmetric about 0, then by (2.24),

$$\begin{aligned} R_{1n} &= n^{-3/2} \sum_{t=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{m}^T} f^{-1}(X_i) \varphi(X_t, Y_t) K_h(X_i - X_t) + O_p(n^{1/2} h^2) \\ &= R_{2n} + O_p(n^{1/2} h^2). \end{aligned}$$

It can be shown that

$$E[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{m}^T} f^{-1}(X_i) K_h(X_i - X_t)] = I_{\beta m}(X_t).$$

Let $\varpi(X_t, Y_t) = I_{\beta m}(X_t) \varphi(X_t, Y_t)$, and $R_{n3} = -n^{-1/2} \sum_{j=1}^{n} \varpi(X_t, Y_t)$, then

$$R_{n2} - R_{n3} \xrightarrow{P} 0,$$

and therefore

$$A_n = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} \{ \frac{\partial \ell(\boldsymbol{m}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta}} - \varpi(X_i, Y_i) \} + o_p(1),$$

given $nh^4 \to 0$.

To complete the proof, we need to find the mean and variance of A_n. Let

$$\Sigma = Var\{ \frac{\partial \ell(\boldsymbol{\theta}(X), Y)}{\partial \boldsymbol{\beta}} - \varpi(X, Y) \},$$

then $Var(A_n) = \Sigma$. It can be easily seen that the elements of $E\{\frac{\partial \ell(\boldsymbol{\theta}(X), Y)}{\partial \boldsymbol{\beta}}\}$ all equal to 0, and $E\{\varpi(X, Y)\} = 0$, and thus, $E(A_n) = 0$. Therefore by (2.23),

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(0, B^{-1} \Sigma B^{-1}).$$

Proof of Theorem 2.2.3.

Define $\hat{\boldsymbol{m}}^* = \sqrt{nh}(\hat{\boldsymbol{m}}(x) - \boldsymbol{m}(x))$, where $\hat{\boldsymbol{m}}(x)$ maximizes (2.7) and $\boldsymbol{m}(x)$ is the true value.

Apply similar arguments as in the proof of Lemma 3, it is easy to see

$$\hat{\boldsymbol{m}}^*(x) = f(x)^{-1} I_m(x)^{-1} \hat{S}_n + o_p(1), \tag{2.25}$$

where

$$\hat{S}_n = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{m}(x), \hat{\boldsymbol{\beta}}, Y_i)}{\partial \boldsymbol{m}} K_h(X_i - x). \tag{2.26}$$

Apply a Taylor's expansion to (2.26) at $\boldsymbol{\beta}$, we have

$$\begin{aligned}\hat{S}_n =& \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{m}} K_h(X_i - x) + \sqrt{\frac{h}{n}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{m} \partial \boldsymbol{\beta}^T} K_h(X_i - x) + o_p(1) \\ \equiv& S_n + D_n + o_p(1).\end{aligned}$$

where S_n is defined in (2.17).

Notice that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$ and

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{m} \partial \boldsymbol{\beta}^T} K_h(X_i - x) = -f(x) I_{\beta m}^T(x) + o_p(1),$$

then

$$\begin{aligned}D_n &= \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sqrt{h} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{m} \partial \boldsymbol{\beta}^T} K_h(X_i - x) \\ &= -\sqrt{h} f(x) I_{\beta m}^T(x) + o_p(1).\end{aligned}$$

Thus, from (2.25)

$$\hat{\boldsymbol{m}}^*(x) = f(x)^{-1} I_m(x)^{-1} S_n + o_p(1).$$

To complete the proof, we need to calculate the expectation and variance of S_n. Let $\Lambda(u|x) =$

$E[\frac{\partial \ell(\boldsymbol{m}(x),\boldsymbol{\beta},Y)}{\partial \boldsymbol{m}}|X=u]$, then

$$E\{\frac{\partial \ell(\boldsymbol{m}(x),\boldsymbol{\beta},Y)}{\partial \boldsymbol{m}}K_h(X-x)\} = E\{E[\frac{\partial \ell(\boldsymbol{m}(x),\boldsymbol{\beta},Y)}{\partial \boldsymbol{m}}K_h(X-x)|X=x_0]\}$$
$$= [\frac{1}{2}f(x)\Lambda''(x|x) + f'(x)\Lambda'(x|x)]\kappa_2 h^2,$$

and so

$$E(S_n) = \sqrt{nh}[\frac{1}{2}f(x)\Lambda''(x|x) + f'(x)\Lambda'(x|x)]\kappa_2 h^2. \tag{2.27}$$

Since

$$E\{\frac{\partial \ell(\boldsymbol{m}(x),\boldsymbol{\beta},Y)}{\partial \boldsymbol{m}}K_h(X-x)\}^2 = E\{E[(\frac{\partial \ell(\boldsymbol{m}(x),\boldsymbol{\beta},Y)}{\partial \boldsymbol{m}}K_h(X-x))^2|X=x_0]\}$$
$$= \frac{1}{h}I_m(x)\nu_0 + o_p(1),$$

then

$$Var(S_n) = f(x)I_m(x)\nu_0. \tag{2.28}$$

To complete the proof, let

$$\Delta(x) = I_m^{-1}(x)[\frac{1}{2}\Lambda''(x|x) + f^{-1}(x)f'(x)\Lambda'(x|x)]\kappa_2 h^2,$$

and $\Delta_m(x)$ be a $k \times 1$ vector whose elements are the first k entries of $\Delta(x)$, then

$$\sqrt{nh}(\hat{\boldsymbol{m}}(x) - \boldsymbol{m}(x) - \Delta_m(x) + o_p(h^2)) \xrightarrow{D} N(0, f^{-1}(x)I_m^{-1}(x)\nu_0).$$

Proof of Theorem 2.2.4.

(i) Assume the latent variables $\{Z_i, i = 1, ..., n\}$ be a random sample from population Z,

then the conditional distribution of Z given Y and $\boldsymbol{\theta}$ is

$$P(Z_i = j|Y, \boldsymbol{\theta}) = \frac{\pi_j \phi(Y|m_j, \sigma_j^2)}{\sum_{j=1}^k \pi_j \phi(Y|m_j, \sigma_j^2)}. \tag{2.29}$$

Given $\boldsymbol{\theta}^{(l)}(X_i) = (\boldsymbol{m}^{(l)}(X_i), \boldsymbol{\pi}^{(l)}(X_i), \boldsymbol{\sigma}^{2(l)}(X_i))$, for any $i = 1, ..., n$, $P(Z_i = j|Y_i, \boldsymbol{\theta}^{(l)}(X_i)) = p_{ij}^{(l+1)}$ and $\sum_{j=1}^k p_{ij}^{(l+1)} = 1$, $i = 1, ..., n$. Therefore,

$$\begin{aligned}
\ell_1(\boldsymbol{\theta}) &= \sum_{i=1}^n \log\{\sum_{j=1}^k \pi_j \phi(Y_i|m_j, \sigma_j^2)\}(\sum_{j=1}^k p_{ij}^{(l+1)}) K_h(X_i - x) \\
&= \sum_{i=1}^n \{\sum_{j=1}^k \log\{\sum_{j=1}^k \pi_j \phi(Y_i|m_j, \sigma_j^2)\} p_{ij}^{(l+1)}\} K_h(X_i - x).
\end{aligned} \tag{2.30}$$

From (2.29), it is easy to show that

$$\log\{\sum_{j=1}^k \pi_j \phi(Y_i|m_j, \sigma_j^2)\} = \log\{\pi_j \phi(Y_i|m_j, \sigma_j^2)\} - \log\{P(Z_i = j|Y, \boldsymbol{\theta})\}. \tag{2.31}$$

Combing (2.30) and (2.31),

$$\begin{aligned}
\ell_1(\boldsymbol{\theta}) = &\sum_{i=1}^n \{\sum_{j=1}^k \log\{\pi_j \phi(Y_i|m_j, \sigma_j^2)\} p_{ij}^{(l+1)}\} K_h(X_i - x) \\
&- \sum_{i=1}^n \{\sum_{j=1}^k \log\{P(Z_i = j|Y, \boldsymbol{\theta})\} p_{ij}^{(l+1)}\} K_h(X_i - x).
\end{aligned} \tag{2.32}$$

Based on the M-step of (2.8), (2.9) and (2.10), we have

$$\begin{aligned}
&n^{-1} \sum_{i=1}^n \{\sum_{j=1}^k \log\{\pi_j^{(l+1)}(x) \phi(Y_i|m_j^{(l+1)}(x), \sigma_j^{2(l+1)}(x))\} p_{ij}^{(l+1)}\} K_h(X_i - x) \\
\geq &n^{-1} \sum_{i=1}^n \{\sum_{j=1}^k \log\{\pi_j^{(l)}(x) \phi(Y_i|m_j^{(l)}(x), \sigma_j^{2(l)}(x))\} p_{ij}^{(l+1)}\} K_h(X_i - x).
\end{aligned}$$

To complete the proof, based on (2.32), we only need to show

$$\limsup_{n\to\infty} n^{-1}\sum_{i=1}^{n}\{\sum_{j=1}^{k}\log\{\frac{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l+1)}(x))}{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(x))}\}p_{ij}^{(l+1)}\}K_h(X_i-x)\leq 0$$

in probability. Define

$$L=n^{-1}\sum_{i=1}^{n}\{\sum_{j=1}^{k}\log\{\frac{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l+1)}(x))}{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(x))}\}p_{ij}^{(l+1)}\}K_h(X_i-x),$$
$$U=n^{-1}\sum_{i=1}^{n}\log\{\sum_{j=1}^{k}\{\frac{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l+1)}(x))}{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(x))}\}p_{ij}^{(l+1)}\}K_h(X_i-x),$$

then, by Jensen's inequality, $L\leq U$. We complete the proof by showing that $U\xrightarrow{P}0$. Without loss of generality, assume that $P(Z_i=j|Y,\boldsymbol{\theta}^{(l)}(x))\geq\delta>0$ for some small value δ. Since

$$\begin{aligned}E(U)&=E\{\log[\sum_{j=1}^{k}\frac{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l+1)}(x))}{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(x))}P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(X_i))]K_h(X_i-x)\}\\&=E\{E[\log\sum_{j=1}^{k}\frac{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l+1)}(x))}{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(x))}P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(X_i))]K_h(X_i-x)|Y]\},\end{aligned}$$

by similar argument as in the proof of Theorem 2.2.2 and Theorem 2.2.3, it can be shown that $E(U)\to 0$. Notice that the leading term of $Var(U)$ is

$$n^{-1}E\{\log[\sum_{j=1}^{k}\frac{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l+1)}(x))}{P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(x))}P(Z_i=j|Y_i,\boldsymbol{\theta}^{(l)}(X_i))]K_h(X_i-x)\}^2,$$

which can be shown to have a order of $O_p((nh)^{-1})$. Therefore, by Chebyshv's inequality, $U=o_p(1)$, and thus completes the proof.

(*ii*) The conditional distribution of Z given Y and $\boldsymbol{m}$ is

$$P(Z_i = j|Y, \boldsymbol{m}, \hat{\boldsymbol{\beta}}) = \frac{\hat{\pi}_j \phi(Y|m_j, \hat{\sigma}_j^2)}{\sum_{j=1}^k \hat{\pi}_j \phi(Y|m_j, \hat{\sigma}_j^2)},$$

and $P(Z_i = j|Y_i, \boldsymbol{m}^{(l)}(X_i), \hat{\boldsymbol{\beta}}) = p_{ij}^{(l+1)}$ and $\sum_{j=1}^k p_{ij}^{(l+1)} = 1$, where $p_{ij}^{(l+1)}$ is defined in (2.11). The rest of the proof is in line with part (i), and thus is omitted here.

(*iii*) Notice that by fixing $\tilde{\boldsymbol{m}}(\cdot) = \boldsymbol{m}^{(l)}(\cdot)$, $\ell^*(\boldsymbol{\pi}, \boldsymbol{m}^{(l)}(\cdot), \boldsymbol{\sigma}^2) = \ell_2(\boldsymbol{\pi}, \boldsymbol{\sigma}^2)$. Therefore, by the ascent property of the ordinary EM algorithm,

$$\ell^*(\boldsymbol{\pi}^{(l+1)}, \boldsymbol{m}^{(l)}(\cdot), \boldsymbol{\sigma}^{2(l+1)}) = \ell_2(\boldsymbol{\pi}^{(l+1)}, \boldsymbol{\sigma}^{2(l+1)}) \geq \ell_2(\boldsymbol{\pi}^{(l)}, \boldsymbol{\sigma}^{2(l)}) = \ell^*(\boldsymbol{\pi}^{(l)}, \boldsymbol{m}^{(l)}(\cdot), \boldsymbol{\sigma}^{2(l)})$$

Thus, to complete the proof, we only need to show

$$\liminf_{n\to\infty} n^{-1}[\ell^*(\boldsymbol{\pi}^{(l+1)}, \boldsymbol{m}^{(l+1)}(\cdot), \boldsymbol{\sigma}^{2(l+1)}) - \ell^*(\boldsymbol{\pi}^{(l+1)}, \boldsymbol{m}^{(l)}(\cdot), \boldsymbol{\sigma}^{2(l+1)})] \geq 0$$

If we fix $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}^{(l+1)}$ and $\hat{\boldsymbol{\sigma}}^2 = \boldsymbol{\sigma}^{2(l+1)}$, then by part (ii),

$$\liminf_{n\to\infty} n^{-1}[\ell_3(\boldsymbol{m}^{(l+1)}(x)) - \ell_3(\boldsymbol{m}^{(l)}(x))] \geq 0$$

in probability for any $x \in \{X_t, t = 1, ..., n\}$. Therefore,

$$\begin{aligned}
&\liminf_{n\to\infty} n^{-2} \sum_{t=1}^n f(X_t)^{-1}[\ell_3(\boldsymbol{m}^{(l+1)}(X_t)) - \ell_3(\boldsymbol{m}^{(l)}(X_t))] \\
\geq &\liminf_{n\to\infty} n^{-1} \sum_{t=1}^n \liminf_{n\to\infty} n^{-1} f(X_t)^{-1}[\ell_3(\boldsymbol{m}^{(l+1)}(X_t)) - \ell_3(\boldsymbol{m}^{(l)}(X_t))] \geq 0.
\end{aligned}$$

Since $K(\cdot)$ is symmetric about 0, $K_h(X_i - X_t) = K_h(X_t - X_i)$, and therefore,

$$n^{-2}\sum_{t=1}^{n} f(X_t)^{-1}\ell_3(\boldsymbol{m}^{(l)}(X_t)) = n^{-2}\sum_{t=1}^{n} f(X_t)^{-1}\sum_{i=1}^{n}\log\{\sum_{j=1}^{k}\hat{\pi}_j\phi(Y_i|m_j^{(l)}(X_t),\hat{\sigma}_j^2)\}K_h(X_i - X_t)$$

$$=n^{-1}\sum_{i=1}^{n}\{n^{-1}\sum_{t=1}^{n} f(X_t)^{-1}\log[\sum_{j=1}^{k}\hat{\pi}_j\phi(Y_i|m_j^{(l)}(X_t),\hat{\sigma}_j^2)]K_h(X_t - X_i)\} = n^{-1}\sum_{i=1}^{n}\Gamma_i^{(l)},$$

where

$$\Gamma_i^{(l)} = n^{-1}\sum_{t=1}^{n} f(X_t)^{-1}\log[\sum_{j=1}^{k}\hat{\pi}_j\phi(Y_i|m_j^{(l)}(X_t),\hat{\sigma}_j^2)]K_h(X_t - X_i).$$

It can be shown that

$$E(\Gamma_i^{(l)}|X_i, Y_i) = \log[\sum_{j=1}^{k}\hat{\pi}_j\phi(Y_i|m_j^{(l)}(X_i),\hat{\sigma}_j^2)](1 + o_p(1))$$

and $Var(\Gamma_i^{(l)}|X_i, Y_i) = O_p((nh)^{-1})$. In addition, it is clear that

$$\sum_{i=1}^{n} E(\Gamma_i^{(l)}|X_i, Y_i) = \ell^*(\boldsymbol{\pi}^{(l+1)}, \boldsymbol{m}^{(l)}(\cdot), \boldsymbol{\sigma}^{2(l+1)}),$$

$$\sum_{i=1}^{n} E(\Gamma_i^{(l+1)}|X_i, Y_i) = \ell^*(\boldsymbol{\pi}^{(l+1)}, \boldsymbol{m}^{(l+1)}(\cdot), \boldsymbol{\sigma}^{2(l+1)}),$$

which completes the proof.

Proof of Theorem 2.2.5.

Since $\hat{\boldsymbol{\beta}}$ has faster convergence rate than $\hat{\boldsymbol{m}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$ has the same asymptotic properties as if $\boldsymbol{\beta}$ were known. Therefore, in the following proof, we study the property of $\hat{\boldsymbol{m}}(\cdot)$ assuming $\boldsymbol{\beta}$ to be known.

Define $\frac{\partial\ell(\theta(X_i),Y_i)}{\partial\theta} = q_{\theta i}$, $\frac{\partial^2\ell(\theta(X_i),Y_i)}{\partial\theta\partial\theta^T} = q_{\theta\theta i}$ and similarly, define q_{mi}, q_{mmi} and so on. Let $\tilde{\boldsymbol{\theta}}$ be the estimator under H_1 (Huang et al., 2013), and $\hat{\boldsymbol{m}}$ be the estimator under H_0 (model

(2.1)). From previous proof, we have

$$\tilde{\boldsymbol{\theta}}(X_i) - \boldsymbol{\theta}(X_i) = \frac{1}{n} f^{-1}(X_i) I_\theta^{-1}(X_i) \sum_{t=1}^{n} q_{\theta i} K_h(X_t - X_i)(1 + o_p(1)) \tag{2.33}$$

$$\hat{\boldsymbol{m}}(X_i) - \boldsymbol{m}(X_i) = \frac{1}{n} f^{-1}(X_i) I_m^{-1}(X_i) \sum_{t=1}^{n} q_{mi} K_h(X_t - X_i)(1 + o_p(1)) \tag{2.34}$$

By (2.33) and (2.34), we can obtain that

$$\begin{aligned}
&\sum_{i=1}^{n} \ell(\tilde{\boldsymbol{\theta}}(X_i), Y_i) - \sum_{i=1}^{n} \ell(\boldsymbol{\theta}(X_i), Y_i) = \{\frac{1}{n} \sum_{i,l} q_{\theta i}^T f^{-1}(X_l) I_\theta^{-1}(X_l) q_{\theta l} K_h(X_i - X_l) \\
&+ \frac{1}{2n^2} \sum_{i,j,l} q_{\theta i}^T f^{-2}(X_l) I_\theta^{-1}(X_l) q_{\theta\theta l} I_\theta^{-1}(X_l) q_{\theta j} K_h(X_i - X_l) K_h(X_j - X_l)\}(1 + o_p(1)), \\
&\sum_{i=1}^{n} \ell(\hat{\boldsymbol{m}}(X_i), Y_i) - \sum_{i=1}^{n} \ell(\boldsymbol{m}(X_i), Y_i) = \{\frac{1}{n} \sum_{i,l} q_{mi}^T f^{-1}(X_l) I_m^{-1}(X_l) q_{ml} K_h(X_i - X_l) \\
&+ \frac{1}{2n^2} \sum_{i,j,l} q_{mi}^T f^{-2}(X_l) I_m^{-1}(X_l) q_{mml} I_m^{-1}(X_l) q_{mj} K_h(X_i - X_l) K_h(X_j - X_l)\}(1 + o_p(1)),
\end{aligned}$$

and so,

$$\begin{aligned}
T =& \frac{1}{n} \sum_{i,l} [q_{\theta i}^T I_\theta^{-1}(X_l) q_{\theta l} - q_{mi}^T I_m^{-1}(X_l) q_{ml}] f^{-1}(X_l) K_h(X_i - X_l) + \frac{1}{2n^2} \sum_{i,j,l} [q_{\theta i}^T I_\theta^{-1}(X_l) q_{\theta\theta l} \\
&\times I_\theta^{-1}(X_l) q_{\theta j} - q_{mi}^T I_m^{-1}(X_l) q_{mml} I_m^{-1}(X_l) q_{mj}] f^{-2}(X_l) K_h(X_i - X_l) K_h(X_j - X_l) \\
\equiv& \Lambda_n + \frac{1}{2} \Gamma_n.
\end{aligned}$$

By similar argument as Fan et al. (2001), it can be shown that under conditions (C9)-(C12),

as $h \to 0$, $nh^{3/2} \to \infty$,

$$\Lambda_n = \frac{2k-1}{h} K(0) E f(X)^{-1} + \frac{1}{n} \sum_{l \neq i} [q_{\theta i}^T I_\theta^{-1}(X_l) q_{\theta l} - q_{mi}^T I_m^{-1}(X_l) q_{ml}] f^{-1}(X_l) K_h(X_i - X_l) + o_p(h^{-1/2}),$$

$$\Gamma_n = -\frac{(2k-1)}{h} E f(X)^{-1} \int K^2(t) dt - \frac{2}{n} \sum_{i<j} [q_{\theta i}^T I_\theta^{-1}(X_i) q_{\theta\theta i} q_{\theta j} - q_{mi}^T I_m^{-1}(X_i) q_{mmi} q_{mj}] f^{-1}(X_i)$$

$$\times K_h * K_h(X_i - X_j) + o_p(h^{-1/2}).$$

Therefore,

$$T = \mu_n + W_n / 2\sqrt{h} + o_p(h^{-1/2}),$$

where $\mu_n = \frac{(2k-1)|\mathscr{X}|}{h}[K(0) - 0.5 \int K^2(t) dt]$,

$$W_n = \frac{\sqrt{h}}{n} \sum_{i \neq j} \{ q_{\theta i}^T I_\theta^{-1}(X_j) [2K_h(X_i - X_j) - q_{\theta\theta j} K_h * K_h(X_i - X_j)] f^{-1}(X_j) q_{\theta j}$$

$$- q_{mi}^T I_m^{-1}(X_j) [2K_h(X_i - X_j) - q_{mmj} K_h * K_h(X_i - X_j)] f^{-1}(X_j) q_{mj}.$$

It can be shown that $\mathrm{Var}(W_n) \to \zeta$, where $\zeta = 2(2k-1) E f^{-1}(X) \int [2K(t) - K * K(t)]^2 dt$. Apply Proposition 3.2 in de Jong (1987), we obtain that

$$W_n \xrightarrow{D} N(0, \zeta),$$

and completes the proof.

Table 2.4: *Pointwise coverage probabilities*

	h		0.111	0.222	0.333	0.444	0.556	0.667	0.778	0.889
	US	m_1	92.00	93.00	93.80	92.40	93.40	93.40	94.20	92.80
$n = 200$		m_2	90.00	92.20	95.20	93.40	94.20	94.40	92.80	93.20
	AS	m_1	92.20	91.40	90.80	87.60	90.00	91.40	93.00	90.40
$(0.5, 0.5)$		m_2	85.40	89.40	85.00	89.00	87.00	84.20	89.40	89.40
	OS	m_1	92.00	77.00	80.80	83.00	87.00	80.80	79.60	89.80
		m_2	58.60	80.20	53.60	76.60	73.60	48.80	80.80	73.00
	US	m_1	93.40	94.40	95.60	93.40	93.00	95.80	96.00	94.00
$n = 400$		m_2	97.20	94.40	94.20	91.60	93.40	94.60	95.00	94.80
	AS	m_1	91.40	93.00	93.60	91.80	90.60	92.40	92.00	91.60
$(0.5, 0.5)$		m_2	89.80	91.80	87.40	90.00	88.40	88.80	89.40	90.40
	OS	m_1	88.80	76.60	81.60	89.00	86.00	80.80	79.60	88.80
		m_2	61.80	82.20	51.40	78.60	79.80	48.60	80.00	73.80
	US	m_1	91.40	97.00	93.40	93.60	93.00	94.80	94.60	93.40
$n = 200$		m_2	89.00	93.20	92.20	91.00	92.80	92.40	93.40	90.80
	AS	m_1	92.40	88.60	91.40	90.20	86.40	89.60	89.60	89.20
$(0.7, 0.3)$		m_2	82.60	89.00	89.40	86.20	84.20	84.20	87.20	86.40
	OS	m_1	91.40	62.20	67.20	82.80	82.00	67.00	62.80	90.00
		m_2	60.60	83.80	63.80	81.60	76.00	57.20	78.20	76.00
	US	m_1	92.40	94.20	93.60	94.60	93.40	96.80	94.00	95.40
$n = 400$		m_2	93.40	95.60	93.00	94.00	93.60	93.60	93.80	93.00
	AS	m_1	91.80	90.80	89.40	91.20	91.80	92.20	88.80	92.20
$(0.7, 0.3)$		m_2	83.60	89.00	87.20	89.20	88.60	85.80	88.20	88.60
	OS	m_1	90.40	60.80	67.00	87.20	86.60	68.20	62.00	87.20
		m_2	56.40	81.00	60.40	78.60	79.60	56.80	81.00	74.00

Table 2.5: *Average of MSPE.*

	5-fold CV	10-fold CV	MCCV s=10	MCCV s=20
US House Price Index Data				
Model (2.1)	0.089	0.107	0.089	0.090
NMR (Huang et al., 2013)	0.109	0.124	0.090	0.091
NO data				
Model (2.1)	0.987	0.940	1.038	1.033
NMR (Huang et al., 2013)	1.851	1.568	1.767	1.920

Chapter 3

Mixture of Regression Models with Single-Index

Abstract

In this chapter, we apply the idea of single-index to the mixture of regression models and propose three new classes of models: mixture of single-index models (MSIM), mixture of regression models with varying single-index proportions (MRSIP), and mixture of regression models with varying single-index proportions and variances (MRSIPV). Backfitting estimates and the corresponding algorithms have been proposed for the new models to achieve the optimal convergence rates for both parameters and nonparametric functions. We show that the nonparametric functions can be estimated 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. Simulation studies and a real data example have been conducted to demonstrate the finite sample performance of the proposed models.

3.1 Introduction

The single-index model has received much attention in recent years due to its application to a variety of fields, such as econometrics, biometrics, and so on. The single-index model has the following form:

$$Y = g(\boldsymbol{\alpha}^T \boldsymbol{x}) + \epsilon,$$

where $Y \in \mathbb{R}$ is a response variable, $\boldsymbol{x} \in \mathbb{R}^p$ are covariates; $g(\cdot)$ is an unknown univariate measurable function, $\boldsymbol{\alpha} \in \mathbb{R}^p$ is an unknown parametric vector; and ϵ is the random error independent of $\boldsymbol{x}$, with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. The appeal of the single-index model 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 Y to estimate the nonparametric link $g(\cdot)$, and thus avoid the curse of dimensionality when nonparametric smoothing is employed. These models are often used as a reasonable compromise between fully parametric and fully nonparametric modeling.

The motivation, importance, and broad potential applications of single-index model are widely discussed in the literature. Härdle and Stoker (1989) and Ichimura (1993) have given examples of classical regression, discrete regression, and censored regression that can all be classified as single-index models. Carroll et al. (1997) have summarized the remarks of Li (1991):

1. As a practical matter, it is important to lower the dimensionality before fitting a data, and single-index models provide a readily interpretable meanings of performing this reduction.

2. If the link function is monotone, then $\boldsymbol{\alpha}$ has the same general meaning as in ordinary linear models.

3. Given an estimated "direction", the multivariate model fitting is reduced to a more manageable low-dimensional modeling problem.

Because of its importance, much efforts have been devoted to studying its estimation and other relevant inference problems. Härdle et al. (1993) employed the kernel smoothing method to study the single-index model, and gave an empirical rule for bandwidth selection. Ichimura (1993) studied the properties of a semiparametric least-squares estimator in a general single-index model. Zhang et al. (2010) extended the generalized likelihood ratio test to the single-index models, basing on the estimates obtained by the local linear method. Stute and Zhu (2005) studied the goodness-of-fit testing of single-index models. Carroll et al. (1997) extended the idea and proposed generalized partially linear single-index models. Using local linear methods, Carroll et al. (1997) proposed estimates of the unknown parameters and the unknown link function, and showed that a semiparametric efficient estimator of the direction can be obtained. Xia and Li (1999) extended the idea to the adaptive varying-coefficient model. They proposed estimating coefficient functions with a given bandwidth and a direction $\boldsymbol{\alpha}$, and then to choose the bandwidth and the direction by cross-validation. Fan et al. (2003) also studied the adaptive varying-coefficient linear models, and proposed a hybrid backfitting algorithm, alternating between estimating the index through a one-step scheme and estimating coefficient functions through one-dimensional local linear smoothing.

In this chapter, we apply the idea of single-index to the mixture of regression models, and propose a mixture of single-index models (MSIM), a mixture of regression models with varying single-index proportions (MRSIP), and a mixture of regression models with varying single-index proportions and variances (MRSIPV). 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 and MRSIPV keep the easy interpretation of the linear component regression functions while assuming that the mixing proportions (and variances) 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. We discuss the selection of the three models in a real data analysis.

The rest of the chapter is organized as follows. In Section 3.2, 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 3.3 and 3.4 discuss the MRSIP and the MRSIPV, respectively. Fully-iterated estimates and their asymptotic properties are also studied. In Section 3.5, 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 chapter.

3.2 Mixture of Single-index Models (MSIM)

3.2.1 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 chapter, 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:

$$Y|_{\boldsymbol{x}} \sim \sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^T\boldsymbol{x})\phi(Y_i|m_j(\boldsymbol{\alpha}^T\boldsymbol{x}), \sigma_j^2(\boldsymbol{\alpha}^T\boldsymbol{x})), \tag{3.1}$$

where $\phi(y|\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. Throughout the chapter, we assume that k is fixed, and refer to model (3.1) as a finite semiparametric mixture of regression models, since $\pi_j(\cdot)$, $m_j(\cdot)$, and $\sigma_j^2(\cdot)$ are all nonparametric. When $k = 1$, model (3.1) 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 (3.1) reduces to a finite mixture of linear regression models (Goldfeld and Quandt, 1973). If $\boldsymbol{x}$ is a scalar, then the model (3.1) reduces to the nonparametric mixture of regression models proposed by Huang et al. (2013). Therefore, the proposed model (3.1) is a natural generalization of many existing popular models.

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 result on identifiability of model (3.1) and its proof is given in Section 3.7.

Theorem 3.2.1. *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 (3.1) 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.

3.2.2 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:

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

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, (3.2) 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:

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

where $K_h(z) = h^{-1}K(z/h)$ 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 (3.3). We can then further update the estimate of $\boldsymbol{\alpha}$ by maximizing

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

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

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 (3.3).

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 apply 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 the l^{th} iteration:

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

M-step:

Update the estimates

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

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 (3.3).

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 (3.4).

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 (3.4) using some numerical methods.

Step 4: Iterate Steps 2 - 3 until convergence.

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

Under further conditions defined in Section 3.7, the properties of the one-step estimator, when $\boldsymbol{\alpha}$ is estimated at the order of $O_p(n^{-1/2})$ (such as by sliced inverse regression), is demonstrated in the following theorem.

Theorem 3.2.2. *Assume that conditions (C1)-(C7) in Section 3.7 hold. Then, as $n \to \infty$, $h \to 0$ and $nh \to \infty$, we have*

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

where $\mathcal{B}_1(z) = \mathcal{I}_{\boldsymbol{\theta}}^{(1)-1}\left\{\frac{f'(z)\Lambda_1'(z|z)}{f(z)} + \frac{1}{2}\Lambda_1''(z|z)\right\}\kappa_2 h^2$, 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 (Carroll et al., 1997). 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 3.2.3. *Assume that conditions (C1)-(C8) in Section 3.7 hold. Then, as $n \to \infty$, $nh^4 \to 0$, and $nh^2/\log(1/h) \to \infty$,*

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

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

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

3.3.1 Model Definition and Identifiability

In order to incorporate the predictor information to the component proportions, Huang and Yao (2012) proposed a semiparametric mixture of linear regression models

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

where z can be the same as or part of $\boldsymbol{x}$ and $\pi_j(\cdot)$ is a smoothing function. Note, however, the nonparametric function $\pi_j(z)$ is difficulty to estimate if the dimension of z is high, due to the "curse of dimensionality".

In this section, we propose a mixture of regression models with varying single-index proportions (MRSIP). 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,

$$Y|\boldsymbol{x} \sim \sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^T\boldsymbol{x}) N(\boldsymbol{x}^T\boldsymbol{\beta}_j, \sigma_j^2). \tag{3.5}$$

Since $\pi_j(\cdot)$s are nonparametric, model (3.5) is also a finite semiparametric mixture of regression models.

Theorem 3.3.1. *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 (3.5) is identifiable.*

3.3.2 Estimation Procedure and Asymptotic Properties

The log-likelihood of the collected data is:

$$\ell^{*(2)}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2)\}, \tag{3.6}$$

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, (3.6) 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:

$$\ell_1^{(2)}(\boldsymbol{\pi}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\}K_h(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_i - z). \tag{3.7}$$

Let $\hat{\boldsymbol{\pi}}(\cdot)$ be the result of maximizing (3.7). We can then further update the estimate of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ by maximizing

$$\ell_2^{(2)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \hat{\pi}_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2)\}. \tag{3.8}$$

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 (3.7).**

E-step:

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

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

M-step:

Update the estimate

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

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 (3.8).

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

E-step:

Calculate the expectations of component identities:

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

M-step:

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

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

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 update the estimate of $\boldsymbol{\alpha}$, using some numerical methods.

Step 3.3: Iterate Steps 3.1-3.2 until convergence.

Step 4: Iterate Steps 2-3 until convergence.

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 $q_{\boldsymbol{\pi}}(z) = \frac{\partial \ell(\boldsymbol{\pi}(z), \boldsymbol{\lambda}, x, y)}{\partial \boldsymbol{\pi}}$, $q_{\boldsymbol{\pi}\boldsymbol{\pi}}(z) = \frac{\partial^2 \ell(\boldsymbol{\pi}(z), \boldsymbol{\lambda}, x, y)}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}^T}$, and similarly, define $q_{\boldsymbol{\lambda}}$, $q_{\boldsymbol{\lambda}\boldsymbol{\lambda}}$, and $q_{\boldsymbol{\pi}\boldsymbol{\eta}}$. Denote $\mathcal{I}_{\boldsymbol{\pi}}^{(2)}(z) = -E[q_{\boldsymbol{\pi}\boldsymbol{\pi}}(Z)|Z = z]$ and $\Lambda_2(u|z) = E[q_{\boldsymbol{\pi}}(z)|Z = u]$.

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 3.3.2. *Assume that conditions (C1)-(C4) and (C9)-(C11) in Section 3.7 hold. Then, as $n \to \infty$, $h \to 0$ and $nh \to \infty$, we have*

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

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

Theorem 3.3.3. *Assume that conditions (C1)-(C4) and (C9)-(C12) in Section 3.7 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,

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

3.4 Mixture of Regression Models with Varying Single-Index Proportions and Variances (MRSIPV)

The MRSIPV 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 $\sigma_j^2(\boldsymbol{\alpha}^T\boldsymbol{x})$. That is,

$$Y|\boldsymbol{x}\sim\sum_{j=1}^{k}\pi_j(\boldsymbol{\alpha}^T\boldsymbol{x})N(\boldsymbol{x}^T\boldsymbol{\beta}_j,\sigma_j^2(\boldsymbol{\alpha}^T\boldsymbol{x})). \tag{3.9}$$

Compared to the model (3.5), model (3.9) relaxes the homogenous assumption of the variance functions over the single-index and thus is more general. In addition, one might also use different indexes for the variance function and the the component proportions but with the cost of more complicated computations.

Theorem 3.4.1. *Assume that (i) $\pi_j(z)>0$ and $\sigma_j(z)>0$ 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) $\boldsymbol{\beta}_j$, $j=1,...,k$, are distinct. Then, model (3.9) is identifiable.*

3.4.1 Estimation Procedure and Asymptotic Properties

The log-likelihood of the collected data is:

$$\ell^{*(3)}(\boldsymbol{\pi},\boldsymbol{\sigma}^2,\boldsymbol{\alpha},\boldsymbol{\beta})=\sum_{i=1}^{n}\log\{\sum_{j=1}^{k}\pi_j(\boldsymbol{\alpha}^T\boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_j,\sigma_j^2(\boldsymbol{\alpha}^T\boldsymbol{x}_i))\}, \tag{3.10}$$

where $\boldsymbol{\pi}(\cdot)=\{\pi_1(\cdot),...,\pi_{k-1}(\cdot)\}^T$, $\boldsymbol{\sigma}^2(\cdot)=\{\sigma_1^2(\cdot),...,\sigma_k^2(\cdot)\}^T$, $\boldsymbol{\beta}=\{\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_k\}^T$. Since $\boldsymbol{\pi}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ consist of nonparametric functions, (3.10) is not ready for maximization.

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

$$\ell_1^{(3)}(\boldsymbol{\pi}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j(\hat{\boldsymbol{\alpha}}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j, \sigma_j^2(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_i))\}K_h(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_i - z). \tag{3.11}$$

Let $\hat{\boldsymbol{\pi}}(\cdot)$ and $\hat{\boldsymbol{\sigma}}^2(\cdot)$ be the result of maximizing (3.11). We can then further update the estimate of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ by maximizing

$$\ell_2^{(3)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \hat{\pi}_j(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\phi(Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \hat{\sigma}_j^2(\boldsymbol{\alpha}^T\boldsymbol{x}_i))\}. \tag{3.12}$$

3.4.2 Computing Algorithm

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

Step 2: Modified EM-type algorithm to maximize $\ell_1^{(3)}$ in (3.11).

E-step:

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

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

M-step:

Update the estimate

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

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

Step 3: Update $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ by maximizing (3.12).

Step 3.1: Given $\hat{\boldsymbol{\alpha}}$, update $\boldsymbol{\beta}$.

E-step:

Calculate the expectations of component identities:

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

M-step:

Update $\boldsymbol{\beta}$:

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

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 $\hat{\boldsymbol{\beta}}$, update $\boldsymbol{\alpha}$.

Given $\hat{\boldsymbol{\beta}}$, maximize $\ell_3^{(3)}(\boldsymbol{\alpha}) = \sum_{i=1}^n \log\{\sum_{j=1}^k \hat{\pi}_j(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\phi(Y_i|\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2(\boldsymbol{\alpha}^T\boldsymbol{X}_i))\}$ to update the estimate of $\boldsymbol{\alpha}$, using some numerical methods.

Step 3.3: Iterate Steps 3.1-3.2 until convergence.

Step 4: Iterate Steps 2-3 until convergence.

Asymptotic Properties

Define $\boldsymbol{\eta} = (\boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T)^T$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$, and $\ell(\boldsymbol{\eta}(z), \boldsymbol{\theta}, \boldsymbol{x}, y) = \log\sum_{j=1}^k \pi_j(z)\phi\{y|\boldsymbol{x}^T\boldsymbol{\beta}_j, \sigma_j^2(z)\}$. Let $q_{\boldsymbol{\eta}}(z) = \frac{\partial\ell(\boldsymbol{\eta}(z),\boldsymbol{\theta},x,y)}{\partial\boldsymbol{\eta}}$, $q_{\boldsymbol{\eta}\boldsymbol{\eta}}(z) = \frac{\partial^2\ell(\boldsymbol{\eta}(z),\boldsymbol{\theta},x,y)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}$, and similarly, define $q_{\boldsymbol{\theta}}$, $q_{\boldsymbol{\theta}\boldsymbol{\theta}}$, and $q_{\boldsymbol{\eta}\boldsymbol{\theta}}$. Denote $\mathcal{I}_{\boldsymbol{\eta}}^{(3)}(z) = -E[q_{\boldsymbol{\eta}\boldsymbol{\eta}}(Z)|Z = z]$, $\Lambda_3(u|z) = E[q_{\boldsymbol{\eta}}(z)|Z = u]$

Under further conditions defined in Section 3.7, the properties of the estimator when $\boldsymbol{\theta}$ 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 3.4.2. *Assume that conditions (C1)-(C4) and (C13)-(C15) in Section 3.7 hold. Then as $n \to \infty$, $h \to 0$ and $nh \to \infty$, we have*

$$\sqrt{nh}\{\hat{\boldsymbol{\eta}}(z) - \boldsymbol{\eta}(z) - \mathcal{B}_3(z) + o_p(h^2)\} \xrightarrow{D} N\{0, \nu_0 f^{-1}(z)\mathcal{I}_\eta^{(3)}(z)\},$$

where $\mathcal{B}_3(z) = \mathcal{I}_\eta^{(3)-1}\left\{\frac{f'(z)\Lambda_3'(z|z)}{f(z)} + \frac{1}{2}\Lambda_3''(z|z)\right\}\kappa_2 h^2$.

Theorem 3.4.3. *Assume that conditions (C1)-(C4) and (C13)-(C16) in Section 3.7 hold. Then, as $n \to \infty$, $nh^4 \to 0$, and $nh^2/\log(1/h) \to \infty$,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(0, \boldsymbol{Q}_3^{-1}),$$

where,

$$\boldsymbol{Q}_3 = E\left[q_{\eta\eta}(Z)\begin{pmatrix}\boldsymbol{x}\boldsymbol{\eta}'(Z)\\ \boldsymbol{I}\end{pmatrix}\left\{\begin{pmatrix}\boldsymbol{x}\boldsymbol{\eta}'(Z)\\ \boldsymbol{I}\end{pmatrix} - \begin{pmatrix}\mathcal{I}_{\boldsymbol{\eta}}^{(3)-1}(Z)E\{q_{\boldsymbol{\eta}\boldsymbol{\eta}}(Z)(\boldsymbol{x}\boldsymbol{\eta}'(Z))^T|Z\}\\ \mathcal{I}_{\boldsymbol{\eta}}^{(3)-1}(Z)E\{q_{\boldsymbol{\eta}\boldsymbol{\eta}}(Z)|Z\}\end{pmatrix}\right\}^T\right].$$

3.5 Numerical Studies

3.5.1 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 (3.1) 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 equally spaced grid points 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.

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 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

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

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 bandwidths 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 under-smoothing, appropriate smoothing and over-smoothing condition, respectively. We also tried the likelihood based cross-validation, the results are similar.

Table 3.1 reports the MSEs of $\hat{\boldsymbol{\alpha}}$ (value times 100). From Table 3.1, 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 3.2 contains the mean and standard deviation of $RASE_\pi$, $RASE_m$, and $RASE_{\sigma^2}$. We see that the fully iterative estimates provide better results than one-step estimate under appropriate-smoothing condition, and the fully iterative estimate is not sensitive to initial values.

Table 3.1: *MSE of $\hat{\boldsymbol{\alpha}}$ (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

Table 3.2: *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	m	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	m	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	m	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)

Example 2. We conduct a simulation for a 2-component MRSIP:

$$\pi_1(z) = 0.5 - 0.35\sin(\pi z) \text{ and } \pi_2(z) = 1 - \pi_1(z),$$
$$m_1(\boldsymbol{x}) = 1 + 3x_2 \text{ and } m_2(\boldsymbol{x}) = -1 + 2x_1 + 3x_3,$$
$$\sigma_1^2 = 0.7 \text{ and } \sigma_2^2 = 0.6,$$

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 3.3 reports the MSEs of parameter estimates, and Table 3.4 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 3.3: *The MSEs of parameters (the values are 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

Example 3. We conduct a simulation for a 2-component MRSIPV:

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

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 $\boldsymbol{\alpha} = (1, 1, 1)/\sqrt{3}$. MRSIPV 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 3.5 reports the MSEs of $\hat{\boldsymbol{\beta}}$s, and Table 3.6 contains the MSEs of $\hat{\boldsymbol{\alpha}}$, the average of $RASE_{\pi}$, and $RASE_{\sigma^2}$. From both tables, we can see that MRSIPV provides much better estimates compared to MixLinReg, both in the estimation of global parameters and the nonparametric functions.

Table 3.4: *The MSEs of direction parameter and the average of $RASE_\pi$ (the values are 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

3.5.2 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 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 CV procedure. Figure 3.1 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

Table 3.5: *The MSEs of parameters (values times 100)*

		β_{10}	β_{11}	β_{12}	β_{13}	β_{20}	β_{21}	β_{22}	β_{23}
$n = 200$	MRSIPV(S)	20.65	13.56	18.56	15.77	5.473	4.128	7.664	4.243
	MRSIPV(T)	17.65	12.83	18.24	14.19	5.093	3.956	6.640	3.792
$h = 0.115$	MixLinReg	31.54	15.38	20.87	20.12	9.134	4.710	7.652	4.641
$n = 400$	MRSIPV(S)	9.192	5.968	9.448	5.822	2.589	1.875	2.901	1.975
	MRSIPV(T)	8.446	6.132	9.567	5.253	2.459	1.785	3.034	1.981
$h = 0.085$	MixLinReg	24.08	8.471	15.72	11.83	6.147	2.349	5.233	2.979
$n = 800$	MRSIPV(S)	3.004	2.287	3.781	2.338	1.303	0.987	0.954	1.007
	MRSIPV(T)	3.166	2.206	4.436	2.380	1.259	0.915	0.966	0.945
$h = 0.062$	MixLinReg	19.01	4.567	12.46	8.242	5.637	1.992	4.203	2.168

(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 models and compared them with some existing 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 $n - s$) and test subsets (with size s). The mean squared prediction error evaluated at the test data sets over 500 replications are reported as boxplots in Figure 3.2. Apparently, the three new models that we proposed have superior prediction power than the linear regression model, the mixture of linear regression models, and the single-index models. In addition, MSIM is more favorable than the MRSIP or MRSIPV for this data set.

3.6 Summary and Future Work

In this chapter we proposed three finite semiparametric mixture of regression models and the corresponding backfitting estimates. We showed that the nonparametric functions can

Table 3.6: *The MSEs of direction parameter and the average of $RASE_\pi$ and $RASE_{\sigma^2}$ (values times 100)*

		α_1	α_2	α_3	RASE_π	RASE_{σ^2}
$n = 200$	MRSIPV(S)	2.439	7.943	3.398	12.50	21.68
	MRSIPV(T)	1.392	1.637	1.590	10.75	19.25
$h = 0.115$	MixLinReg	-	-	-	22.01	30.78
$n = 400$	MRSIPV(S)	0.890	3.441	2.116	9.312	16.82
	MRSIPV(T)	0.536	0.480	0.504	7.736	15.32
$h = 0.085$	MixLinReg	-	-	-	21.80	30.93
$n = 800$	MRSIPV(S)	0.477	2.357	1.818	7.613	14.07
	MRSIPV(T)	0.217	0.229	0.242	5.920	12.42
$h = 0.062$	MixLinReg	-	-	-	21.62	30.87

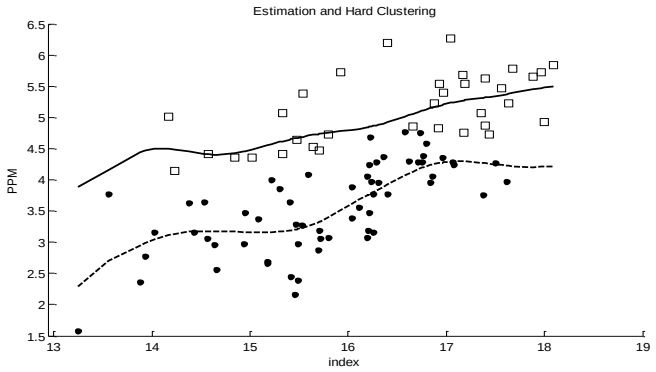

Figure 3.1: *NBA data: Estimated mean functions and a hard-clustering result.*

be estimated as if the parameters were known and the parameters can be estimated with root-n convergence rate. In this chapter, 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 also interesting to build some formal test to compare the three semiparametric mixture models. One way is to apply generalized likelihood ratio statistic proposed by Fan, et al. (2001). In case of categorical covariates, we can further explore mixture models with partial linear mean functions.

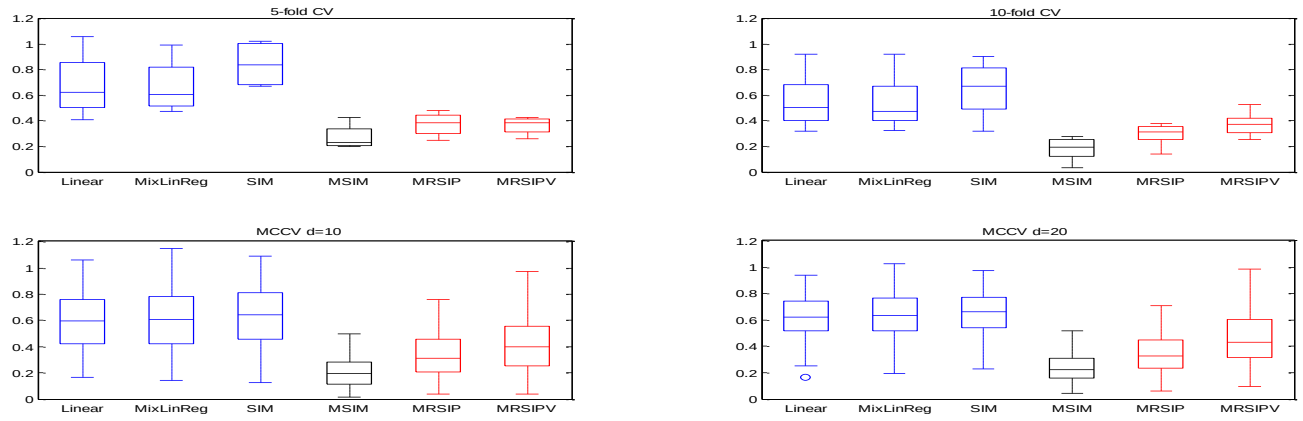

Figure 3.2: *Prediction accuracy: (a) 5-fold CV; (b) 10-fold CV;(c) MCCV s=10;(d) MCCV s=20.*

3.7 Proofs

. **Technical Conditions:**

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

(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) $\boldsymbol{\theta}''(\cdot)$ 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.

(C13) The third derivative $|\partial^3 \ell(\boldsymbol{\eta}, y)/\partial \pi_i \partial \pi_j \partial \pi_k| \leq M(y)$ for all y and all $\boldsymbol{\eta}$ in a neighborhood of $\boldsymbol{\eta}(z)$, and $E[M(y)] < \infty$.

(C14) The unknown functions $\boldsymbol{\eta}(z)$ have continuous second derivative. For $j = 1, ..., k$, $\pi_j(z) > 0$ and $\sigma_j(z) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$.

(C15) For all i and j, the following conditions hold:

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

(C16) $\boldsymbol{\eta}''(\cdot)$ is continuous at the point z.

Proof of Theorem 3.2.1.

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 3.2.2.

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

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

If $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{\sigma}}^2)^T$ maximizes (3.3), then $\hat{\boldsymbol{\theta}}^*$ maximizes

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

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

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

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 $\boldsymbol{A}_{1n} = -f(z)\mathcal{I}^{(1)}_{\boldsymbol{\theta}}(z) + o_p(1)$. Therefore,

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

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

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

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

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

Similarly, we can show that $\text{Cov}(\boldsymbol{W}_{1n}) = f(z)\mathcal{I}_{\boldsymbol{\theta}}^{(1)}(z)\nu_0 + o_p(1)$, 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 3.2.3.

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 (3.3), 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

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

By similar argument as in the previous proof,

$$\begin{aligned}\hat{\boldsymbol{\theta}}(z_0;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(z_0) = {} & n^{-1}f^{-1}(z_0)\mathcal{I}_\theta^{(1)-1}(z_0)\sum_{i=1}^n q_{1i}(Z_i)K_h(Z_i - z_0) \\ & - \mathcal{I}_\theta^{(1)-1}(z_0)E\{q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^T|Z = z_0\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + o_p(n^{-1/2}). \qquad (3.13)\end{aligned}$$

Note that

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^T\boldsymbol{x}_i) &= \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^T\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^T\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^T\boldsymbol{x}_i) \\ &= (\hat{\boldsymbol{\theta}}'(\boldsymbol{\alpha}^T\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}))^T(\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}^T)\boldsymbol{x}_i + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^T\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}_0^T\boldsymbol{x}_i) + o_p(n^{-1/2}) \\ &= (\boldsymbol{\theta}'(\boldsymbol{\alpha}^T\boldsymbol{x}_i))^T(\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}^T)\boldsymbol{x}_i + \hat{\boldsymbol{\theta}}(\boldsymbol{\alpha}^T\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) - \boldsymbol{\theta}(\boldsymbol{\alpha}^T\boldsymbol{x}_i) + o_p(n^{-1/2}), \qquad (3.14)\end{aligned}$$

where the second part is handled by (3.13).

Since $\hat{\boldsymbol{\alpha}}$ maximizes (3.4), 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 (3.14), 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 (3.13),

$$\begin{aligned}
\mathbf{0} &= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{1i}(Z_i) + n^{1/2}A_\alpha(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\
&- n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)\mathcal{I}_\theta^{-1}(Z_i)E\{q_2(Z)[\boldsymbol{x}\boldsymbol{\theta}'(Z)]^T|Z = Z_i\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\
&+ n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)n^{-1}f^{-1}(Z_i)\mathcal{I}_\theta^{-1}(Z_i)\sum_{t=1}^{n} q_{1t}(Z_t)K_h(Z_t - Z_i) + o_p(1) \\
&= \lambda\hat{\boldsymbol{\alpha}} + n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{1i}(Z_i) + \boldsymbol{Q}_1 n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\
&+ n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)q_{2i}(Z_i)n^{-1}f^{-1}(Z_i)\mathcal{I}_\theta^{(1)-1}(Z_i)\sum_{t=1}^{n} q_{1t}(Z_t)K_h(Z_t - Z_i) + o_p(1). \quad (3.15)
\end{aligned}$$

Interchanging the summations in the last term, we get

$$n^{-1/2}\sum_{i=1}^{n}\left[n^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{\theta}'(Z_t)q_{2t}(Z_t)K_h(Z_t-Z_i)f^{-1}(Z_t)\mathcal{I}_\theta^{-1}(Z_t)q_{1i}(Z_i)\right]$$
$$=n^{-1/2}\sum_{i=1}^{n}E[\boldsymbol{x}\boldsymbol{\theta}'(Z)q_2(Z)|Z_i]\mathcal{I}_\theta^{(1)-1}(Z_i)q_{1i}(Z_i)+o_p(1). \tag{3.16}$$

Let $\Gamma_\alpha = I-\boldsymbol{\alpha}\boldsymbol{\alpha}^T+o_p(1)$. Combining (3.15) and (3.16), and multiply by Γ_α, we have

$$\Gamma_\alpha\boldsymbol{Q}_1 n^{1/2}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})=n^{-1/2}\sum_{i=1}^{n}\Gamma_\alpha\{\boldsymbol{x}_i\boldsymbol{\theta}'(Z_i)+E[\boldsymbol{x}\boldsymbol{\theta}'(Z)q_2(Z)|Z_i]\mathcal{I}_{\boldsymbol{\theta}}^{(1)-1}(Z_i)\}q_{1i}(Z_i)+o_p(1) \tag{3.17}$$

It can be shown that the right-hand side of (3.17) has the covariance matrix $\Gamma_\alpha\boldsymbol{Q}_1\Gamma_\alpha$, and therefore, completes the proof. □

Proof of Theorem 3.3.1

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 3.3.2

This proof is similar to the proof of Theorem 3.2.2.

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

Proof of Theorem 3.3.3

The proof is similar to the proof of Theorem 3.2.3. It can be shown that

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

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

and therefore,

$$\hat{\boldsymbol{\pi}}(\hat{Z}_i; \hat{\boldsymbol{\lambda}}) - \boldsymbol{\pi}(Z_i) = \{\boldsymbol{x}_i \boldsymbol{\pi}'(Z_i)\}^T (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \hat{\boldsymbol{\pi}}(Z_i; \hat{\boldsymbol{\lambda}}) - \boldsymbol{\pi}(Z_i) + o_p(n^{-\frac{1}{2}}). \tag{3.18}$$

Since $\hat{\boldsymbol{\lambda}}$ maximizes (3.8), it is the solution to

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

where γ is the Lagrange multiplier. By Taylor series and (3.18)

$$\begin{aligned}
\mathbf{0} =& \gamma \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \mathbf{0} \end{pmatrix} + n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{1i} q_{\boldsymbol{\pi} i}(Z_i) + n^{\frac{1}{2}} \boldsymbol{Q}_2 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix} \\
& + n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{1i} q_{\boldsymbol{\pi}\boldsymbol{\pi} i}(Z_i) n^{-1} f^{-1}(Z_i) \mathcal{I}_{\boldsymbol{\pi}}^{(2)-1}(Z_i) \sum_{j=1}^{n} q_{\boldsymbol{\pi} j}(Z_j) K_h(Z_j - Z_i) + o_p(1) \\
=& \gamma \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \mathbf{0} \end{pmatrix} + n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{1i} q_{\boldsymbol{\pi} i}(Z_i) + n^{\frac{1}{2}} \boldsymbol{Q}_2 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix} \\
& + n^{-\frac{1}{2}} \sum_{i=1}^{n} E[\boldsymbol{\Lambda}_{1i} q_{\boldsymbol{\pi}\boldsymbol{\pi}}(Z_i)] \mathcal{I}_{\boldsymbol{\pi}}^{(2)-1}(Z_i) q_{\boldsymbol{\pi} i}(Z_i) + o_p(1). \qquad (3.19)
\end{aligned}$$

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 $\Gamma_\alpha = \begin{pmatrix} \boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\alpha}^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} + o_p(1)$. By (3.19), and multiply by Γ_α, we have

$$n^{\frac{1}{2}} \Gamma_\alpha \boldsymbol{Q}_2 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \Gamma_\alpha \left\{ \boldsymbol{\Lambda}_{1i} - \mathcal{I}_{\boldsymbol{\pi}}^{(2)-1}(Z_i) E[\boldsymbol{\Lambda}_{1i}(Z_i) q_{\boldsymbol{\pi}\boldsymbol{\pi}}(Z_i) | Z_i] \right\} q_{\boldsymbol{\pi} i}(Z_i) + o_p(1). \qquad (3.20)$$

It can be shown that the right-hand side of (3.20) has the covariance matrix $\Gamma_\alpha \boldsymbol{Q}_2 \Gamma_\alpha$, and thus, completes the proof. □

Proof of Theorem 3.4.1

Ichimura (1993) have shown that under conditions (i)-(iv), $\boldsymbol{\alpha}$ is identifiable. Furthermore, with condition (v), $(\boldsymbol{\pi}(\cdot), \boldsymbol{\beta}, \boldsymbol{\sigma}^2(\cdot))$ are identifiable. Thus completes the proof. □

Proof of Theorem 3.4.2

This proof is similar to the proof of Theorem 3.2.2

Let $\hat{\pi}_j^* = \sqrt{nh}\{\hat{\pi}_j - \pi_j(z)\}$, $j = 1, ..., k-1$, and $\hat{\sigma}_j^{2*} = \sqrt{nh}\{\hat{\sigma}_j^2 - \sigma_j^2(z)\}$, $j = 1, ..., k$. Define $\hat{\boldsymbol{\pi}}^* = (\hat{\pi}_1^*, ..., \hat{\pi}_{k-1}^*)^T$, $\hat{\boldsymbol{\sigma}}^{2*} = (\hat{\sigma}_1^{2*}, ..., \hat{\sigma}_k^{2*})^T$, and $\hat{\boldsymbol{\eta}}^* = (\hat{\boldsymbol{\pi}}^{*T}, (\hat{\boldsymbol{\sigma}}^{2*})^T)^T$. It can be shown that

$$\hat{\boldsymbol{\eta}}^* = f(z)^{-1}\mathcal{I}_{\boldsymbol{\eta}}^{(3)-1}(z)\boldsymbol{W}_{3n} + o_p(1).$$

where

$$\boldsymbol{W}_{3n} = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\frac{\partial \ell(\boldsymbol{\eta}(z), \hat{\boldsymbol{\theta}}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\eta}}K_h(\hat{Z}_i - z).$$

To complete the proof, notice that

$$\begin{aligned} E(\boldsymbol{W}_{3n}) =& \sqrt{nh}E\left\{E[\frac{\partial \ell(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{x}_i, Y_i)}{\partial \boldsymbol{\eta}}K_h(Z_i - z)|Z = z_0]\right\} \\ =& \sqrt{nh}[\frac{1}{2}f(z)\Lambda_3''(z|z) + f'(z)\Lambda_3'(z|z)]\kappa_2 h^2. \end{aligned}$$

and $\text{Cov}(\boldsymbol{W}_{3n}) = f(z)\mathcal{I}_{\boldsymbol{\eta}}^{(3)}(z)\nu_0 + o_p(1)$. The rest of the proof follows a standard argument. □

Proof of Theorem 3.4.3

The proof is similar to the proof of Theorem 3.2.3. It can be shown that

$$\begin{aligned} &\hat{\boldsymbol{\eta}}(z_0; \hat{\boldsymbol{\theta}}) - \boldsymbol{\eta}(z_0) = n^{-1}f^{-1}(z_0)\mathcal{I}_{\boldsymbol{\eta}}^{(3)-1}(z_0)\sum_{i=1}^{n} q_{\boldsymbol{\eta}i}(Z_i)K_h(Z_i - z_0) \\ &-\mathcal{I}_{\boldsymbol{\eta}}^{(3)-1}(z_0)E\{q_{\boldsymbol{\eta}\boldsymbol{\eta}}(Z)[\boldsymbol{x}\boldsymbol{\eta}'(Z)]^T|Z = z_0\}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - \mathcal{I}_{\boldsymbol{\eta}}^{(3)-1}(z_0)E\{q_{\boldsymbol{\eta}\boldsymbol{\beta}}(Z)|Z = z_0\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(n^{-1/2}), \end{aligned}$$

and therefore,

$$\hat{\boldsymbol{\eta}}(\hat{Z}_i; \hat{\boldsymbol{\theta}}) - \boldsymbol{\eta}(Z_i) = \{\boldsymbol{x}_i\boldsymbol{\eta}'(Z_i)\}^T(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \hat{\boldsymbol{\eta}}(Z_i; \hat{\boldsymbol{\theta}}) - \boldsymbol{\eta}(Z_i) + o_p(n^{-\frac{1}{2}}). \tag{3.21}$$

Since $\hat{\boldsymbol{\theta}}$ maximizes (3.12), it is the solution to

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

where γ is the Lagrange multiplier. By Taylor series and (3.21)

$$\begin{aligned}
\mathbf{0} =& \gamma \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \mathbf{0} \end{pmatrix} + n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{2i} q_{\boldsymbol{\eta}i}(Z_i) + n^{\frac{1}{2}} \boldsymbol{Q}_3 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{pmatrix} \\
& + n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{2i} q_{\boldsymbol{\eta}\boldsymbol{\eta}i}(Z_i) n^{-1} f^{-1}(Z_i) \mathcal{I}_{\boldsymbol{\eta}}^{(3)-1}(Z_i) \sum_{j=1}^{n} q_{\boldsymbol{\eta}j}(Z_j) K_h(Z_j - Z_i) + o_p(1) \\
=& \gamma \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \mathbf{0} \end{pmatrix} + n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\Lambda}_{2i} q_{\boldsymbol{\eta}i}(Z_i) + n^{\frac{1}{2}} \boldsymbol{Q}_3 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{pmatrix} \\
& + n^{-\frac{1}{2}} \sum_{i=1}^{n} E[\boldsymbol{\Lambda}_{2i} q_{\boldsymbol{\eta}\boldsymbol{\eta}}(Z_i)] \mathcal{I}_{\boldsymbol{\eta}}^{(3)-1}(Z_i) q_{\boldsymbol{\eta}i}(Z_i) + o_p(1). \qquad (3.22)
\end{aligned}$$

where $\boldsymbol{\Lambda}_{2i} = \begin{pmatrix} \boldsymbol{x}_i \boldsymbol{\eta}'(Z_i) \\ \mathbf{I} \end{pmatrix}$, 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 (3.22), and multiply by Γ_α, we have

$$n^{\frac{1}{2}} \Gamma_\alpha \boldsymbol{Q}_3 \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{pmatrix} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \Gamma_\alpha \left\{ \boldsymbol{\Lambda}_{2i} - \mathcal{I}_{\boldsymbol{\eta}}^{(2)-1}(Z_i) E[\boldsymbol{\Lambda}_{2i}(Z_i) q_{\boldsymbol{\eta}\boldsymbol{\eta}}(Z_i) | Z_i] \right\} q_{\boldsymbol{\eta}i}(Z_i) + o_p(1). \qquad (3.23)$$

It can be shown that the right-hand side of (3.23) has the covariance matrix $\Gamma_\alpha \boldsymbol{Q}_3 \Gamma_\alpha$, and thus, completes the proof. □

Bibliography

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate-a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289-300.

[2] Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, 5, 445-463.

[3] Bordes, L., Delmas, C. and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics*, 33, 733-752.

[4] Botev, Z. I., Grotowski, J. F., and Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, 2010, 2916-2957.

[5] Cao, J and Yao, W (2012). Semiparametric mixture of binomial regression with a degenerate component. *Statistica Sinica*, 22, 27-46.

[6] Carroll, R. J., Fan, J., Gijbels, I., and Wand, M.P. (1997). Generalized partially Linear Single-index Models. *Journal of American Statistical Association*, 92, 477-489.

[7] 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.

[8] de Boor, C. (2001). A practical guide to splines. Springer. New York.

[9] de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probablity Theory and Related Fields*, 75, 261-277.

[10] Doetsch, G.(1928). Die elimination des dopplereffekts auf spektroskopische feinstrukturen und exakte bestimmung der komponenten. *Zeitschrift für Physik* 49, 705-730.

[11] Donoho, D. L. and Liu, D. C. (1988). The automatic robustness of minimum distance functionals. *Annals of Statistics*, 16(2), 552-586.

[12] Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its application. Chapman & Hall/ CRC.

[13] Fan, J. and Huang, T. (2005). Profile Likelihood Inference on Semiparametric Varying-Coefficient Partially Linear Models. *Bernoulli*, 11, 1031-1057.

[14] Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of Royal Statistical Society*, 65, 57-80.

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

[16] Fisher,R.A. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenic*, 7, Part II, 179-188.

[17] Frühwirth-Schnatter, S. (2001). Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of American and Statisical Association*. 96, 194-209.

[18] García-Escudero, L. A., Gordaliza, A., and Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2), 434-449.

[19] García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3), 1324-1345.

[20] García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2), 89-109.

[21] Goldfeld, S.M. and Quandt, R.E. (1973). A Markov Model for Switching Regressions. *Journal of Econometrics*, 1, 3-6.

[22] Green, P.J. and Richardson, S. (2002). A Markov Model for Switching Regression. *Journal of American and Statistical Association*, 97, 1055-1070.

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

[24] Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of aerage derivatives. *Journal of the American Statistical Association*, 84, 986-995.

[25] Hastie, T., Tibshirani, R. and Friedman, J. (2003). The Elements of Statistical Learning, Data Mining, Inference and Prediction. Springer. New York.

[26] Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, B., Borg, A. and Trent, J. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicion*, 344, 539-548.

[27] Hennig, C. (2000). Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17, 273-296.

[28] Huang, M., Li, R. and Wang, S. (2013). Nonparametric Mixture of Regression Models. *Journal of the American Statistical Association*, 108, 929-941.

[29] 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.

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

[31] Lagarias, J. C., Reeds, J. A., Wright, M. H. and Wright, P. E. (1998). Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, 9(1), 112-147.

[32] Langaas, M., Lindqvist, B. H. and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with appliation to DNA microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4), 555-572.

[33] Leroux, B. G. (1992). Consistent Estimation of a Mixing Distribution. *Annals of Statistics*, 20, 1350-1360.

[34] Li, K. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414), 316-327.

[35] Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance estimation and related Methods. *Annals of Statistics*, 22, 1081-1114.

[36] McLachlan, G.J. and Peel, D. (2000). Finite Mixture Models. Wiley, New York.

[37] McLachlan, G. J. and Wockner, L. (2010). Use of mixture models in multiple hypothesis testing with applications in Bioinformatics. In Hermann Locarek-Junge and Claus Weihs (Ed.), *Classification as a Tool for Research: Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft fr Klassifikation* (pp. 177-184) Heidelberg, Germany: Springer-Verlag.

[38] McLachlan, G. J., Bean, R. W., and Ben-Tovim Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrys. *Bioinformatics*, 22, 1608-1615.

[39] Pak, R. (1996). Minimum Hellinger distance estimation in simple linear regression models; distribution and efficiency. *Statistics & Probability Letters*, 26, 263-269.

[40] Pearson, K. (1894). Contributions to the mathematical theory of evolution. Philosophical. *Transactions of the Royal Society of London* A 185 , 71-110.

[41] Pollard, D. (1991). Asymptotics for Least Absolute Deviation Regression Estimators. *Econometric Theory*, 7, 186-199.

[42] Punzo, A. and McNicholas, P.D. (2013). Outlier detection via parsimonious mixtures of contaminated Gaussian distributions. ArXiv:1305.4669.

[43] Severini, T.A. and Staniswalis J.G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89, 501-511.

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

[45] Simpson, D. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, 82, 802-807.

[46] Simpson, D. (1989). Hellinger deviance tests - efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, 84, 107-113.

[47] Song, J. and Nicolae, D. (2009). A sequential clustering algorithm with applications to gene expression data. *Journal of the Korean Statistical Society*, 38, 175-184.

[48] Song, S., Nicolae, D.L. and Song, J. (2010). Estimating the mixing proportion in a semiparametric mixture model. *Computational Statistics and Data Analysis*, 54, 2276-2283.

[49] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 3889-3894.

[50] Stute, W. and Zhu, L. (2005). Nonparametric checks for single-index models. *The Annals of Statistics*, 33, 1048-1083.

[51] Swanepoel, J. W. H. (1999). The limiting behavior of a modified maximal symmetric 2s-spacing with applications. *Annals of Statistics*, 27, 24-35.

[52] Tamura, R. N. and Boos, D. D. (1983). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81, 223-229.

[53] Titterington, D., Smith, A., and Makov, U. (1985). Statistical Analysis of Finite Mixture Distributions. Wiley.

[54] Wang, J., Xue, L., Zhu, L. and Chong, Y.S. (2010). Estimation for a partial-linear single-index model. *Annals of Statistics*, 38, 246-274.

[55] Wedel, M. and DeSarbo, W.S. (1993). A Latent Class Binomial Logit Methodology for the Analysis of Paired Comparison Data. *Decision Sciences*, 24, 1157-1170.

[56] Wong, H., Ip, W. and Zhang, R. (2008). Varying-coefficient single-index model. *Computational Statistics & Data Analysis*, 52, 1458-1476.

[57] Woodward, W. A., Whitney, P. and Eslinger, P. W. (1995). Minimum Hellinger distance estimation of mixture proportions. *Journal of Statistical Planning and Inference*, 48, 303-319.

[58] Wu, J., Schick, A., and Karunamuni, R.J. (2011). Profile Hellinger distance estimation. Technical Report.

[59] Xia, Y. and Hardle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97, 1162-1184.

[60] Xia, Y. and Li, W.K. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, 94, 1275-1285.

[61] Xiang, L., Yau, K.K.W., Hui, Y.V. and Lee, A.H. (2008). Minimum Hellinger distance estimation for k-component poisson mixture with random effects. *Biometrics*, 64, 508-518.

[62] Xue, L. and Zhu, L. (2006). Empirical likelihood for single-index models. *Journal of Multivariate Analysis*, 97, 1295-1312.

[63] Yang, S. (2008). Minimum Hellinger distance estimation of parameter in the random censorship model. *Analysis of Statistics*, 19, 579-602.

[64] Yao, W. and Lindsay, B.G. (2009). Bayesian Mixture Labeling by Highest Posterior Density. *Journal of American Statistical Association*, 104, 758-767.

[65] Ying, Z. (1992). Minimum Hellinger-type distance estimation for censored-data. *Analysis of statistics*, 20, 1361-1390.

[66] Young, D.S. and Hunter, D.R. (2010). Mixtures of Regressions with Predictor-dependent Mixing Proportions. *Computational Statistics and Data Analysis*, 54, 2253-2266.

[67] Zhang, R., Huang, Z. and Lv, Y. (2010). Statistical inference for the index parameter in single-index models. *Journal of Multivariate Analysis*, 101, 1026-1041.

[68] Zhu, L. and Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *Journal of Royal Statistical Society*, 68, 549-570.