Mixture of Regression Models with Varying Mixing Proportions: A Semiparametric Approach

Mian Huang and Weixin Yao

Abstract

In this paper, we study a class of semiparametric mixtures of regression models, in which the regression functions are linear functions of the predictors, but the mixing proportions are smoothing functions of a covariate. We propose a one-step backfitting estimation procedure to achieve the optimal convergence rates for both regression parameters and the nonparametric functions of mixing proportions. We derive the asymptotic bias and variance of the one-step estimate, and further establish its asymptotic normality. A modified EM-type estimation procedure is investigated. We show that the modified EM algorithms preserve the asymptotic ascent property. Numerical simulations are conducted to examine the finite sample performance of the estimation procedures. The proposed methodology is further illustrated via an analysis of a real dataset.

Keywords: Mixture of regression models, EM algorithm, Kernel regression, Semiparametric model, Nonparametric regression

[1]Huang Mian is Assistant Professor, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, P. R. China. Email:huang.mian@shufe.edu.cn. Huang's research is partially supported by a National Science Foundation grant DMS 0348869, a funding through Projet 211 Phase 3 of SHUFE, and Shanghai Leading Academic Discipline Project, B803. Weixin Yao is the corresponding author and Assistant Professor, Department of Statistics, Kansas State University, Manhattan, Kansas 66506. Email: wxyao@ksu.edu.

1. INTRODUCTION

Mixtures of regression models are well known as switching regression models in econometrics literature, which were introduced by Goldfeld and Quandt (1973). These models are useful to study the relationship between some interested variables coming from several unknown latent components. The model setting can be stated as follows. Let $\mathcal{C}$ be a latent class variable with $P(\mathcal{C} = c \mid X = x) = \pi_c$ for $c = 1, 2, \cdots, C$, where x is a p-dimentional vector. Given $\mathcal{C} = c$, suppose that the response y depends on x in a linear way $y = \mathbf{x}^T\boldsymbol{\beta}_c + \epsilon_c$, where $\mathbf{x} = (1, x^T)^T$, $\boldsymbol{\beta}_c = (\beta_{0c}, \beta_{1c}, \ldots, \beta_{pc})^T$, and $\epsilon_c \sim N(0, \sigma_c^2)$. Then the conditional distribution of Y given $X = x$ can be written as

$$Y|_{X=x} \sim \sum_{c=1}^{C} \pi_c N(\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2). \tag{1.1}$$

Mixture models including model (1.1) are comprehensively summarized in McLachlan and Peel (2000). Frühwirth-Schnatter (2006) and Hurn et al. (2003) focus on the Bayesian approaches for model (1.1), including the selection of number of components C. Many applications can be found in literature, i.e., in econometrics (Wedel and DeSarbo, 1993; Frühwirth-Schnatter, 2001), and in biology and epidemiology (Wang et al., 1996; Green and Richardson, 2002).

In this paper, we study a class of mixtures of regression models by allowing the mixing proportions to depend on a covariate z nonparametrically, where z can be either from x or not. Consider the analysis of a CO_2-GDP dataset published by World Resource Institute. As shown in Figure 3(a), the CO_2-GDP dataset contains two related variables of 171 countries in year 2005. The response variable is the CO_2-emission per capita in year 2005, and the predictor is the GDP per capita in the same year, measured by the current US dollars. From Figure 3(a), we can see that likely there are two homogenous groups, and thus we may consider fitting a two-component mixture of regression models for the data. The purpose of the analysis is to identify the group of countries through their development path as featured by the relationship of GDP and CO2-emission. However, we can also observe that the data are more likely from the lower group when the predictor is larger. Therefore, the mixing proportions for the two components may depend on $z = x$, which violates the constant

proportion assumption of the model (1.1).

The ideas that allow the proportions to depend on the covariates in a mixture model can be found in literature, e.g., the hierarchical mixtures of experts model (Jordan and Jacobs, 1994) in machine learning. Huang (2009) and Huang and Li (2010) proposed a fully nonparametric mixture of regression models by assuming the mixing proportions, the regression functions, and the variance functions are nonparametric functions of a covariate. Young and Hunter (2010) used kernel regression to model covariates-dependent proportions for mixture of linear regression models. In Young and Hunter (2010), mixing proportions may depend on a multivariate covariate z, however, there lacks of theoretical results, and such extension may not be very useful in practice for the reason of "curse of dimensionality".

In this paper, we systematically study the mixture of regression models with varying proportions. Since the mixing proportions are nonparametric, while the regression function and variance of each component are parametric, the proposed model indeed is a semiparametric model. Compared to the nonparametric mixture of regression models of Huang (2009) and Huang and Li (2010), the new semiparametric model offers more flexibility by combining both parametric and nonparametric information together. However, the new model poses more challenge for estimation since it contains both global parameters and nonparametric functions. To estimate the unknown smoothing function $\pi_c(z)$, we introduce kernel regression technique and local likelihood method (Fan and Gijbels, 1996). To achieve the optimal convergence rate for the global parameters $\boldsymbol{\beta}_c$s and σ_c^2s and the nonparametric functions $\pi_c(z)$s, we propose a one-step backfitting estimation procedure. A fully iterative estimation procedure is also investigated. For the mixture of regression models with varying proportions, this paper makes the following major contributions to the literature:

(a) We show that mixture of regression models with varying mixing proportions are identifiable under certain conditions.

(b) We propose a new one step backfitting estimation procedure for the proposed model. In addition, we prove that the one-step estimators for the regression coefficients and variance parameters are $\sqrt{n}$ consistent, and follow an asymptotic normal distribution;

the kernel estimates for the proportion functions based upon the root-n consistent estimates of $\boldsymbol{\beta}_c$s and σ_c^2s have the same first order asymptotic bias and variance as the kernel estimates with true values of $\boldsymbol{\beta}_c$s and σ_c^2s.

(c) We develop a fast modified EM algorithm for the estimation procedure, and show that the proposed algorithm preserves the ascent property for local likelihoods and global likelihood in an asymptotic sense.

The rest of this paper is structured as follows. We present the semiparametric mixture of regression model and the estimation procedure in Section 2. In particular, we develop a one step backfitting estimation procedure for the proposed model using modified EM algorithm and kernel regression. The asymptotic properties for the resulting estimates and the ascent properties of the proposed EM-type algorithms are investigated. Simulation studies and a real data application are presented in Section 3. In Section 4, we give some discussion. Technical conditions and proofs are given in Section 5.

2. ESTIMATION PROCEDURE AND ASYMPTOTIC PROPERTIES

2.1 The Semiparametric Mixture of Regressions

Suppose that $\{(X_i, Y_i, Z_i), i = 1, \cdots, n\}$ is a random sample from population (X, Y, Z). Throughout this paper, X is p-dimensional and Y and Z are univariate. Let $\mathcal{C}$ be a latent class variable, and assume that conditioning on $X = x, Z = z$, $\mathcal{C}$ has a discrete distribution $P(\mathcal{C} = c|X = x, Z = z) = \pi_c(z)$ for $c = 1, 2, \cdots, C-1$. Here, Z can be part of X. We assume that $\pi_c(z)$s are smooth functions of z for $c = 1, 2, \cdots, C$, and $\sum_{c=1}^{C} \pi_c(z) = 1$ for all z. Given $\mathcal{C} = c$, $X = x$, and $Z = z$, Y follows a normal distribution with mean $\mathbf{x}^T\boldsymbol{\beta}_c$ and variance σ_c^2. In other words, conditioning on $X = x$ and $Z = z$, the response variable Y follows a finite mixture of normals

$$Y|_{X=x,Z=z} \sim \sum_{c=1}^{C} \pi_c(z) N\left(\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right), \tag{2.1}$$

where $\mathbf{x} = (1, x^T)^T$. When $\pi_c(z)$s are constant, model (2.1) reduces to a finite mixture of linear regression model (Goldfeld and Quandt, 1973). So model (2.1) can be regarded as a

natural extension of traditional finite mixture of linear regression models. In this article, we will mainly consider one dimensional Z. But the method and the results proposed in this article can be easily extended to multivariate Z. However, such extension is less desirable due to the "curse of dimensionality".

Identifiability is a major concern for most mixture models. Section 3.1 of Titterington et al. (1985) provided detailed accounts of the identifiability of finite mixture of distributions. In particular, mixture of univariate normals is identifiable up to relabeling. However, identifiability of mixture of regression models does not directly follow the result of univariate normal mixture. To achieve identifiability for finite mixture of regression models, the variability of $\mathbf{x}$ can not be too small; see Hening (2000) and section 8.2.2 of Frühwirth-Schnatter (2006) for detail. For model (2.1), we have the following identifiability result. Its proof is given in Section 5.

Theorem 1 *Assume that $\pi_c(z) > 0$ are continuous functions, $c = 1, \ldots, C$, and $(\boldsymbol{\beta}_c, \sigma_c^2)$, $c = 1, \ldots, C$, are distinct pairs. In addition, assume that the domain $\mathcal{X}$ of x contains an open set in $\mathbb{R}^p$, and the domain $\mathcal{Z}$ of z has no isolated points. Then model (2.1) is identifiable.*

Denote by $\ell^*(\boldsymbol{\pi}(\cdot), \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ the log-likelihood function of the collected data $\{(X_i, Y_i, Z_i), i = 1, \cdots, n\}$. That is,

$$\ell^*(\boldsymbol{\pi}(\cdot), \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c(Z_i) \phi(Y_i | \mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2) \right\}, \tag{2.2}$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_C^T\}^T$, $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \cdots, \sigma_C^2\}^T$, and $\boldsymbol{\pi}(\cdot) = \{\pi_1(\cdot), \cdots, \pi_{C-1}(\cdot)\}^T$. Since $\boldsymbol{\pi}(\cdot)$ consists of nonparametric functions, (2.2) is not yet ready for maximization. In order to estimate this semiparametric model, we propose a one-step backfitting procedure. Specifically, we first estimate $\boldsymbol{\pi}(\cdot)$ locally by maximizing the following local likelihood function

$$\ell_1(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \phi(Y_i | \mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2) \right\} K_h(Z_i - z), \tag{2.3}$$

where $K_h(t) = h^{-1} K(t/h)$ and $K(t)$ is a kernel density function. For each local model at z, we may adapt the conventional constraints and conditions imposed on the finite mixture

of linear regressions, so that the corresponding local likelihood functions are bounded (See Hathaway, 1985).

Let $\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\sigma}}^2$ be the solution of maximizing (2.3). Then $\tilde{\pi}_c(z) = \tilde{\pi}_c, \tilde{\boldsymbol{\beta}}_c(z) = \tilde{\boldsymbol{\beta}}_c$, and $\tilde{\sigma}_c(z) = \tilde{\sigma}_c$. Since the global parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ are estimated locally, they do not have root-n consistency. To improve the efficiency, parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ can be estimated globally by maximizing the following likelihood function (2.4), which replaces $\pi_c(z)$ with its estimate $\tilde{\pi}_c(z)$ in (2.2),

$$\ell_2(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \tilde{\pi}_c(Z_i) \phi\{Y_i | \mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2\} \right\}. \tag{2.4}$$

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$ be the solution of maximizing (2.4). Their root n consistency will be established in the next section under certain regularity conditions. After getting the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$, we can further improve the estimate of $\boldsymbol{\pi}(z)$ by maximizing the following local likelihood

$$\ell_3(\boldsymbol{\pi}) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \phi(Y_i | \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2) \right\} K_h(Z_i - z). \tag{2.5}$$

Let $\hat{\pi}_c(z) = \hat{\pi}_c$ be the solution of (2.5). We refer to $\hat{\pi}_c(z)$, $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\sigma}}^2$ as the proposed one-step backfitting estimates.

In semiparametric modeling, one-step estimation procedure provides convenience for deriving asymptotic properties and achieves the optimal convergence rates for both global parameters and nonparametric regression functions. Given undersmoothing conditions we are able to estimate the parametric part in the rate of $n^{-1/2}$. In section 2.2, we will show that the one-step backfitting estimates achieve the optimal convergence rates for the parameters, and the nonparametric functions can be estimated as good as if the parameters were known.

2.2 Asymptotic Properties

In this section, we first study the sampling properties of the proposed one-step backfitting estimators $\hat{\pi}_c(z)$, $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\sigma}}^2$. We will show that the one-step estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$ are root n consistent and follow an asymptotic normal distribution. In addition, we will provide the

asymptotic bias and variance of the estimator $\hat{\boldsymbol{\pi}}(\cdot)$, and show that it has smaller asymptotic covariance compared to $\tilde{\boldsymbol{\pi}}(\cdot)$.

Let $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T, \boldsymbol{\beta}^T)^T$, $\boldsymbol{\eta} = \{(\boldsymbol{\sigma}^2)^T, \boldsymbol{\beta}^T\}^T$, and thus $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, \boldsymbol{\eta}^T)^T$. Let

$$\rho(y|x,\boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c \phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right), \quad \ell(\boldsymbol{\theta}, x, y) = \log \rho(y|x,\boldsymbol{\theta}).$$

$$q_\theta\{\boldsymbol{\theta}, x, y\} = \frac{\partial \ell(\boldsymbol{\theta}, x, y)}{\partial \boldsymbol{\theta}}, \quad q_{\theta\theta}\{\boldsymbol{\theta}, x, y\} = \frac{\partial^2 \ell(\boldsymbol{\theta}, x, y)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}.$$

Similarly, we can define $q_\eta, q_{\eta\eta}, q_{\eta\pi}$, and $q_{\pi\pi}$. Furthermore, define

$$\begin{aligned}
\mathcal{I}_\theta(z) &= -\mathrm{E}\left[q_{\theta\theta}\{\boldsymbol{\theta}(z), X, Y\}\middle| Z = z\right], \\
\mathcal{I}_\eta(z) &= -\mathrm{E}\left[q_{\eta\eta}\{\boldsymbol{\theta}(z), X, Y\}\middle| Z = z\right], \\
\mathcal{I}_\pi(z) &= -\mathrm{E}\left[q_{\pi\pi}\{\boldsymbol{\theta}(z), X, Y\}\middle| Z = z\right], \\
\mathcal{I}_{\eta\pi}(z) &= -\mathrm{E}\left[q_{\eta\pi}\{\boldsymbol{\theta}(z), X, Y\}\middle| Z = z\right],
\end{aligned}$$

and

$$\Lambda(u|z) = \mathrm{E}\left[q_\pi\{\boldsymbol{\theta}(z), X, Y\}\middle| Z = u\right],$$

where $\boldsymbol{\theta}(z) = (\boldsymbol{\pi}(z)^T, (\boldsymbol{\sigma}^2)^T, \boldsymbol{\beta}^T)^T$. Let $\hat{\boldsymbol{\eta}}$ be the one-step estimate of $\boldsymbol{\eta}$. Denote by $\boldsymbol{\psi}(x, y, z)$ the vector which consists of the first $(C-1)$ elements of $\mathcal{I}_\theta^{-1}(z)\frac{\partial}{\partial \boldsymbol{\theta}}\ell(\boldsymbol{\theta}(z), x, y)$.

Theorem 2 *Suppose that $nh^4 \to 0$, $nh^2 \log(1/h) \to \infty$, and Conditions (A)—(H) in Section 5 hold. Then we have the asymptotic normality*

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

where $B = \mathrm{E}\{\mathcal{I}_\eta(Z)\}$, and

$$\Sigma = Var\left\{\frac{\partial \ell\left(\boldsymbol{\pi}(Z), \boldsymbol{\eta}, X, Y\right)}{\partial \boldsymbol{\eta}} - \boldsymbol{\omega}(X, Y, Z)\right\},$$

where $\boldsymbol{\omega}(x, y, z) = \mathcal{I}_{\eta\pi}(z)\boldsymbol{\psi}(x, y, z)$.

Define

$$\kappa_l = \int u^l K(u)\, du \quad \text{and} \quad \nu_l = \int u^l K^2(u)\, du.$$

Theorem 3 *Assume that Conditions (A)—(H) in Section 5 hold. Then as* $n \to \infty$, $h \to 0$, $nh \to \infty$, *we have the asymptotic normality results for* $\hat{\boldsymbol{\pi}}(z)$

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

where $\mathcal{B}_\pi(z)$, *is a* $(C-1)\times 1$ *vector, with the elements taken from* $[1^{th}, \cdots, (C-1)^{th}]$ *entries of* $\mathcal{B}(z)$, *where*

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

Based on the above theorem, we can see that estimating $\boldsymbol{\eta}$ does not have first order effect on $\hat{\boldsymbol{\pi}}(z)$, which is obvious since $\hat{\boldsymbol{\pi}}(z)$ is the result of nonparametric estimation with a slower rate than $\hat{\boldsymbol{\eta}}$. Therefore, $\hat{\boldsymbol{\pi}}(z)$ is more efficient than $\tilde{\boldsymbol{\pi}}(z)$, which needs to account for the uncertainty of estimating $\boldsymbol{\eta}$.

2.3 Computing Algorithms and Their Properties

EM-type algorithm for (2.3)

We first propose a modified EM algorithm to maximize (2.3) to obtain estimates $\tilde{\boldsymbol{\pi}}(Z_i)$. In the l-th cycle of the EM algorithm iteration, we have $\boldsymbol{\beta}_c^{(l)}(\cdot)$, $\sigma_c^{2(l)}(\cdot)$, and $\pi^{(l)}(\cdot)$. In the E-step, we calculate expectation of component identities

$$r_{ic}^{(l+1)} = \frac{\pi_c^{(l)}(Z_i)\phi\{Y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)}(Z_i), \sigma_c^{2(l)}(Z_i)\}}{\sum_{c=1}^C \pi_c^{(l)}(Z_i)\phi\{Y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)}(Z_i), \sigma_c^{2(l)}(Z_i)\}}, c = 1, \ldots, C. \tag{2.6}$$

Let $\{u_1, \cdots, u_N\}$ be a set of grid points at which the unknown functions are evaluated, where N is the number of grid points. In the M-step, we update for $z \in \{u_j, j = 1, \cdots, N\}$,

$$\pi_c^{(l+1)}(z) = \frac{\sum_{i=1}^n r_{ic}^{(l+1)} K_h(Z_i - z)}{\sum_{i=1}^n K_h(Z_i - z)}, \tag{2.7}$$

$$\boldsymbol{\beta}_c^{(l+1)}(z) = (\mathbf{S}^T W_c^{(l+1)}\mathbf{S})^{-1}\mathbf{S}^T W_c^{(l+1)}\mathbf{y}, \tag{2.8}$$

$$\sigma_c^{2(l+1)}(z) = \frac{\sum_{i=1}^n w_{ic}^{(l+1)}\{Y_i - \mathbf{x}_i^T\boldsymbol{\beta}_c^{(l+1)}\}^2}{\sum_{i=1}^n w_{ic}^{(l+1)}}, \tag{2.9}$$

where $c = 1, \ldots, C, w_{ic}^{(l+1)} = r_{ic}^{(l+1)} K_h(Z_i - z)$, $W_c^{(l+1)} = \text{diag}\{w_{1c}^{(l+1)}, \cdots, w_{nc}^{(l+1)}\}$, $\mathbf{y} = (Y_1, \cdots, Y_n)^T$, and $\mathbf{S} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. Furthermore, we update $\pi_c^{(l+1)}(Z_i)$, $\boldsymbol{\beta}_c^{(l+1)}(Z_i)$, and

$\sigma_c^{2(l+1)}(Z_i)$, $i = 1, \cdots, n$ by linearly interpolating $\pi_c^{(l+1}(u_j)$, $\boldsymbol{\beta}_c^{(l+1)}(u_j)$, and $\sigma_c^{2(l+1)}(u_j)$, $j = 1, \cdots, N$, respectively. In practice, if n is not very large, we may directly set the observed $\{X_1, \cdots, X_n\}$ to be the grid points. We also set grid points to be $\{X_1, \cdots, X_n\}$ when deriving the asymptotic ascent properties for the proposed algorithm.

In (2.7), for simplicity of presentation and computation, we use the same bandwidth for all $\pi_c(z)$'s. One might use different bandwidths for $\pi_c(z)$'s to improve the estimation accuracy but with much more complexity of computation and bandwidth selection. Note that in the M-step, the nonparametric functions are estimated simultaneously at a set of grid points; thus, the classification probabilities in the E-Step can be estimated globally to avoid the label switch problem (See, for example, Stephens, 2000; Celeux et al., 2000; Yao and Lindsay, 2009). The classical EM algorithm estimates the nonparametric functions separately for a set of grid points, which makes it difficult to assign the same component labels for these estimators across all the grid points.

EM algorithm for (2.4)

Given the estimate $\tilde{\boldsymbol{\pi}}(z)$, we maximize (2.4) by a regular EM algorithm to get the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$. In the E-step, we calculate the expectation of component identities

$$r_{ic}^{(l+1)} = \frac{\tilde{\pi}_c(Z_i)\phi(Y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)}, \sigma_c^{2(l)})}{\sum_{c=1}^{C}\tilde{\pi}_c(Z_i)\phi(Y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)}, \sigma_c^{2(l)})}, c = 1, \ldots, C. \tag{2.10}$$

Then in the M-step, we update $\boldsymbol{\beta}_c$s and σ_c^2s,

$$\boldsymbol{\beta}_c^{(l+1)} = (\mathbf{S}^T R_c^{(l+1)}\mathbf{S})^{-1}\mathbf{S}^T R_c^{(l+1)}\mathbf{y}, \tag{2.11}$$

$$\sigma_c^{2(l+1)} = \frac{\sum_{i=1}^{n} r_{ic}^{(l+1)}(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}_c^{(l+1)})^2}{\sum_{i=1}^{n} r_{ic}^{(l+1)}}, \tag{2.12}$$

where $c = 1, \ldots, C, R_c^{(l+1)} = \text{diag}\{r_{1c}^{(l+1)}, \cdots, r_{nc}^{(l+1)}\}$. The ascent property of the above algorithm follows the theory of ordinary EM algorithm.

EM algorithm for (2.5)

Given $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}$, we would maximize (2.5) to obtain the estimate $\hat{\boldsymbol{\pi}}(z)$. Since $\hat{\boldsymbol{\beta}}_c$ and $\hat{\sigma}_c$ are well labeled, we can use the regular EM algorithm without worrying about the label

switching problem. In the E-step of l-th cycle, the expectation of component identities are given by

$$r_{ic}^{(l+1)}(z) = \frac{\pi_c^{(l)}(z)\phi(Y_i|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}{\sum_{c=1}^{C}\pi_c^{(l)}(z)\phi(Y_i|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}, c = 1, \ldots, C. \tag{2.13}$$

In the M-step, we update $\boldsymbol{\pi}(z)$ by

$$\pi_c^{(l+1)}(z) = \frac{\sum_{i=1}^{n} r_{ic}^{(l+1)}(z)K_h(Z_i - z)}{\sum_{i=1}^{n} K_h(Z_i - z)},\ c = 1, \ldots, C. \tag{2.14}$$

We may also use the idea of the modified EM algorithm for (2.3) to estimate $\boldsymbol{\pi}(\cdot)$ simultaneously in a set of grid points, and speed up the computation.

A computational accelerating scheme

To avoid extensive computation, many researchers prefer to using one-step estimate in semiparametric modeling, e.g., in partially linear model (Hunsberger, 1994; Severini and Staniswalis, 1994), generalized partially linear single-index model (Carroll et al., 1997), and generalized varying-coefficient partially linear model (Li and Liang, 2008). However, the fully iterated estimation procedure is of great interest if extensive computation can be avoid. Next, we discuss one approach to approximate the fully iterated estimation procedure with less computation.

In the E-step of l-th cycle,

$$r_{ic}^{(l+1)} = \frac{\pi_c^{(l)}(Z_i)\phi(Y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)}, \sigma_c^{2(l)})}{\sum_{c=1}^{C}\pi_c^{(l)}(Z_i)\phi(Y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)}, \sigma_c^{2(l)})}, c = 1, \ldots, C\,. \tag{2.15}$$

In the M-step, we simultaneously update $\boldsymbol{\beta}, \boldsymbol{\sigma}$, and $\boldsymbol{\pi}(z)$ by

$$\boldsymbol{\beta}_c^{(l+1)} = (\mathbf{S}^T R_c^{(l+1)}\mathbf{S})^{-1}\mathbf{S}^T R_c^{(l+1)}\mathbf{y}, \tag{2.16}$$

$$\sigma_c^{2(l+1)} = \frac{\sum_{i=1}^{n} r_{ic}^{(l+1)}(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)})^2}{\sum_{i=1}^{n} r_{ic}^{(l+1)}}, \tag{2.17}$$

$$\pi_c^{(l+1)}(z) = \frac{\sum_{i=1}^{n} r_{ic}^{(l+1)}K_h(Z_i - z)}{\sum_{i=1}^{n} K_h(Z_i - z)},\ z \in \{u_j, j = 1, \cdots, N\}, \tag{2.18}$$

where $c = 1, \ldots, C, R_c^{(l+1)} = \text{diag}\{r_{1c}^{(l+1)}, \cdots, r_{nc}^{(l+1)}\}$. Furthermore, we update $\pi_c^{(l+1)}(Z_i)$, $i = 1, \cdots, n$ by linearly interpolating $\pi_c^{(l+1)}(u_j)$, $j = 1, \cdots, N$.

In the following theorem, we provide the ascending properties for the EM algorithms proposed in this section. Its proof is given in Section 5.

Theorem 4 *(a) For EM type algorithm of (2.6)— (2.9), supposing* $nh \to \infty$ *as* $n \to \infty$ *and* $h \to 0$*, we have*

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

in probability, for any given point z*, where* $\ell_1(\cdot)$ *is defined in (2.3).*

(b) Each iteration of the algorithm from (2.13) to (2.14) will monotonically increase the local likelihood (2.5), i.e., $\ell_3(\boldsymbol{\pi}^{(l+1)}(z)) \geq \ell_3(\boldsymbol{\pi}^{(l)}(z))$*, for all* l*, where* $\ell_3(\cdot)$ *is given in (2.5).*

(c) The iterations of (2.15)—(2.18) have the following property:

$$\liminf_{n\to\infty} \; n^{-1} \left[\ell^*\{\boldsymbol{\pi}^{(l+1)}(\cdot), \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\sigma}^{2(l+1)}\} - \ell^*\{\boldsymbol{\pi}^{(l)}(\cdot), \boldsymbol{\beta}^{(l)}, \boldsymbol{\sigma}^{2(l)}\} \right] \geq 0 \tag{2.19}$$

in probability, where $\ell^*(\cdot)$ *is defined in (2.2).*

Theorem 4 (a) implies that when the sample size n is large enough, the algorithm of (2.6)–(2.9) possesses the ascent property for $\ell_1\{\boldsymbol{\theta}(z)\}$ at any given z. Theorem 4 (c) implies that the iterations of (2.15)—(2.18) possess similar asymptotic ascent property for the global log-likelihood (2.2).

3. SIMULATION AND APPLICATION

In this section, we conduct simulation studies to test the performance of the proposed methodologies. The performance of the estimates of the mixing proportion functions $\pi_c(z)$s is measured by the square root of the average square errors (RASE),

$$\text{RASE}_\pi^2 = N^{-1} \sum_{c=1}^{C-1} \sum_{j=1}^{N} \{\hat{\pi}_c(u_j) - \pi_c(u_j)\}^2,$$

where $\{u_j, j = 1, \cdots, N\}$ are the grid points at which the unknown functions $\pi_c(\cdot)$ are evaluated. In simulation, we set $N = 100$. The same set of grid points are used for the

algorithm proposed in Section 2.3. For simplification, the grid points are taken evenly on the range of the z-variable.

To apply our proposed methodologies, we need to first select a proper bandwidth for estimating $\boldsymbol{\pi}(\cdot)$. In practice, data driven methods can be used for bandwidth selection, such as cross-validation (CV). Denote by $\mathcal{D}$ as the full data set. We then partition $\mathcal{D}$ into a training set $\mathcal{R}_j$ and a test set $\mathcal{T}_j$, i.e., $\mathcal{D} = \mathcal{T}_j \cup \mathcal{R}_j$ for $j = 1, \cdots, J$. We use the training set $\mathcal{R}_j$ to obtain the estimates $\{\hat{\pi}_c(\cdot), \hat{\sigma}_c^2, \hat{\boldsymbol{\beta}}_c\}$. Then we can estimate $\pi_c(z)$ for the data points belonging to the corresponding test set. For $(x_l, y_l, z_l) \in \mathcal{T}_j$,

$$\hat{\pi}_c(z_l) \quad = \quad \frac{\sum_{\{i:Z_i \in \mathcal{R}_j\}} r_{ic} K_h(Z_i - z_l)}{\sum_{\{i:Z_i \in \mathcal{R}_j\}} r_{ic}}.$$

Based on the estimated $\hat{\pi}_c(z_l)$ of test set $\mathcal{T}_j$, we consider a likelihood version CV, which is given by

$$CV = \sum_{j=1}^{J} \sum_{l \in \mathcal{T}_j} \log \left\{ \sum_{q=1}^{C} \hat{\pi}_q(z_l) \phi(y_l | \mathbf{x}_l^T \hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2) \right\}. \tag{3.1}$$

In practice, we usually set the value of J to be 5 or 10, and randomly partition the data. Since different random partitions may lead to different selected bandwidth, we suggest repeating the procedure 30 times, and taking the average of the selected bandwidth as the optimal bandwidth. Note that the required under-smoothing conditions for the proposed procedure are $nh^4 \to 0$ and $nh^2 \log(1/h) \to \infty$ in order to get the root n consistency for the global parameters. The optimal bandwidth $\hat{h}$ selected by CV will be of order $n^{-1/5}$, which does not satisfy the under-smoothing conditions. As suggested by Li and Liang (2008), a good adjusted bandwidth is given by $\tilde{h} = \hat{h} \times n^{-2/15} = O(n^{-1/3})$. This bandwidth satisfies the under-smoothing requirement. In our simulation study, both cases of appropriate smoothing and under-smoothing will be investigated.

When fitting a mixture of regression model with varying proportions, it is natural to ask whether the mixing proportions actually depend on the covariates. This leads to the following testing hypothesis problem:

$$H_0 : \pi_c(z) \equiv \pi_c, c = 1, \cdots, C-1.$$

Denote by $\ell^*(H_0)$ and $\ell^*(H_1)$ the log-likelihood functions computed under null and alternative hypothesis, respectively. Then we can construct a likelihood ratio test statistic

$$T = 2\{\ell^*(H_1) - \ell^*(H_0)\}.$$

This likelihood ratio is different from the parametric likelihood ratio, since the alternative is a semiparametric model, and the number of parameters under H_1 is undefined. One approach is to study the asymptotic distribution of T. Alternatively, here we consider the conditional bootstrap method (Cai et al., 2000) to construct the null distribution. Let $\{\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\sigma}}^2\}$ be the MLE under null hypothesis. For given x_i, we can generate Y_i^* from the distribution $\sum_{c=1}^{C} \bar{\pi}_c N(\mathbf{x}_i^T \bar{\boldsymbol{\beta}}_c, \bar{\sigma}_c^2)$. For each bootstrap sample, we calculate the test statistics T, and then obtain its approximate distribution. If the asymptotic null distribution is independent of the nuisance parameters $\pi_c, c = 1, \cdots, C-1$, then the conditional bootstrap method is valid. Although a solid theoretical research is out of the scope in this paper, we investigate the Wilk's phenomenon (Fan et al., 2001) via Monte Carlo simulation. Our simulation results show that the Wilk's type of results continue to hold for the proposed model (2.1). Therefore, the conditional bootstrap method is applicable. This provides a convenience way to conduct the likelihood ratio test for the above testing problem.

In addition, we use a bootstrap procedure to construct confidence intervals for the parameters and point-wise confidence intervals for the proportion functions. For given covariates, the response variable Y_i^* can be generated from the distribution $\sum_{c=1}^{C} \hat{\pi}_c(z_i) N(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)$. We apply the proposed estimation procedure to each of the bootstrap samples, and further obtain the confidence intervals. The bootstrap approach to construct confidence intervals for nonparametric regression has been studies by many authors, such as Härdle and Bowman (1988), Härdle and Marron (1991), Eubank and Speckman (1993), Neumann and Polzehl (1998), Xia (1998), and Claeskens and Van Keilegom (2003). It is well known that theoretically the traditional bootstrap fails for kernel estimates when the bandwidth is chosen to be of order $n^{-1/5}$ (Davison and Hinkley (1997), page 226). To account for bias, Härdle and Bowman (1988) proposed to adjust the constructed interval using an estimated bias; Härdle and Marron (1991) proposed to estimate the simulation model curve by over-smoothing and

then smooth the bootstrapped data using the appropriate smoothing; Neumann and Polzehl (1998) proposed to use only one under-smoothing bandwidth for the whole procedure. Our simulation studies will investigate the under-smoothing, appropriate smoothing, and over-smoothing situations.

Example 1. In the following example, we conduct a simulation for a 2-component mixture of regression model with varying mixing proportions:

$$\begin{aligned}
&\pi_1(x) = 0.1 + 0.8\sin(\pi x) \text{ and } \pi_2(x) = 1 - \pi_1(x),\\
&m_1(x) = 4 - 2x \text{ and } m_2(x) = 3x,\\
&\sigma_1^2 = 0.09 \text{ and } \sigma_2^2 = 0.16,
\end{aligned}$$

where $m_1(x)$ and $m_2(x)$ are the regression functions for the first and second components, respectively. Therefore, in this example, $z = x$, $\boldsymbol{\beta}_1 = (4, -2)$, and $\boldsymbol{\beta}_2 = (0, 3)$. The sample sizes $n = 200$ and 400 were conducted with 500 replicates. The predictor x was generated from one dimensional uniform distribution in $[0, 1]$. The Epanechnikov kernel is used in our simulation. The selected bandwidth was obtained from the following strategy: we first generate several simulation datasets for a given sample size, and then apply the CV bandwidth selector to determine the optimal bandwidth for each dataset. The selected bandwidth, denoted by $\hat{h}$, was the average of these CV bandwidths with rounding. In the simulation, we consider three different bandwidths: $\hat{h} \times n^{-2/15}$, $\hat{h}$, $2\hat{h}$, which correspond to the under-smoothing, appropriate smoothing, and over-smoothing, respectively. It was shown that the asymptotic distribution of the non-parametric functional estimates does not have to account for the variability due to the estimation of the parametric components. We examine this via simulation studies in finite samples. In the tables, the line marked with "M1" gives the results given by the proposed method, while "M2" gives the results assuming $\boldsymbol{\eta}$ were known.

Table 1 displays the MSE of regression parameter estimates and the average of RASE_π over 500 simulations (the values are times 100). For comparison, we also report the results based on the fully parametric mixture of linear regression model (denoted by "PAR" in Table 1), which assumes the mixing proportions are constant. From Table 1, we can see that

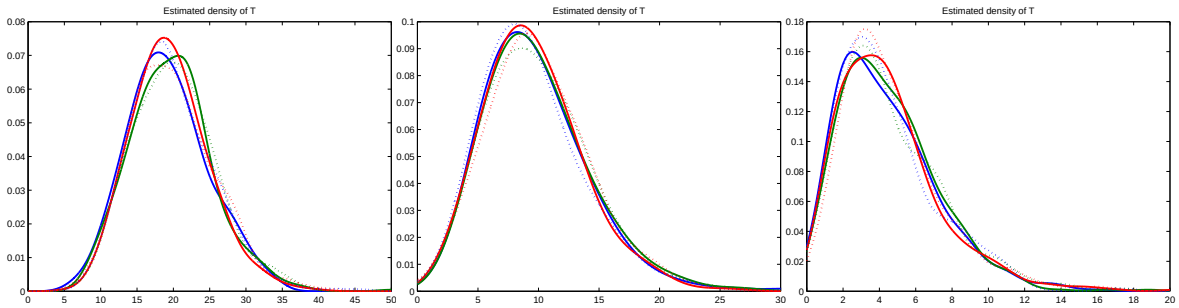
the proposed procedure gives better results compared to mixture of linear regression models, e.g., RASE_π, and the MSE of $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$ are significantly reduced. In addition, it can be seen that the proposed procedure for estimating the nonparametric function $\hat{\boldsymbol{\pi}}(\cdot)$ works almost as well as if the true value of $\boldsymbol{\eta}$ were known and works better if it is not under-smoothing.

Table 1: The averages of MSEs of parameters and RASE_π (the values are times 100)

	bandwidth ($n = 200$)				*bandwidth* ($n = 400$)			
MSE	0.04	0.08	0.16	PAR	0.03	0.07	0.14	PAR
β_{10}	0.568	0.554	0.550	0.726	0.274	0.267	0.266	0.374
β_{11}	2.290	2.176	2.156	3.840	1.151	1.113	1.122	2.396
β_{20}	0.641	0.638	0.635	0.648	0.295	0.293	0.297	0.320
β_{21}	2.587	2.392	2.382	4.237	1.114	1.026	1.079	3.156
σ_1^2	0.018	0.017	0.017	0.017	0.010	0.011	0.010	0.010
σ_2^2	0.089	0.086	0.086	0.095	0.040	0.040	0.040	0.048
$RASE_\pi$								
M1	14.61	10.71	9.722	25.93	12.32	8.304	7.613	25.73
M2	14.14	10.13	9.143	–	11.83	7.841	7.034	–

Table 2 summarizes the performance of the bootstrap method for the standard errors of estimate of parameters. The standard deviation of 500 estimates, denoted by SD, can be viewed as the true standard errors. To test the accuracy of the the proposed standard error estimate via bootstrap method, we calculated the average and standard deviation of the 500 estimated standard errors, denoted by SE and STD. The coverage probabilities for all the parameters are obtained based on the estimated standard errors. From the results, we find that the proposed bootstrap procedure estimates the true standard deviation quite well, and the coverage probabilities are close to the nominal level for most of cases. However, with moderate n, the coverage levels are a bit low for σ_1 and σ_2.

The bootstrap procedure also enables us to investigate the point-wise coverage probabilities for the proportion functions. For a set of grid points evenly distributed in the support of x, Table 3 shows the results at the level of 95% for both "M1" and "M2". For most



(a) (b) (c)

Figure 1: The estimated density of unconditional null distributions of T (solid lines), and the estimated density of conditional null distributions of T (dotted lines); the bandwidth is 0.04, 0.08, 0.16 in (a) ,(b), and (c), respectively.

points, the cases of under-smoothing and appropriate smoothing give better performance than over-smoothing case. However, for $n = 200$ the coverage levels are a bit low for point 0.5, but a bit high and thus conservative for points 0.7 and 0.8. In addition, based on Table 2 and Table 3, we can see that the over-smoothing does not provide very satisfactory coverage levels.

We next conduct a simulation to investigate whether the Wilk's type of phenomenon holds for the proposed model. Under the null hypothesis H_0, the mixing proportion π_1 is a constant. For 3 different values of $\pi_1 \in \{0.25, 0.5, 0.75\}$, we compute the unconditional null distribution with $n = 200$ via 500 Monte Carlo simulations. The resulting 3 densities were very close, as plotted in solid lines in Figure 1. This suggests that the asymptotic distribution of T under the null hypothesis was not sensitive to the true value of $\boldsymbol{\pi}$. To validate the conditional bootstrap method, we select 3 typical samples generated from the 3 values of π_1s. For each typical sample, we compute the conditional null distribution based on its 500 bootstrap samples. The resulting 3 densities were depicted as dotted curves in the same figures. From Figure 1, we can see that our conditional bootstrap method worked reasonably well to approximate the true null distribution.

The power of the proposed test is also of interest. We evaluate the power function under

Table 2: Standard errors and coverage probabilities

	SD	SE(STD)	95%	SD	SE(STD)	95%
	$n = 200, h = 0.04$			$n = 400, h = 0.03$		
β_{10}	0.074	0.069(0.008)	94.00	0.050	0.049(0.004)	94.20
β_{11}	0.154	0.142(0.019)	92.00	0.103	0.100(0.010)	93.80
β_{20}	0.079	0.078(0.010)	94.60	0.060	0.055(0.005)	94.20
β_{21}	0.151	0.153(0.024)	94.60	0.111	0.107(0.012)	93.80
σ_1	0.022	0.021(0.002)	87.60	0.015	0.015(0.001)	93.20
σ_2	0.037	0.036(0.004)	91.80	0.027	0.026(0.002)	92.20
	$n = 200, h = 0.08$			$n = 400, h = 0.07$		
β_{10}	0.074	0.069(0.008)	93.00	0.050	0.049(0.004)	94.20
β_{11}	0.151	0.140(0.019)	92.60	0.100	0.099(0.009)	93.80
β_{20}	0.079	0.079(0.010)	95.00	0.059	0.056(0.005)	93.80
β_{21}	0.148	0.153(0.024)	94.80	0.106	0.106(0.012)	94.60
σ_1	0.023	0.021(0.002)	88.00	0.015	0.015(0.001)	93.80
σ_2	0.036	0.036(0.004)	92.40	0.027	0.025(0.002)	91.60
	$n = 200, h = 0.16$			$n = 400, h = 0.14$		
β_{10}	0.073	0.066(0.007)	90.60	0.049	0.047(0.004)	92.40
β_{11}	0.149	0.131(0.016)	90.80	0.099	0.094(0.008)	91.80
β_{20}	0.079	0.080(0.010)	95.60	0.058	0.056(0.005)	93.60
β_{21}	0.143	0.156(0.025)	95.40	0.100	0.108(0.012)	94.00
σ_1	0.022	0.021(0.002)	90.40	0.015	0.015(0.001)	94.40
σ_2	0.036	0.036(0.004)	92.20	0.027	0.025(0.002)	91.40

Table 3: The pointwise coverage probabilities

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	$n = 200, h = 0.04$								
M1	96.90	96.70	95.80	92.70	88.60	94.10	98.80	100.00	97.70
M2	96.80	96.40	97.20	92.40	87.20	93.00	98.40	100.00	97.60
	$n = 200, h = 0.08$								
M1	97.40	97.10	97.40	96.20	95.80	96.60	97.80	99.30	97.70
M2	97.80	96.40	97.80	96.20	94.40	95.00	98.20	98.60	97.20
	$n = 200, h = 0.16$								
M1	91.00	96.40	95.50	95.00	91.30	90.40	96.20	97.40	79.20
M2	92.40	96.20	97.60	95.00	91.80	93.40	96.00	96.80	85.20
	$n = 400, h = 0.03$								
M1	96.60	97.20	96.20	94.80	91.80	95.60	98.80	100.00	96.40
M2	96.60	97.20	96.20	94.80	91.60	95.00	98.80	100.00	97.60
	$n = 400, h = 0.07$								
M1	97.60	96.60	97.20	98.00	95.60	97.40	99.20	99.20	96.80
M2	97.60	96.60	97.20	98.00	96.20	97.20	98.80	99.40	98.40
	$n = 400, h = 0.14$								
M1	90.80	95.10	96.20	92.20	87.70	84.90	92.80	97.90	75.40
M2	91.40	94.60	96.60	94.00	91.40	90.80	95.20	97.20	85.00

a sequence of local alternatives indexed by λ:

$$H_0 : \pi_1(x) \equiv \pi_1 \qquad vs \qquad H_1 : \pi_1(x) = 0.1 + 0.8\lambda \sin(\pi x)/\sqrt{nh},$$

and $\pi_2(x) = 1 - \pi_1(x)$, where $\lambda/\sqrt{nh} \in [0, 1]$. In Figure 2, we plot three power functions at three different significance levels: 0.10, 0.05, and 0.01, based on 500 simulations for sample size $n = 200, 400$. The results show that the powers increase rapidly as λ increases. When $\lambda = 0$, the alternative collapses into the null hypothesis, and the powers at $\lambda = 0$ for the three significance levels are close to the nominal level. This shows that the proposed bootstrap method approximately provides the right levels of the test.

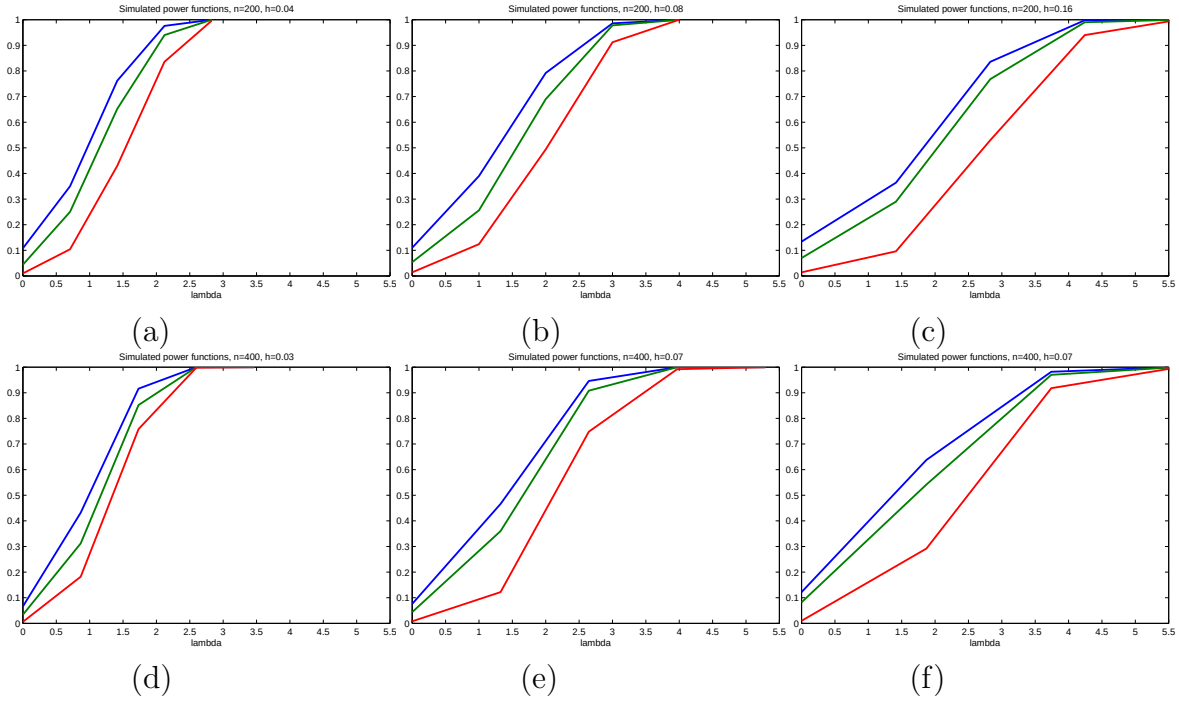
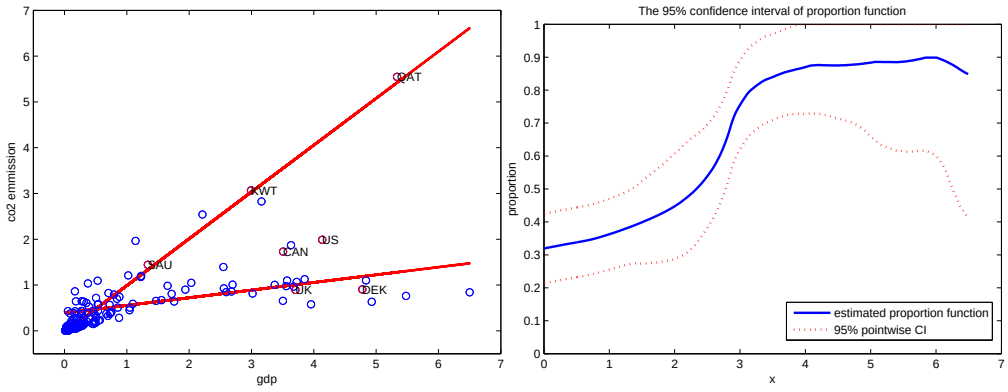


Figure 2: The power functions of the test against local alternatives; (a) $n = 200, h = 0.04$; (b) $n = 200, h = 0.08$; (c) $n = 200, h = 0.16$; (d) $n = 400, h = 0.03$; (e) $n = 400, h = 0.07$; (f) $n = 400, h = 0.14$.

Example 2. CO2-GDP Data Application.

We illustrate the proposed methodology by an analysis of the CO2-GDP Data described in Section 1. This dataset was published by World Resource Institute. We know that GDP is a measure of the size of a nation's economy, and Carbon dioxide (CO2) is an important greenhouse gas which causes the greenhouse effect and may relate to global warming. Development with high GDP per capita and relative low CO2-emission is a desired goal and consensus for modern governments. It is of interest to study the relationship between a country's CO2-emission from its industrial activities and the economy size per capita. In the analysis, we set CO2-emission per capita (Y) to be the response variable, and the GDP per capita (X) to be predictor. Note that both variables have positive observed values. We divide Y by 10000 and divide X by 10, so that they have comparable numerical scale.

For this dataset, we consider a two-component mixture of regression models with varying



(a) (b)

Figure 3: (a) The CO_2-GDP data, year 2005. y: CO_2 emission per capita; x: GDP per capita; (b) The estimated proportion function of the lower component and confidence interval.

Table 4: Estimated parameters and confidence intervals

	estimate	Bootstrap CI	estimate	Bootstrap CI	estimate	Bootstrap CI
	$h = 1.44$		$h = 2.85$		$h = 5.70$	
β_{10}	0.421	(0.275, 0.584)	0.388	(0.258, 0.515)	0.353	(0.255, 0.452)
β_{11}	0.157	(0.106, 0.212)	0.167	(0.120, 0.222)	0.177	(0.127, 0.236)
β_{20}	-0.035	(-0.063, -0.011)	-0.033	(-0.063, -0.009)	-0.032	(-0.062, -0.005)
β_{21}	1.021	(0.986, 1.050)	1.022	(1.001, 1.053)	1.024	(1.004, 1.041)

mixing proportions. An optimal bandwidth is selected at 2.85 by CV procedure, and the under-smoothing bandwidth and over-smoothing bandwidth are selected at 1.44 and 5.70. For the optimal bandwidth, we first test whether the mixing proportions vary by using the proposed conditional bootstrap method. Based on 500 conditional bootstrap simulations, the resulting test statistics T is 26.10, and the approximate p-value of the test is less than 0.001. In fact, the testing procedure rejects the constant proportion hypothesis under a wide range of bandwidths, including both the under-smoothing and over-smoothing bandwidths. This suggests that it is appropriate to use a mixture of regression models with varying proportions.

The resulting estimate of $\boldsymbol{\beta}$ along with its 95% confidence interval (CI) are shown in

Table 4. Take the results of bandwidth 1.44 for illustration. The lower component has an estimated slope $\hat{\beta}_{11} = 0.157$. We may conclude that for countries within this component, increasing in GDP per capita for a thousand dollar may be on average associated with increment of 0.157 ton CO2-emission per capita, and a 95% CI of such CO2-emission increment per capita is from 0.106 to 0.212 ton. Most developed countries are of this component, and the representatives include US, UK, Canada, Australia, etc. The upper component has an estimated slope $\hat{\beta}_{21} = 1.021$. For countries within this component, increasing in GDP per capita for a thousand dollar may be on average associated with increment of 1.021 metric ton CO2-emission per capita, and a 95% CI is from 0.986 to 1.050 ton. Representatives countries of this component include Kuwait, Saudi Arabia, Qatar, etc. The functional estimate of the mixing proportion function of the lower component together with its 95% bootstrap pointwise confidence interval are depicted in Figure 3(b). The result shows that as GDP per capita increases, the proportion of low CO2 emission counties increases, which indicates that high GDP-per-capita countries tend to develop in a relative low-CO2-emission path.

4. DISCUSSION

In this article, we assume that the number of components C is known. However, in many cases, C might be unknown and we need to estimate both C and bandwidth h. One might first select C and then select the bandwidth h after C is given. Choosing the number of components in mixture model is an important problem, which attracts many attentions in statistical research. For parametric mixture models, many methods have been proposed to deal with this selection issue. One popular and simple approach is the information criteria, such as AIC and BIC. Leroux (1992) proved the weak consistency of the maximum penalized likelihood estimators for the mixing distribution. For other references, see McLachlan and Peel (2000), Chen et al. (2004), and Chen and Li (2009).

The choice of the number of components is related to degrees of freedom. However, the degrees of freedom of the proposed model is not clear. In practice, we may use the results of traditional parametric mixture models. Note that locally in covariate z, the mixing

proportions of model (2.1) can be considered as constant. Therefore, one might apply the information criteria to the partial data in a local area. We may take several typical local areas, and determine C by comparing several selection results. Since the variance of Y tends to increase when the separation of mixture components increases, the local areas can be those with relatively large variation of Y. More research are needed on how to choose the number of components for model (2.1).

5. PROOFS

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

$$\sum_{c=1}^{C} \pi_c N(\mu_c, \sigma_c^2) = \sum_{d=1}^{D} \lambda_d N(\nu_d, \tau_d^2),$$

where the parameters satisfy $\pi_c > 0$, $c = 1, \ldots, C$, $\sigma_1^2 \leq \cdots \leq \sigma_C^2$, and if $\sigma_i^2 = \sigma_j^2$ and $i < j$, then $\mu_i < \mu_j$; similarly, $\lambda_d > 0$, $d = 1, \ldots, D$, $\tau_1^2 \leq \cdots \leq \tau_D^2$, and if $\tau_i^2 = \tau_j^2$ and $i < j$, then $\nu_i < \nu_j$. Then $C = D$ and $(\pi_c, \mu_c, \sigma_c^2) = (\lambda_c, \nu_c, \tau_c^2)$, $c = 1, \ldots, C$. (See Titterington et al. (1985), p. 38, Example 3.1.4)

Proof of Theorem 1. Suppose that model (2.1) admits another representation

$$Y|_{X=x,Z=z} \sim \sum_{d=1}^{D} \lambda_d(z) N(\mathbf{x}^T \boldsymbol{\gamma}_d, \delta_d^2),$$

where $\lambda_d(z) > 0$, $d = 1, \ldots, D$, and $(\boldsymbol{\gamma}_d, \delta_d^2)$, $d = 1, \ldots, D$, are distinct.

For any two distinct pairs of parameters $(\boldsymbol{\beta}_a, \sigma_a^2)$ and $(\boldsymbol{\beta}_b, \sigma_b^2)$, if $\sigma_a^2 = \sigma_b^2$, then $\boldsymbol{\beta}_a \neq \boldsymbol{\beta}_b$, therefore, the set $\{x \in \mathbb{R}^p : \mathbf{x}^T \boldsymbol{\beta}_a = \mathbf{x}^T \boldsymbol{\beta}_b\}$ is either an empty set or a $(p-1)$-dimensional hyperplane in $\mathbb{R}^p$, and thus has zero Lebesgue measure in $\mathbb{R}^p$. This implies that there are at most a finite number of $(p-1)$-dimensional hyperplanes on which $(\mathbf{x}^T \boldsymbol{\beta}_a, \sigma_a^2) = (\mathbf{x}^T \boldsymbol{\beta}_b, \sigma_b^2)$ for some a, b. Hence the union of these finite number of hyperplanes has zero Lebesgue measure in $\mathbb{R}^p$. The same thing is true for the set of parameters $(\boldsymbol{\gamma}_d, \delta_d^2)$, $d = 1, \ldots, D$.

From Lemma 1, for any given (x, z) such that both sets of parameters $(\mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2)$, $c = 1, \ldots, C$, and $(\mathbf{x}^T \boldsymbol{\gamma}_d, \delta_d^2)$, $d = 1, \ldots, D$, are distinct pairs, respectively, model (2.1)

conditioning on $t = (x, z)$ is identifiable. Therefore, $C = D$ and there exists a permutation $\omega_t = \{\omega_t(1), \ldots, \omega_t(C)\}$ of set $\{1, \ldots, C\}$ depending on t, such that $\lambda_{\omega_t(c)}(z) = \pi_c(z)$, $\mathbf{x}^T\boldsymbol{\gamma}_{\omega_t(c)} = \mathbf{x}^T\boldsymbol{\beta}_c$, $\delta^2_{\omega_t(c)} = \sigma^2_c$, $c = 1, \ldots, C$. Consider any permutation $\omega = \{\omega(1), \ldots, \omega(C)\}$ such that

$$\mathbf{x}^T\boldsymbol{\gamma}_{\omega(c)} = \mathbf{x}^T\boldsymbol{\beta}_c,\ \delta^2_{\omega(c)} = \sigma^2_c,\ \ c = 1, \ldots, C. \tag{5.1}$$

for some $\mathbf{x}$ values. If $\boldsymbol{\gamma}_{\omega(c)} \neq \boldsymbol{\beta}_c$ for some c, then the set $\{x \in \mathbb{R}^p : \mathbf{x}^T\boldsymbol{\gamma}_{\omega(c)} = \mathbf{x}^T\boldsymbol{\beta}_c\}$ is contained in a $(p-1)$-dimensional hyperplane in $\mathbb{R}^p$ and has a zero Lebesgue measure. Since there are only a finite number $(C!)$ of possible permutations of $\{1, 2, \ldots, C\}$ and the domain $\mathcal{X}$ of x contains an open set in $\mathbb{R}^p$, there must exist a permutation $\omega^* = \{\omega^*(1), \ldots, \omega^*(C)\}$, such that (5.1) holds on a subset of $\mathcal{X}$ with nonzero Lebesgue measure. Hence, $\boldsymbol{\gamma}_{\omega^*(c)} = \boldsymbol{\beta}_c, \delta^2_{\omega^*(c)} = \sigma^2_c, c = 1, \ldots, C$. Because that $(\boldsymbol{\beta}_c, \sigma^2_c)$, $c = 1, \ldots, C$ are distinct and $(\boldsymbol{\gamma}_c, \delta^2_c)$, $c = 1, \ldots, C$ are distinct, it follows that ω^* is the unique permutation such that (5.1) holds on a subset of $\mathcal{X}$ with nonzero Lebesgue measure. If z is not from x, then $\lambda_{\omega^*(c)}(z) = \pi_c(z)$, $c = 1, \ldots, C$ for any $z \in \mathcal{Z}$. If z is from x, $\lambda_{\omega^*(c)}(z) = \pi_c(z)$, $c = 1, \ldots, C$, for all $z \in \mathcal{Z}$ but points where some hyperplanes intersect. Because $\pi_c(z)$ are continuous and the domain of z has no isolated points, the values of $\pi_c(z)$ at those points where some hyperplanes intersect are also uniquely determined. This completes the proof.

We next outline the key steps of proofs for Theorems 2 to 4. Note that $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T, \boldsymbol{\beta}^T)^T$ is a $((p+3)C-1)\times 1$ vector. Whenever necessary, we rewrite $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{(p+3)C-1})^T$ without changing the order of $\boldsymbol{\pi}, \boldsymbol{\sigma}^2$, and $\boldsymbol{\beta}$.

Regularity Conditions

A. The sample $\{(X_i, Y_i, Z_i),\ i = 1, \cdots, n\}$ is independent and identically distributed from the joint density $f(x, y, z)$ with finite sixth moments. The support for z, denoted by $\mathcal{Z}$, is closed and bounded of $\mathbb{R}^1$.

B. The joint density $f(x, y, z)$ has continuous first derivative and is positive in its support.

C. The third derivative $|\partial^3\ell(\boldsymbol{\theta}, x, y, z)/\partial\theta_j\partial\theta_k\partial\theta_l| \leq M_{jkl}(x, y, z)$, where $\mathrm{E}\{M_{jkl}(X, Y, Z)\}$ is bounded for all j, k, l, and all X and Y.

D. The unknown functions $\pi_c(z), c = 1, \cdots, C-1$, have continuous second derivative.

E. The kernel density function $K(\cdot)$ is symmetric, continuous, and has a closed and bounded support.

F. For $c = 1, \cdots, C$, $\sigma_c^2 > 0$, and $\pi_c(z) > 0$ hold for all $z \in \mathcal{Z}$.

G. The second derivative matrix $-\mathrm{E}\{\partial^2 \ell(\boldsymbol{\theta}(z), x, y)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T \mid Z = z\}$ is positive definite, where $\boldsymbol{\theta}(z) = (\boldsymbol{\pi}^T(z), (\boldsymbol{\sigma}^2)^T, \boldsymbol{\beta}^T)^T$.

H. $\mathrm{E}(Z^{2r}) < \infty$ for some $\varepsilon < 1 - r^{-1}$, $n^{2\varepsilon-1}h \to \infty$.

All the above conditions are mild conditions and have been used in the literature of local likelihood estimation and mixture models. Let

$$\ell(\boldsymbol{\theta}) = \log\left\{\sum_{c=1}^{C} \pi_c \phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)\right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T, \boldsymbol{\beta}^T)^T$ and $\phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)$ is the normal density of y with mean $\mathbf{x}^T\boldsymbol{\beta}_c$ and variance σ_c^2. Then

$$\begin{aligned}
\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\beta}_c &= \frac{\pi_c\phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)(y - \mathbf{x}^T\boldsymbol{\beta}_c)\mathbf{x}/\sigma^2}{\sum_{c=1}^{C}\pi_c\phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)} \\
\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_c\partial\boldsymbol{\beta}_c^T} &= \left[\left\{\sum_{c=1}^{C}\pi_c\phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)\right\}\left\{\pi_c\phi^2\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)(y - \mathbf{x}^T\boldsymbol{\beta}_c)^2\mathbf{x}\mathbf{x}^T/\sigma^4\right.\right. \\
&\quad \left.\left. -\pi_c\phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)\mathbf{x}\mathbf{x}^T/\sigma^2\right\} - \pi_c^2\phi^2\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)(y - \mathbf{x}^T\boldsymbol{\beta}_c)^2\mathbf{x}\mathbf{x}^T/\sigma^4\right] \\
&\quad \times\left\{\sum_{c=1}^{C}\pi_c\phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)\right\}^{-2}
\end{aligned}$$

Note that $\phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)$ and $\phi\left(y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right)(y - \mathbf{x}^T\boldsymbol{\beta}_c)^k$ is bounded for any c and $k > 0$. Then we have

$$\sup_z \mathrm{E}\left[\left|\frac{\partial^2\ell(\boldsymbol{\theta}(z), x, y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|^3 \mid Z = z\right] < \infty,$$

and

$$E\left(|\partial\ell(\boldsymbol{\theta}, X, Y, Z)/\partial\theta_j|^3\right) < \infty$$

if X have sixth finite moments.

The following lemma is taken from Lemma A.1 of Fan and Huang (2005) and will be used throughout the proofs of this section.

Lemma 2. Let $\{(X_i, Y_i), i = 1 \cdots, n\}$ be i.i.d random vectors from (X, Y), where X is a random vector and Y is a scalar random variable. Denote f^* to be the joint density of (X, Y), and further assume that $\mathrm{E}|Y|^r < \infty$ 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

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

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

To establish asymptotic properties of $\hat{\boldsymbol{\eta}}$, we first study the asymptotic behaviors of $\{\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\sigma}}^2, \tilde{\boldsymbol{\beta}}\}$, the maximum local likelihood estimator of (2.3). Denote

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_c^* &= \sqrt{nh}\{\tilde{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_c\}, \\
\tilde{\sigma}_c^{2*} &= \sqrt{nh}\{\tilde{\sigma}_c^2 - \sigma_c^2\}, \\
\tilde{\pi}_c^* &= \sqrt{nh}\{\tilde{\pi}_c - \pi_c(z)\},\ c = 1, \ldots, C-1 \\
\tilde{\pi}_C^* &= \sqrt{nh}\{\tilde{\pi}_C - \pi_C(z)\} = \sqrt{nh}[1 - \sum_{c=1}^{C-1}\{\tilde{\pi}_c - \pi_c(z)\}],
\end{aligned}$$

Let $\tilde{\boldsymbol{\beta}}^* = \{(\tilde{\boldsymbol{\beta}}_1^*)^T, \cdots, (\tilde{\boldsymbol{\beta}}_C^*)^T\}^T$, $\tilde{\boldsymbol{\sigma}}^{2*} = (\tilde{\sigma}_1^{2*} \cdots, \tilde{\sigma}_C^{2*})^T$, and $\tilde{\boldsymbol{\pi}}^* = (\tilde{\pi}_1^*, \cdots, \tilde{\pi}_{C-1}^*)^T$. Define $\tilde{\boldsymbol{\theta}}^* = \{(\tilde{\boldsymbol{\pi}}^*)^T, (\tilde{\boldsymbol{\sigma}}^{2*})^T, (\tilde{\boldsymbol{\beta}}^*)^T\}^T$.

Lemma 3. Assume that Conditions (A)—(H) hold, in addition with $nh \to \infty$ as $n \to \infty$, $h \to 0$, then for all z in the support $\mathcal{Z}$, we have

$$\sup_{z \in \mathcal{Z}} |\tilde{\boldsymbol{\theta}}^* - f^{-1}(z)\mathcal{I}_\theta^{-1}(z)\Delta_n| = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\},$$

where Δ_n is defined in (5.4), and

$$\mathcal{I}_\theta(z) = -\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}, x, y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mid Z = z\right].$$

Proof.

If $\{\tilde{\boldsymbol{\pi}}_0, \tilde{\boldsymbol{\sigma}}_0^2, \tilde{\boldsymbol{\beta}}_0\}$ maximizes (2.3), then $\tilde{\boldsymbol{\theta}}^*$ maximizes

$$\ell_n^*(\boldsymbol{\theta}^*) = h\sum_{i=1}^{n}\{\ell(\boldsymbol{\theta}(z) + \gamma_n\boldsymbol{\theta}^*, X_i, Y_i) - \ell(\boldsymbol{\theta}(z), X_i, Y_i)\}K_h(Z_i - z), \tag{5.2}$$

where $\boldsymbol{\theta}(z) = \{(\boldsymbol{\pi}(z))^T, (\boldsymbol{\sigma}^2)^T, (\boldsymbol{\beta})^T\}^T$. By the Taylor expansion and some calculation,

$$\ell_n^*(\boldsymbol{\theta}^*) = \Delta_n\boldsymbol{\theta}^* + \frac{1}{2}\boldsymbol{\theta}^{*T}\Gamma_n\boldsymbol{\theta}^* + o_p(1), \tag{5.3}$$

where

$$\Delta_n = \sqrt{\frac{h}{n}}\sum_{i=1}^{n} q_\theta(\boldsymbol{\theta}(z), X_i, Y_i)K_h(Z_i - z), \tag{5.4}$$

$$\Gamma_n = \frac{1}{n}\sum_{i=1}^{n} q_{\theta\theta}(\boldsymbol{\theta}(z), X_i, Y_i)K_h(Z_i - z). \tag{5.5}$$

By the SLLN and some calculations, it follows that $\Gamma_n = -f(z)\mathcal{I}_\theta(z) + o_p(1)$. Therefore,

$$\ell_n^*(\boldsymbol{\theta}^*) = \Delta_n\boldsymbol{\theta}^* - \frac{1}{2}f(z)\boldsymbol{\theta}^{*T}\mathcal{I}_\theta(z)\boldsymbol{\theta}^* + o_p(||\theta^*||^2). \tag{5.6}$$

Since each element in Γ_n is sum of i.i.d. random variables, by Lemma 2 and condition (G), we can show that Γ_n converge to $-f(z)\mathcal{I}_\theta(z)$ uniformly for all $z \in \mathcal{Z}$. By (5.3) and condition (G), we know $\ell_n^*(\boldsymbol{\theta}^*)$ is a concave function of $\boldsymbol{\theta}^*$ for large n. Then by condition (F), when n is large enough, $-\ell_n^*(\boldsymbol{\theta}^*)$ is a convex function defined on a convex open set. Thus, by the convexity lemma (Pollard, 1991),

$$\sup_{z\in\mathcal{Z}}\left|(\Delta_n\boldsymbol{\theta}^* + \frac{1}{2}\boldsymbol{\theta}^{*T}\Gamma_n\boldsymbol{\theta}^*) - (\Delta_n\boldsymbol{\theta}^* - \frac{1}{2}f(z)\boldsymbol{\theta}^{*T}\mathcal{I}_\theta(z)\boldsymbol{\theta}^*)\right| \xrightarrow{P} 0 \tag{5.7}$$

holds uniformly for all $z \in \mathcal{Z}$ and $\boldsymbol{\theta}^*$ in any compact set Ω. We know that $f^{-1}(z)\mathcal{I}_\theta^{-1}(z)\Delta_n$ is a unique maximizer of (5.6), and is continuous in z; $\tilde{\boldsymbol{\theta}}^*$ is a maximizer of (5.3). Then by Lemma A.1 of Carroll et al. (1997), we have

$$\sup_{z\in\mathcal{Z}}|\tilde{\boldsymbol{\theta}}^* - f^{-1}(z)\mathcal{I}_\theta^{-1}(z)\Delta_n| \xrightarrow{P} 0. \tag{5.8}$$

Then by the definition of $\tilde{\boldsymbol{\theta}}^*$,

$$\left.\frac{\partial\ell_n^*(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^*}\right|_{\boldsymbol{\theta}^*=\tilde{\boldsymbol{\theta}}^*} = h\gamma_n\sum_{i=1}^{n} q_\theta\{\tilde{\boldsymbol{\theta}}^*(z), X_i, Y_i\}K_h(Z_i - z) = 0. \tag{5.9}$$

By a Taylor expansion, we have

$$\Delta_n + \Gamma_n \tilde{\boldsymbol{\theta}}^* + \frac{h\gamma_n^3}{2} \sum_{i=1}^{n} \sum_{j,l} \frac{\partial^2 q_\theta(\boldsymbol{\theta}(z) + \tilde{\xi}_i)}{\partial \theta_j^* \partial \theta_l^*} \tilde{\theta}_j^* \tilde{\theta}_l^{*T} K_h(Z_i - z) = 0, \tag{5.10}$$

where $\boldsymbol{\theta}^*$ is rewritten as $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_{(p+3)C-1}^*)^T$. $\tilde{\xi}_i$ is a vector between 0 and $\gamma_n \boldsymbol{\theta}^*$. The last term of (5.10) is of order $O_p(\gamma_n ||\tilde{\boldsymbol{\theta}}^*||^2)$. Again it can be deduced from Lemma 2, for each element of Γ_n,

$$\sup_{z \in \mathcal{Z}} |\Gamma_n(i,j) - \mathrm{E}\{\Gamma_n(i,j)\}| = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}. \tag{5.11}$$

By (5.10), $\Gamma_n \tilde{\boldsymbol{\theta}}^* + O_p(\gamma_n ||\tilde{\boldsymbol{\theta}}^*||^2) = -\Delta_n$, then

$$\{\Gamma_n - \mathrm{E}(\Gamma_n)\}\tilde{\boldsymbol{\theta}}^* + O_p(\gamma_n ||\tilde{\boldsymbol{\theta}}^*||^2) = -\Delta_n + f(z)\mathcal{I}_\theta(z)\tilde{\boldsymbol{\theta}}^*. \tag{5.12}$$

By (5.8), it is obvious that $\sup_{z \in \mathcal{Z}} |\tilde{\boldsymbol{\theta}}^*| = O_p(1)$. Thus for the left side of (5.12), we have

$$\sup_{z \in \mathcal{Z}} |\{\Gamma_n - \mathrm{E}(\Gamma_n)\}\tilde{\boldsymbol{\theta}}^*| + O_p(\gamma_n) = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}.$$

It follows that the order also holds for the right side of (5.12), i.e.,

$$\sup_{z \in \mathcal{Z}} |f(z)\mathcal{I}_\theta(z)\tilde{\boldsymbol{\theta}}^* - \Delta_n| = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}.$$

The proof is completed by the conditions that $f(z)$ and $\mathcal{I}_\theta(z)$ are bounded and continuous functions in a closed set of $\mathcal{Z}$.

Proof of Theorem 2. Denote $\hat{\boldsymbol{\eta}}^* = \sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is the true value. Further, define

$$\begin{aligned}
\ell(\tilde{\boldsymbol{\pi}}(Z_i), \boldsymbol{\eta}, X_i, Y_i) &= \log\left\{\sum_{c=1}^{C} \tilde{\pi}_c(Z_i)\phi(Y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c, \sigma_c^2)\right\}, \\
\ell(\tilde{\boldsymbol{\pi}}(Z_i), \hat{\boldsymbol{\eta}} + \boldsymbol{\eta}^*/\sqrt{n}, X_i, Y_i) &= \log\left\{\sum_{c=1}^{C} \tilde{\pi}_c(Z_i)\phi\{Y_i|\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_c + \boldsymbol{\beta}_c^*/\sqrt{n}), \hat{\sigma}_c^2 + \sigma_c^{*2}/\sqrt{n}\right\}.
\end{aligned}$$

Then $\hat{\boldsymbol{\eta}}^*$ maximizes

$$\ell_n(\boldsymbol{\eta}^*) = \sum_{i=1}^{n} \{\ell(\tilde{\boldsymbol{\pi}}(Z_i), \boldsymbol{\eta} + \boldsymbol{\eta}^*/\sqrt{n}, X_i, Y_i) - \ell(\tilde{\boldsymbol{\pi}}(Z_i), \boldsymbol{\eta}, X_i, Y_i)\}. \tag{5.13}$$

By a Taylor expansion and some calculation,

$$\ell_n(\boldsymbol{\eta}^*) = A_n\boldsymbol{\eta}^* + \frac{1}{2}\boldsymbol{\eta}^{*T}B_n\boldsymbol{\eta}^* + o_p(1), \tag{5.14}$$

where

$$\begin{aligned} A_n &= n^{-1/2}\sum_{i=1}^{n}\frac{\partial \ell(\tilde{\boldsymbol{\pi}}(Z_i),\boldsymbol{\eta},X_i,Y_i)}{\partial \boldsymbol{\eta}}, \\ B_n &= n^{-1}\sum_{i=1}^{n}\frac{\partial^2 \ell(\tilde{\boldsymbol{\pi}}(Z_i),\boldsymbol{\eta},X_i,Y_i)}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}. \end{aligned}$$

For B_n, it can be shown that

$$B_n = -\mathrm{E}\{\mathcal{I}_\eta(X)\} + o_p(1).$$

Then by (5.14), we have

$$\ell_n(\boldsymbol{\eta}^*) = A_n\boldsymbol{\eta}^* - \frac{1}{2}\boldsymbol{\eta}^{*T}B\,\boldsymbol{\eta}^* + o_p(1). \tag{5.15}$$

Next, we expand A_n as

$$\begin{aligned} A_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell(\boldsymbol{\pi}(Z_i),\boldsymbol{\eta},X_i,Y_i)}{\partial \boldsymbol{\eta}} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^2 \ell(\boldsymbol{\pi}(Z_i),\boldsymbol{\eta},X_i,Y_i)}{\partial \boldsymbol{\eta}\partial \boldsymbol{\pi}^T}\{\tilde{\boldsymbol{\pi}}(Z_i)-\boldsymbol{\pi}(Z_i)\} + O_p(d_{1n}) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell(\boldsymbol{\pi}(Z_i),\boldsymbol{\eta},X_i,Y_i)}{\partial \boldsymbol{\eta}} + T_{n1} + O_p(d_{1n}). \end{aligned}$$

where $d_{1n} = n^{-1/2}||\tilde{\boldsymbol{\pi}}-\boldsymbol{\pi}||_\infty^2$. By Lemma 2, we have

$$\tilde{\boldsymbol{\theta}}(Z_i) - \boldsymbol{\theta}(Z_i) = \frac{1}{n}f^{-1}(Z_i)\mathcal{I}_\theta^{-1}(Z_i)\sum_{j=1}^{n}\frac{\partial \ell(\boldsymbol{\theta}(Z_i),X_j,Y_j)}{\partial \boldsymbol{\theta}}K_h(Z_j-Z_i) + O_p(d_{n2}),$$

where $d_{n2} = \gamma_n h^2 + \gamma_n^2\sqrt{\log(1/h)}$. Let $\boldsymbol{\psi}(X_j,Y_j,Z_j)$ be a $(C-1)\times 1$ vector, in which the elements are taken from the first $C-1$ entries of $\mathcal{I}_\theta^{-1}(z_j)\times\{\partial\ell(\boldsymbol{\theta}(Z_j),X_j,Y_j)/\partial\boldsymbol{\theta}\}$.

By condition $nh^2/\log(1/h)\to\infty$, we have $O_p(n^{1/2}d_{n2}) = o_p(1)$. Since $\boldsymbol{\pi}(Z_i)-\boldsymbol{\pi}(Z_j) = O(Z_i-Z_j)$ and $K(\cdot)$ is symmetric about 0, we have

$$\begin{aligned} T_{n1} &= n^{-3/2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial^2 \ell(\boldsymbol{\pi}(Z_i),\boldsymbol{\eta},X_i,Y_i)}{\partial \boldsymbol{\eta}\partial \boldsymbol{\pi}^T}f^{-1}(Z_i)\boldsymbol{\psi}(X_j,Y_j,Z_j)K_h(Z_i-Z_j) + O_p(n^{1/2}h^2) \\ &= T_{n2} + O_p(n^{1/2}h^2). \end{aligned}$$

It can be shown, by calculating the second moment, that

$$T_{n2} - T_{n3} \xrightarrow{P} 0, \tag{5.16}$$

where $T_{n3} = -n^{-1/2}\sum_{j=1}^{n}\boldsymbol{\omega}(X_j, Y_j, Z_j)$, with

$$\begin{aligned}\boldsymbol{\omega}(X_j, Y_j, Z_j) &= -\mathrm{E}\left\{\frac{\partial^2\ell(\boldsymbol{\pi}(Z), \boldsymbol{\eta}, X, Y)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\pi}^T} \mid Z = Z_j\right\}\boldsymbol{\psi}(X_j, Y_j, Z_j)\\ &= \mathcal{I}_{\eta\pi}(Z_j)\boldsymbol{\psi}(X_j, Y_j, Z_j).\end{aligned}$$

By condition $nh^4 \to 0$, we know

$$A_n = n^{-1/2}\sum_{i=1}^{n}\left\{\frac{\partial\ell(\boldsymbol{\pi}(Z_i), \boldsymbol{\eta}, X_i, Y_i)}{\partial\boldsymbol{\eta}} - \boldsymbol{\omega}(X_i, Y_i, Z_i)\right\} + o_p(1).$$

By (5.15) and quadratic approximation lemma,

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

Then we calculate the mean and variance of A_n. It is obvious that $\mathrm{Var}(A_n) = \Sigma$, and

$$\mathrm{E}(A_n) = \sqrt{n}\,\mathrm{E}\left\{\frac{\partial\ell(\pi(Z), \boldsymbol{\eta}, X, Y)}{\partial\boldsymbol{\eta}} - \boldsymbol{\omega}(X, Y, Z)\right\}.$$

We can show that the elements of $\mathrm{E}(\partial\ell(\pi(Z), \boldsymbol{\eta}, X, Y)/\partial\boldsymbol{\eta})$ are equal to 0, and

$$\mathrm{E}\left\{\boldsymbol{\omega}(X, Y, Z)\right\} = \mathrm{E}\left\{\mathcal{I}_{\eta\pi}(Z)\boldsymbol{\psi}(X, Y, Z)\right\},$$

where $\boldsymbol{\psi}(X, Y, Z)$ are the $[1^{th}, \cdots, (C-1)^{th}]$ elements of $\mathcal{I}_\theta^{-1}(Z) \times \{\partial\ell(\boldsymbol{\theta}(Z), X, Y)/\partial\boldsymbol{\theta}\}$. Further calculation shows that $\mathrm{E}\left\{\boldsymbol{\omega}(X, Y, Z)\right\} = 0$. So we have $\mathrm{E}(A_n) = 0$. By the Central Limit Theorem we complete the proof of Theorem 2.

Proof of Theorem 3. Using similar arguments in the proof of Lemma 3, we have

$$\sqrt{nh}\{\hat{\boldsymbol{\pi}}(z) - \boldsymbol{\pi}(z)\} = f(z)^{-1}\mathcal{I}_\pi(z)^{-1}\hat{\Delta}_n + o_p(1), \tag{5.17}$$

where

$$\hat{\Delta}_n = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\boldsymbol{\pi}(z), \hat{\boldsymbol{\eta}}, X_i, Y_i)}{\partial\boldsymbol{\pi}}K_h(Z_i - z).$$

It can be calculated that

$$\hat{\Delta}_n = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\frac{\partial\ell(\boldsymbol{\pi}(z), \boldsymbol{\eta}, X_i, Y_i)}{\partial\boldsymbol{\pi}}K_h(Z_i - z) + D_n + o_p(1),$$

where

$$D_n = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\pi}(z), \boldsymbol{\eta}, X_i, Y_i)}{\partial \boldsymbol{\pi} \partial \boldsymbol{\eta}^T} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) K_h(Z_i - z)$$

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

$$D_n = -\sqrt{h} \mathcal{I}_{\eta\pi}^T(z) f(z) = o_p(1).$$

Hence

$$\sqrt{nh}\{\hat{\boldsymbol{\pi}}(z) - \boldsymbol{\pi}(z)\} = f(z)^{-1} \mathcal{I}_{\pi}(z)^{-1} \Delta_n + o_p(1),$$

where

$$\Delta_n = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\pi}(z), \boldsymbol{\eta}, X_i, Y_i)}{\partial \boldsymbol{\pi}} K_h(Z_i - z).$$

We can show that

$$\mathrm{Var}(\Delta_n) = \mathcal{I}_{\pi}(z) f(z) \nu_0$$

and

$$\mathrm{E}(\Delta_n) = \frac{\sqrt{nh}}{2} \left\{ \Lambda''(z|z) f(z) + 2\Lambda'(z|z) f'(z) \right\} \kappa_2 h^2,$$

where $\kappa_l = \int u^l K(u)\, du$, and $\nu_l = \int u^l K^2(u)\, du$. Then the result of Theorem 3 follows a standard argument.

Proof of Theorem 4.

(a) We assume the unobserved data $(\mathcal{C}_i, i = 1, \cdots, n)$ are random samples from population $\mathcal{C}$, and the complete data $\{(X_i, Y_i, Z_i, \mathcal{C}_i), i = 1, 2, \cdots, n\}$ are random samples from $(X, Y, Z, \mathcal{C})$. The conditional distribution of $\mathcal{C}$ given X, Y, and $\boldsymbol{\theta}$ is

$$g\{c|X, Y, \boldsymbol{\theta}\} = \frac{\pi_c \phi(Y|\mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2)}{\sum_{c=1}^{C} \pi_c \phi(Y|\mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2)}. \tag{5.18}$$

For given $\boldsymbol{\theta}^{(l)}(Z_i) = \{\boldsymbol{\pi}^{(l)}(Z_i), \boldsymbol{\beta}^{(l)}(Z_i), \boldsymbol{\sigma}^{2(l)}(Z_i)\}$, we have $g\{c|X_i, Y_i, \boldsymbol{\theta}^{(l)}(Z_i)\} = r_{ic}^{(l+1)}$,

and $\sum_{c=1}^{C} r_{ic}^{(l+1)} = 1$, $i = 1, \cdots, n$. Then

$$\begin{aligned}\ell_1(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \phi(Y_i | \mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2) \right\} \left(\sum_{c=1}^{C} r_{ic}^{(l+1)} \right) K_h(Z_i - z) \\ &= \sum_{i=1}^{n} \left\{ \sum_{c=1}^{C} \log \left\{ \sum_{c=1}^{C} \pi_c \phi(Y_i | \mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2) \right\} r_{ic}^{(l+1)} \right\} K_h(Z_i - z). \end{aligned} \tag{5.19}$$

By (5.18), we also have

$$\log \left\{ \sum_{c=1}^{C} \pi_c \phi(Y_i | \mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2) \right\} = \log\{\pi_c \phi(Y_i | \mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2)\} - \log[g\{c | X_i, Y_i, \boldsymbol{\theta}\}]. \tag{5.20}$$

Thus, we have

$$\begin{aligned}\ell_1(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left\{ \sum_{c=1}^{C} \log\{\pi_c \phi(Y_i | \mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2)\} r_{ic}^{(l+1)} \right\} K_h(Z_i - z) \\ &- \sum_{i=1}^{n} \left\{ \sum_{c=1}^{C} \log[g\{c | X_i, Y_i, \boldsymbol{\theta}\}] r_{ic}^{(l+1)} \right\} K_h(Z_i - z), \end{aligned} \tag{5.21}$$

Based on the M-step of (2.7) — (2.9) we have

$$\begin{aligned}&\frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{c=1}^{C} \log\{\pi_c^{(l+1)}(z) \phi(Y_i | \mathbf{x}^T \boldsymbol{\beta}_c^{(l+1)}(z), \sigma_c^{2(l+1)}(z)\} r_{ic}^{(l+1)} \right\} K_h(Z_i - z) \\ \geq\ &\frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{c=1}^{C} \log\{\pi_c^{(l)}(z) \phi(Y_i | \mathbf{x}^T \boldsymbol{\beta}_c^{(l)}(z), \sigma_c^{2(l)}(z))\} r_{ic}^{(l+1)} \right\} K_h(Z_i - z). \end{aligned}$$

It suffices to show that

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{c=1}^{C} \log \left\{ \frac{g\{c | X_i, Y_i, \boldsymbol{\theta}^{(l+1)}(z)\}}{g\{c | X_i, Y_i, \boldsymbol{\theta}^{(l)}(z)\}} \right\} r_{ic}^{(l+1)} \right] K_h(Z_i - z) \leq 0 \tag{5.22}$$

in probability. Define

$$L_g = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{c=1}^{C} \log \left\{ \frac{g\{c | X_i, Y_i, \boldsymbol{\theta}^{(l+1)}(z)\}}{g\{c | X_i, Y_i, \boldsymbol{\theta}^{(l)}(z)\}} \right\} r_{ic}^{(l+1)} \right] K_h(Z_i - z),$$

and

$$L_J = \frac{1}{n} \sum_{i=1}^{n} \log \left[\sum_{c=1}^{C} \left\{ \frac{g\{c | X_i, Y_i, \boldsymbol{\theta}^{(l+1)}(z)\}}{g\{c | X_i, Y_i, \boldsymbol{\theta}^{(l)}(z)\}} \right\} r_{ic}^{(l+1)} \right] K_h(Z_i - z).$$

By Jensen's inequality, $L_g \leq L_J$. Next we show that $L_J \to 0$ in probability. For the simplicity of proof, we assume $g\{c|X, Y, \boldsymbol{\theta}^{(l)}(Z)\} \geq a > 0$ for some small value a, which can always be done in practice . To this end, we first calculate the expectation of L_J.

$$E(L_J) = E\left(\log\left[\sum_{c=1}^{C} \frac{g\{c|X, Y, \boldsymbol{\theta}^{(l+1)}(z)\}}{g\{c|X, Y, \boldsymbol{\theta}^{(l)}(z)\}} g\{c|X, Y, \boldsymbol{\theta}^{(l)}(Z)\}\right] K_h(Z - z)\right).$$

By a standard argument, we know

$$\Delta_n(X, Y) \triangleq E\left(\log\left[\sum_{c=1}^{C} \frac{g\{c|X, Y, \boldsymbol{\theta}^{(l+1)}(z)\}}{g\{c|X, Y, \boldsymbol{\theta}^{(l)}(z)\}} g\{c|X, Y, \boldsymbol{\theta}^{(l)}(Z)\}\right] K_h(Z - z)\Big|X, Y\right) \to 0.$$

Noting that $\Delta_n(X, Y)$ is bounded, we have

$$E(L_J) = E(\Delta_n(X, Y)) \to 0.$$

We next calculate the variance of L_J. Note that the variance of L_J is dominated by the following term

$$\frac{1}{n} E\left(\log\left[\sum_{c=1}^{C} \frac{g\{c|X, Y, \boldsymbol{\theta}^{(l+1)}(z)\}}{g\{c|X, Y, \boldsymbol{\theta}^{(l)}(z)\}} g\{c|X, Y, \boldsymbol{\theta}^{(l)}(Z)\}\right] K_h(Z - z)\right)^2,$$

which can be shown to have the order $O_p\{(nh)^{-1}\}$. Then we have $L_J = o_p(1)$ by Chebyshev inequality. This completes the proof.

(b)

$$\begin{aligned}
\ell_3(\boldsymbol{\pi}^{(l+1)}) - \ell_3(\boldsymbol{\pi}^{(l)}) &= \sum_{i=1}^{n} \log\left\{\frac{\sum_{c=1}^{C} \pi_c^{(l+1)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}{\sum_{c=1}^{C} \pi_c^{(l)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}\right\} K_h(Z_i - z) \\
&= \sum_{i=1}^{n} \log \sum_{c=1}^{C} \left\{\frac{\pi_c^{(l)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}{\sum_{c=1}^{C} \pi_c^{(l)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)} \frac{\pi_c^{(l+1)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}{\pi_c^{(l)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}\right\} K_h(Z_i - z) \\
&= \sum_{i=1}^{n} \log \sum_{c=1}^{C} \left\{r_{ic}^{(l+1)} \frac{\pi_c^{(l+1)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}{\pi_c^{(l)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}\right\} K_h(Z_i - z)
\end{aligned}$$

Based on the Jensen's inequality, we have

$$\ell_3(\boldsymbol{\pi}^{(l+1)}) - \ell_3(\boldsymbol{\pi}^{(l)}) \geq \sum_{i=1}^{n} \sum_{c=1}^{C} r_{ic}^{(l+1)} \log\left\{\frac{\pi_c^{(l+1)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}{\pi_c^{(l)} \phi(Y_i|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)}\right\} K_h(Z_i - z).$$

Based on the M-step of (2.14), we have

$$\ell_3(\boldsymbol{\pi}^{(l+1)}) - \ell_3(\boldsymbol{\pi}^{(l)}) \geq 0.$$

(c) By fixing $\hat{\boldsymbol{\pi}}(\cdot) = \boldsymbol{\pi}^{(l)}(\cdot)$, $\ell^*(\boldsymbol{\pi}^{(l)}(\cdot), \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ is equal to $\ell_1(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$. Then by the ascent property of the ordinary EM algorithm, we have

$$\ell^*\{\boldsymbol{\pi}^{(l)}(\cdot), \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\sigma}^{2(l+1)}\} \;\; \geq \;\; \ell^*\{\boldsymbol{\pi}^{(l)}(\cdot), \boldsymbol{\beta}^{(l)}, \boldsymbol{\sigma}^{2(l)}\}.$$

Therefore, we only need to show

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

Fix $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(l+1)}$ and $\hat{\boldsymbol{\sigma}}^2 = \boldsymbol{\sigma}^{2(l+1)}$, and take $z \in \{Z_j, j = 1, \cdots, n\}$. By similar arguments of Theorem 4(a), we can show that for any given z,

$$\liminf_{n\to\infty} \; n^{-1}\left[\ell_3\{\boldsymbol{\pi}^{(l+1)}(z)\} - \ell_3\{\boldsymbol{\pi}^{(l)}(z)\}\right] \geq 0$$

in probability. Hence,

$$\begin{aligned}
& \liminf_{n\to\infty} \frac{1}{n^2}\sum_{j=1}^{n} f(Z_j)^{-1}\left[\ell_3\{\boldsymbol{\pi}^{(l+1)}(Z_j)\} - \ell_3\{\boldsymbol{\pi}^{(l)}(Z_j)\}\right] \\
\geq \;\; & \liminf_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \liminf_{n\to\infty} \frac{1}{n} f(Z_j)^{-1}\left[\ell_3\{\boldsymbol{\pi}^{(l+1)}(Z_j)\} - \ell_3\{\boldsymbol{\pi}^{(l)}(Z_j)\}\right] \\
\geq \;\; & 0.
\end{aligned}$$

Since $K_h(Z_i - Z_j) = K_h(Z_j - Z_i)$, it can be shown that

$$\begin{aligned}
& \frac{1}{n^2}\sum_{j=1}^{n} f(Z_j)^{-1}\ell_3\{\boldsymbol{\pi}^{(l)}(Z_j)\} \\
= \;\; & \frac{1}{n^2}\sum_{j=1}^{n} f(Z_j)^{-1}\sum_{i=1}^{n} \log\left\{\sum_{c=1}^{C} \pi_c^{(l)}(Z_j)\phi(Y_i|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2)\right\} K_h(Z_i - Z_j) \\
= \;\; & \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n} f(Z_j)^{-1}\log\left[\sum_{c=1}^{C} \pi_c^{(l)}(Z_j)\phi\{Y_i|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2\}\right] K_h(Z_j - Z_i)\right) \\
= \;\; & \frac{1}{n}\sum_{i=1}^{n} D_i^{(l)},
\end{aligned}$$

where

$$D_i^{(l)} = \frac{1}{n}\sum_{j=1}^{n} f(Z_j)^{-1}\log\left[\sum_{c=1}^{C} \pi_c^{(l)}(Z_j)\phi\{Y_i|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2\}\right] K_h(Z_j - Z_i).$$

By treating (X_i, Y_i, Z_i) as fixed in $D_i^{(l)}$, we can further show that

$$\mathrm{E}(D_i^{(l)}|X_i, Y_i, Z_i) = \log\left[\sum_{c=1}^{C} \pi_c^{(l)}(Z_i)\phi\{Y_i|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_c, \hat{\sigma}_c^2\}\right](1 + o_p(1)),$$

and $\mathrm{Var}\{\mathrm{E}(D_i^{(l)}|X_i, Y_i, Z_i)\}$ is of order $O_p\{(nh)^{-1}\}$. It is easy to see that

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{E}(D_i^{(l)}|X_i, Y_i, Z_i) &= \ell^*\{\boldsymbol{\pi}^{(l)}(\cdot), \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\sigma}^{2(l+1)}\}(1 + o_p(1)),\\ \sum_{i=1}^{n} \mathrm{E}(D_i^{(l+1)}|X_i, Y_i, Z_i) &= \ell^*\{\boldsymbol{\pi}^{(l+1)}(\cdot), \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\sigma}^{2(l+1)}\}(1 + o_p(1)).\end{aligned}$$

This completes the proof of Theorem 4(c).

6. ACKNOWLEDGEMENTS

The authors are grateful to the editor, the associate editor, and the referees for their insightful comments and suggestions, which greatly improved this article.

REFERENCES

Cai, Z, Fan, J., and Li, R. (2000). Efficient Estimation and Inferences for Varying-coefficient Models. *Journal of American and Statisical Association*. 95, 888-902.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized Partially Linear Single-Index Models. *Journal of American and Statisical Association*. 92, 477-489.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95, 957-970.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society*, Series B, 66, 95-115.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*. 37, 2523-2542.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. Annals of Statistics, 31(6):1852-1884.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Eubank, R. L. and Speckman, P. L.(1993). Confidence bands in nonparametric regression. *Journal of American and Statisical Association*, 88, 1287-1301.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. Chapmam and Hall, London.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli, 11, 1031-1057.

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *Annals of Statistics*. 29, 153-193.

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.

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

Goldfeld, S. M. and Quandt, R. E. (1973). A Markov Model for Switching Regression. *Journal of Econometrics*. 1, 3-15

Green, P. J. and Richardson, S. (2002). Hidden Markov Models and Disease Mapping. *Journal of American and Statisical Association*. 97, 1055-1070.

Härdle, W. and Bowman, A. W. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. Journal of the American Statistical Association, 83(401):102- 110.

Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. Annals of Statistics, 19(2):778-796.

Hathaway, R. J. (1985). A Constrained Formulation of Maximum-likelihood Estimation for Normal Mixture Distributions. *Annals of Statistics*. 13, 795-800.

Hennig, C. (2000). Identifiability of Models for Clusterwise Linear Regression. *Journal of Classification*. 17, 273-296.

Huang, M. (2009). *Nonparametric Techniques in Mixture of Regression Models.* Ph.D. Dissertation, The Pennsylvania State University.

Huang, M. and Li, R. (2010). *Nonparametric mixture of regression models.* Submitted.

Hunsberger, S. (1994). Semiparametric regression in likelihood-based models. *Journal of American and Statisical Association*. 89, 1354-1365.

Hurn, M., Justel, A., and Robert, C. (2003). Estimating Mixture of Regressions. *Journal of Computational and Graphical Statistics*. 12, 55-79.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*. 6, 181C214.

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

Li, R. and Liang, H. (2008). Variable Selection in Semiparametric Modeling. *Annals of Statistics*. 36, 261-286.

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

Neumann, M. H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. Journal of Nonparametric Statistics, 9:307-333.

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

Severini, T. A. and Staniswalis, J. G. (1994). Quasilikelihood estimation in semiparametric models. *Journal of American and Statisical Association*. 89, 501-511.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of Royal Statistical Society, Ser B.*, 62, 795-809.

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

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.

Wang, P., Puterman, M. L., Cockburn, I., and Le, N. (1996). Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics*. 52, 381-400.

Xia, Y. C. (1998). Bias-corected confidence bands in nonparametric regression. *Journal of American and Statisical Association*, 60, 797-811.

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

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