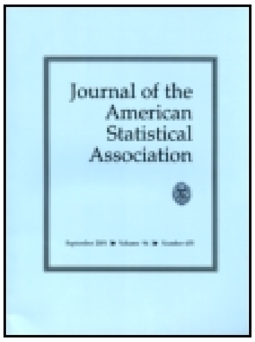


Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Nonparametric Mixture of Regression Models

Mian Huang , Runze Li & Shaoli Wang

Nonparametric Mixture of Regression Models

Mian HUANG, Runze LI, and Shaoli WANG

Motivated by an analysis of U.S. house price index (HPI) data, we propose nonparametric finite mixture of regression models. We study the identifiability issue of the proposed models, and develop an estimation procedure by employing kernel regression. We further systematically study the sampling properties of the proposed estimators, and establish their asymptotic normality. A modified EM algorithm is proposed to carry out the estimation procedure. We show that our algorithm preserves the ascent property of the EM algorithm in an asymptotic sense. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed estimation procedure. An empirical analysis of the U.S. HPI data is illustrated for the proposed methodology.

KEY WORDS: EM algorithm; Kernel regression; Mixture models; Nonparametric regression.

1. INTRODUCTION

Mixture models have been widely used in econometrics and social science, and the theories for mixture models have been well studied (Lindsay 1995). As a useful class of mixture models, finite mixture of linear regression models have been applied in various fields in the literature since its introduction by Goldfeld and Quandt (1973). For example, there are applications in econometrics and marketing (Wedel and DeSarbo 1993; Frühwirth-Schnatter 2001; Rossi, Allenby, and McCulloch 2005), in epidemiology (Green and Richardson 2002), and in biology (Wang et al. 1996). Bayesian approaches for mixture regression models are summarized in the article by Frühwirth-Schnatter (2006). Many efforts have been made to these models and their extensions such as finite mixture of generalized linear models (Hurn, Justel, and Robert 2003).

Motivated by an analysis of U.S. HPI data in Section 5, we propose nonparametric finite mixture of regression models. Compared with finite mixture of linear regression models, the newly proposed models relax the linearity assumption on the regression functions, and allow the regression function in each component to be an unknown but smooth function of covariates. In this article, we consider the situation in which the mixing proportions, the mean functions, and the variance functions are all nonparametric ones. Under certain conditions, we first show that the proposed model is identifiable. To estimate the unknown functions, we develop an estimation procedure via local-likelihood approach. Local-likelihood estimation (Tibshirani and Hastie 1987) extends the idea of nonparametric kernel regression to likelihood-based regression models. Fan, Heckman, and Wand (1995) studied local polynomial regression in quasi-likelihood model. Aerts and Claeskens (1997) studied multiparameter local-likelihood model. Fan and Gijbels (1996) gave a comprehensive account on this method.

For any estimation procedure of nonparametric functions, it is desirable to estimate the whole curves over a set of grid points. One may naively implement an EM algorithm (Dempster, Laird, and Rubin 1977) by maximizing each of the local-likelihood functions. However, the naive implementation of the EM algorithm does not ensure that the component labels match correctly at different grid points. This is similar to the label-switching problem in previous applications of mixture modeling (Stephens 2000; Yao and Lindsay 2009). To solve the problem, we modify the EM algorithm to simultaneously maximize the local-likelihood functions at a set of grid points. We further show that the modified EM algorithm possesses the monotone ascent property enjoyed by the ordinary EM algorithm in an asymptotic sense. The modified EM algorithm works well in our simulations and a real-data analysis.

The sampling properties of the proposed estimation procedure are investigated. We derive the asymptotic bias and variance of the local-likelihood estimate, and establish its asymptotic normality. To select the number of components, we consider implementing the information criterion approach. A bandwidth selector is proposed for the local-likelihood estimate using a multifold cross-validation (CV) method. We use a bootstrap method to obtain the standard error of the resulting estimate. Numerical simulations are conducted to examine the performance of the proposed procedure and test the accuracy of the proposed standard error estimation method. We further demonstrate the proposed model and estimation procedure by an empirical analysis of U.S. HPI data.

The rest of this article is structured as follows. In Section 2, we present the nonparametric finite mixture of regression models, and then derive the identifiability result. In Section 3, we further develop an estimation procedure for the proposed model using kernel regression and a modified EM algorithm. Model selection problems are addressed in Section 4. Simulation results and an empirical analysis of a real dataset are presented in Section 5. Some discussions are provided in Section 6. Technical conditions and proofs are given in the Appendix.

Mian Huang is Associate Professor, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, P. R. China (E-mail: *huang.mian@shufe.edu.cn*). Runze Li is the corresponding author and Distinguished Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111 (E-mail: *rzli@psu.edu*). Shaoli Wang is Associate Professor, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, P. R. China (E-mail: *swang@shufe.edu.cn*). Huang's and Wang's research is supported by a funding through Projet 211 Phase 4 of SHUFE, and Shanghai Leading Academic Discipline Project, B803. Li's research was supported by National Institute on Drug Abuse (NIDA) grants R21 DA024260 and P50-DA10075 and National Natural Science Foundation of China grant 11028103. The authors thank the editor, the AE, and the reviewers for their constructive comments, which have led to a dramatic improvement of the earlier version of this article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NIDA.

Journal of the American Statistical Association
September 2013, Vol. 108, No. 503, Theory and Methods
DOI: 10.1080/01621459.2013.772897

2. MODEL DEFINITION AND IDENTIFIABILITY

2.1 Model Definition

Assume that $\{(X_i, Y_i), i = 1, \ldots, n\}$ is a random sample from the population (X, Y). Throughout this article, we assume covariate X is univariate. The proposed methodology and theoretical results can be extended to multivariate covariate X, but the extension is less useful due to the "curse of dimensionality." Let $\mathcal{C}$ be a latent class variable. We assume that for $c = 1, 2, \ldots, C$, $\mathcal{C}$ has a discrete distribution $P(\mathcal{C} = c|X = x) = \pi_c(x)$ given $X = x$. Conditioning on $\mathcal{C} = c$ and $X = x$, Y follows a normal distribution with mean $m_c(x)$ and variance $\sigma_c^2(x)$. We further assume that $\pi_c(\cdot)$, $m_c(\cdot)$, and $\sigma_c^2(\cdot)$ are unknown but smooth functions. Hence, conditioning on $X = x$, Y follows a finite mixture of normals:

$$Y|_{X=x} \sim \sum_{c=1}^{C} \pi_c(x) N\left\{m_c(x), \sigma_c^2(x)\right\}. \tag{2.1}$$

In this article, we assume that C is fixed, and refer to model (2.1) as a nonparametric finite mixture of regression models because $\pi_c(\cdot)$, $m_c(\cdot)$, and $\sigma_c^2(\cdot)$ are nonparametric. When $C = 1$, model (2.1) is a nonparametric regression model. When $\pi_c(x)$ and $\sigma_c^2(x)$ are constant, and $m_c(x)$ is linear in x, model (2.1) reduces to a finite mixture of linear regression models (Goldfeld and Quandt 1973). Thus, model (2.1) can be regarded as a natural extension of nonparametric regression models and finite mixture of linear regression models. Although the component distribution in Equation (2.1) is assumed to be normal in this article, it is easy to extend our results to different parametric families for the component distribution.

Huang and Yao (2012) studied a semiparametric mixture of regression models with the mixing proportions being smooth functions of a covariate. Their model assumes that given $\mathbf{x}$ and $Z = z$, Y follows a finite mixture of normals:

$$Y|_{\mathbf{x}, Z=z} \sim \sum_{c=1}^{C} \pi_c(z) N\left(\alpha_c + \mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2\right). \tag{2.2}$$

The linearity assumption in each component allows for multivariate predictor $\mathbf{x}$. This would remarkably widen the application of model (2.2), since any smooth regression function may be approximated by a linear basis expansion of derived variables. However, a comprehensive study of nonparametric mixture of regression models (2.1) is of great importance and interest. Model (2.1) provides a general framework for mixture models. When the predictor $\mathbf{x}$ is one-dimensional, and $Z = \mathbf{x}$, model (2.2) is a restricted version of model (2.1). Theoretical study of identifiability shall offer justification for model estimation, including the aforementioned methodology of basis approximation for smooth regression functions. Model (2.1) also provides some guidelines for further research on more complex mixture models, for example, mixture models with each component being a varying coefficient model, or a nonparametric additive model, or a partial linear model, etc.

2.2 Identifiability

Identifiability is a critical issue for most mixture models. The identifiability of finite mixture distributions was studied in detail in Section 3.1 of the literature by Titterington et al. (1985). Hennig (2000) and section 8.2.2 of the literature by Frühwirth-Schnatter (2006) investigated the identifiability of finite mixture of regression models. There are useful related results, for example, mixture of univariate normals is identifiable up to relabeling, and finite mixture of regression models is identifiable up to relabeling provided that covariates have a certain level of variability. To derive the identifiability result for model (2.1), we first introduce the concept of transversality.

Definition. Let $x \in \mathbb{R}$, and let $\mathbf{a}(x)$ and $\mathbf{b}(x)$ be two smooth curves in $\mathbb{R}^2$. That is, $\mathbf{a}(x) = (a_1(x), a_2(x))$, $\mathbf{b}(x) = (b_1(x), b_2(x))$, and $a_i(x)$, $b_i(x)$ are differentiable, $i = 1, 2$. We say that $\mathbf{a}(x)$ and $\mathbf{b}(x)$ are *transversal* if $\|\mathbf{a}(x) - \mathbf{b}(x)\|^2 + \|\mathbf{a}'(x) - \mathbf{b}'(x)\|^2 \neq 0$, for any $x \in \mathbb{R}$.

The transversality of two smooth curves $\mathbf{a}(x)$ and $\mathbf{b}(x)$ implies that if $\mathbf{a}(x) = \mathbf{b}(x)$, then $\mathbf{a}'(x) \neq \mathbf{b}'(x)$. In other words, we impose a condition that the mean and variance functions of any two components cannot be tangent to each other. For more complex structures, the identifiability issue deserves further consideration (see the discussion in Section 6).

Theorem 1. Assume that: (i) $\pi_c(x) > 0$ are continuous functions, and $m_c(x)$ and $\sigma_c^2(x)$ are differentiable functions, $c = 1, \ldots, C$; (ii) any two curves $(m_i(x), \sigma_i^2(x))$ and $(m_j(x), \sigma_j^2(x))$, $i \neq j$, are transversal; and (iii) the range $\mathcal{X}$ of x is an interval in $\mathbb{R}$. Then model (2.1) is identifiable.

3. ESTIMATION PROCEDURE

3.1 Local-Likelihood Estimation

Denote by $\phi(y|\mu, \sigma^2)$ the density function of $N(\mu, \sigma^2)$. The likelihood function for the collected data $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$ is

$$\mathcal{L} = \sum_{i=1}^{n} \log\left[\sum_{c=1}^{C} \pi_c(X_i)\phi\left\{Y_i|m_c(X_i), \sigma_c^2(X_i)\right\}\right]. \tag{3.1}$$

Note that $\pi_c(\cdot)$, $m_c(\cdot)$, and $\sigma_c^2(\cdot)$ are nonparametric functions. In this article, we will employ kernel regression for model (2.1). Suppose that we want to estimate the unknown functions at x. In kernel regression, we first use local constants π_c, σ_c^2, and m_c to approximate $\pi_c(x)$, $\sigma_c^2(x)$, and $m_c(x)$. Let $K_h(\cdot) = h^{-1}K(\cdot/h)$ be a rescaled kernel of a kernel function $K(\cdot)$ with a bandwidth h. The corresponding local log-likelihood function for data $\{(X_i, Y_i): i = 1, 2, \ldots, n\}$ is

$$\ell_n(\boldsymbol{\pi}, \boldsymbol{\sigma}^2, \mathbf{m}; x) = \sum_{i=1}^{n} \log\left\{\sum_{c=1}^{C} \pi_c \phi\left(Y_i|m_c, \sigma_c^2\right)\right\} K_h(X_i - x), \tag{3.2}$$

where $\mathbf{m} = (m_1, \ldots, m_C)^T$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_C^2)^T$, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{C-1})^T$. One may also apply local linear regression or local polynomial regression techniques for the estimation of $\pi_c(x)$, $m_c(x)$, and $\sigma_c^2(x)$. Local linear regression has several nice statistical properties (Fan and Gijbels 1996). However, in our model setting, local linear regression does not yield a closed-form solution for variance functions and mixing proportion functions in the M-step of the proposed EM algorithm (see Equations (3.7), (3.8), and (3.9) for more details).

3.2 Asymptotic Properties

Let $\{\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\sigma}}^2, \tilde{\mathbf{m}}\}$ be the maximizer of the local-likelihood function (3.2). Then the estimates of $\pi_c(x)$, $\sigma_c^2(x)$, and $m_c(x)$ are

$$\tilde{\pi}_c(x) = \tilde{\pi}_c, \quad \tilde{\sigma}_c^2(x) = \tilde{\sigma}_c^2, \quad \text{and} \quad \tilde{m}_c(x) = \tilde{m}_c.$$

In this section, we study the asymptotic properties of $\tilde{\pi}_c(x)$, $\tilde{\sigma}_c^2(x)$, and $\tilde{m}_c(x)$. Let $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T, \mathbf{m}^T)^T$, and denote

$$\eta(y|\boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c \phi\left\{y|m_c, \sigma_c^2\right\}, \quad \ell(\boldsymbol{\theta}, y) = \log \eta(y|\boldsymbol{\theta}).$$

Let $\boldsymbol{\theta}(x) = \{\boldsymbol{\pi}^T(x), \boldsymbol{\sigma}^2(x)^T, \mathbf{m}(x)^T\}^T$, and denote

$$q_1\{\boldsymbol{\theta}(x), y\} = \frac{\partial \ell\{\boldsymbol{\theta}(x), y\}}{\partial \boldsymbol{\theta}}, \quad q_2\{\boldsymbol{\theta}(x), y\} = \frac{\partial^2 \ell\{\boldsymbol{\theta}(x), y\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T},$$
$$\mathcal{I}(x) = -\mathrm{E}[q_2\{\boldsymbol{\theta}(X), Y\}|X = x], \quad \text{and}$$
$$\Lambda(u|x) = \int_Y q_1\{\boldsymbol{\theta}(x), y\}\eta\{y|\boldsymbol{\theta}(u)\}dy.$$

Let $\gamma_n = (nh)^{-1/2}$, and for $c = 1, \ldots, C$, denote

$$\tilde{m}_c^* = \{\tilde{m}_c - m_c(x)\},$$
$$\tilde{\sigma}_c^{2*} = \left\{\tilde{\sigma}_c^2 - \sigma_c^2(x)\right\}.$$

For $c = 1, \ldots, C - 1$, denote

$$\tilde{\pi}_c^* = \{\tilde{\pi}_c - \pi_c(x)\}.$$

Let $\tilde{\mathbf{m}}^* = (\tilde{m}_1^*, \ldots, \tilde{m}_C^*)^T$, $\tilde{\boldsymbol{\sigma}}^{2*} = (\tilde{\sigma}_1^{2*} \ldots, \tilde{\sigma}_C^{2*})^T$, and $\tilde{\boldsymbol{\pi}}^* = (\tilde{\pi}_1^*, \ldots, \tilde{\pi}_{C-1}^*)^T$, and $\tilde{\boldsymbol{\theta}}^* = \{(\tilde{\boldsymbol{\pi}}^*)^T, (\tilde{\boldsymbol{\sigma}}^{2*})^T, (\tilde{\mathbf{m}}^*)^T\}^T$. The asymptotic bias, variance, and normality of the resulting estimators are given in the following theorem. The proof is given in the Appendix.

Theorem 2. Suppose that conditions (A)–(G) in the Appendix hold. It follows that

$$\sqrt{nh}\{\tilde{\boldsymbol{\theta}}^* - \mathcal{B}(x) + o(h^2)\} \xrightarrow{D} N\{0, \nu_0 f^{-1}(x)\mathcal{I}^{-1}(x)\},$$

where $f(\cdot)$ is the marginal density function of X, $\nu_0 = \int K^2(u)\,du$, and

$$\mathcal{B}(x) = \mathcal{I}^{-1}(x)\left\{\frac{f'(x)\Lambda_u'(x|x)}{f(x)} + \frac{1}{2}\Lambda_u''(x|x)\right\}\kappa_2 h^2,$$

with $\kappa_2 = \int u^2 K(u)\,du$.

3.3 An Effective EM Algorithm

For a given x, one may maximize the local-likelihood function (3.2) using an EM algorithm easily. In practice, we typically want to evaluate the unknown functions at a set of grid points over an interval of x, which requires us to maximize the local-likelihood function (3.2) at different grid points. However, if we naively implement the EM algorithm for each local model, we may suffer a problem similar to the label-switching problem of mixture models. Consider a two-component model with the mean function of one component consistently above the other, for example, solid lines in Figure 1(a). For any location x, a direct maximization of Equation (2.3) yields two sets of estimates due to label switching, that is, $\{\hat{m}_1(x), \hat{m}_2(x), \hat{\sigma}_1^2(x), \hat{\sigma}_2^2(x), \hat{\pi}_1(x)\}$, and $\{\hat{m}_2(x), \hat{m}_1(x), \hat{\sigma}_2^2(x), \hat{\sigma}_1^2(x), 1 - \hat{\pi}_1(x)\}$. Hence, for the three locations u_{j-1}, u_j, and u_{j+1} in Figure 1(a), there are eight possible configurations when we link the mean functions of the two components, while only two of them match the true model correctly. Directly maximizing Equation (2.3) at each location does not ensure correctly matching of the components that one lies consistently above the other. Based on our experience with simulation study, the naive EM algorithm results in very wiggly estimates for the $m_c(\cdot)$ and $\sigma_c^2(\cdot)$. This implies that the naive EM algorithm does not work at all.

In this section, we propose an effective EM algorithm to deal with the issue. We first introduce component labels for each of the observation, and define a set of local complete log-likelihood with the same labels. In the E-step of the EM algorithm, we estimate these labels. In the M-step, we simultaneously update

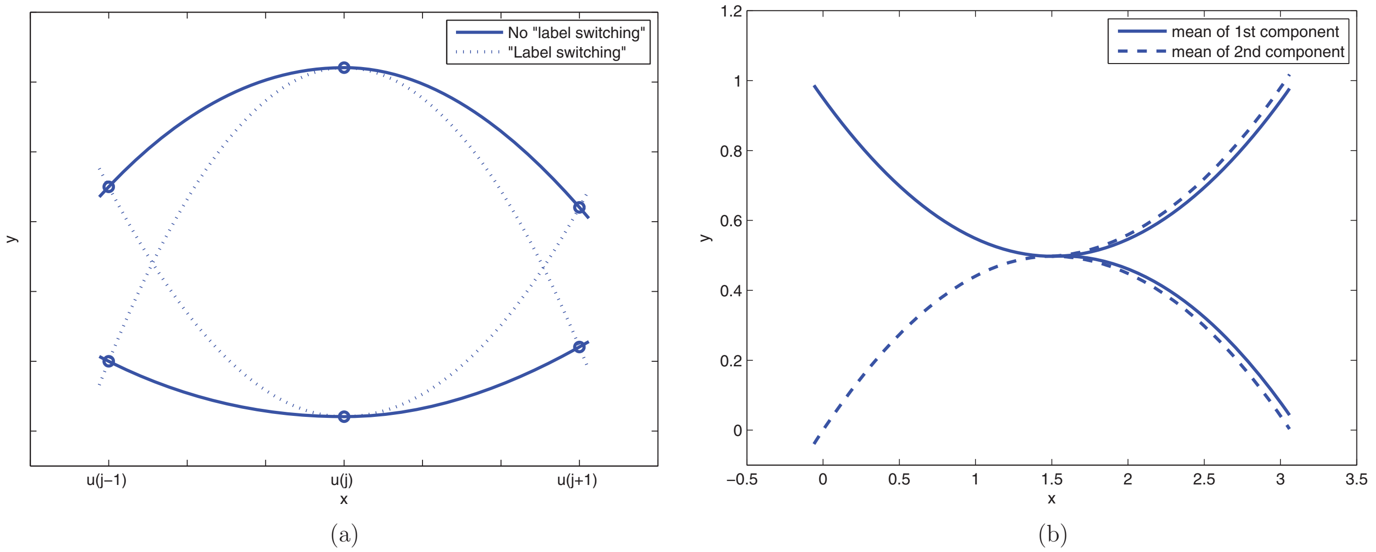


Figure 1. (a) An illustration of mismatching labels of naive implementation of EM algorithm. Solid curves are obtained when the labels at u_{j-1}, u_j, and u_{j+1} match correctly. Dotted curves are obtained when the labels at u_j does not match the ones at u_{j-1} and u_{j+1} correctly. (b) An illustration of a complex structure with crossed mean functions.

the estimated curves at all grid points for the same probabilistic label obtained in the E-step. This ensures that the resulting functional estimates are continuous and smooth at each iteration of the EM algorithm.

In the EM framework, the mixture problem is formulated as an incomplete-data problem. We view the observed data (X_i, Y_i)s as being incomplete, and introduce the unobserved Bernoulli random variables

$$z_{ic} = \begin{cases} 1, & \text{if } (X_i, Y_i) \text{ is in the } c\text{th group,} \\ 0, & \text{otherwise} \end{cases}$$

and $\mathbf{z}_i = (z_{i1}, \ldots, z_{iC})^T$, the associated component identity or label of (X_i, Y_i). The complete data are $\{(X_i, Y_i, \mathbf{z}_i), i = 1, 2, \ldots, n\}$, and the complete log-likelihood function corresponding to Equation (3.1) is

$$\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}\left[\log \pi_c(X_i) + \log \phi\left\{Y_i | m_c(X_i), \sigma_c^2(X_i)\right\}\right].$$

For $x \in \{u_1, \ldots, u_N\}$, the set of grid points at which the unknown functions are to be evaluated, define a local complete log-likelihood as

$$\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}\left[\log \pi_c + \log \phi\left\{Y_i | m_c, \sigma_c^2\right\}\right] K_h(X_i - x).$$

Note that z_{ic}'s do not depend on the choice of x. In the lth cycle of the EM algorithm iteration, we have $m_c^{(l)}(\cdot)$, $\sigma_c^{2(l)}(\cdot)$, and $\pi_c^{(l)}(\cdot)$. Then in the E-step of $(l+1)$-th cycle, the expectation of the latent variable z_{ic} is given by

$$r_{ic}^{(l+1)} = \frac{\pi_c^{(l)}(X_i)\phi\left\{Y_i | m_c^{(l)}(X_i), \sigma_c^{2(l)}(X_i)\right\}}{\sum_{c=1}^{C} \pi_c^{(l)}(X_i)\phi\left\{Y_i | m_c^{(l)}(X_i), \sigma_c^{2(l)}(X_i)\right\}}. \tag{3.3}$$

In the M-step of the $(l+1)$th cycle, we maximize

$$\sum_{i=1}^{n}\sum_{c=1}^{C} r_{ic}^{(l+1)}\left[\log \pi_c + \log \phi\left\{Y_i | m_c, \sigma_c^2\right\}\right] K_h(X_i - x), \tag{3.4}$$

for $x = u_j$, $j = 1, \ldots, N$. In practice, if n is not very large, one may choose the observed $\{X_1, \ldots, X_n\}$ to be the grid points. In such case, $N = n$.

The maximization of Equation (3.4) is equivalent to maximizing

$$\sum_{i=1}^{n}\sum_{c=1}^{C} r_{ic}^{(l+1)} \log \pi_c K_h(X_i - x), \tag{3.5}$$

and for $c = 1, \ldots, C$,

$$\sum_{i=1}^{n} r_{ic}^{(l+1)} \log \phi\left\{Y_i | m_c, \sigma_c^2\right\} K_h(X_i - x), \tag{3.6}$$

separately. For $x \in \{u_j, j = 1, \ldots, N\}$, the solution for maximization of Equation (3.5) is

$$\pi_c^{(l+1)}(x) = \frac{\sum_{i=1}^{n} r_{ic}^{(l+1)} K_h(X_i - x)}{\sum_{i=1}^{n} K_h(X_i - x)}, \tag{3.7}$$

and the closed-form solution for Equation (3.6) is

$$m_c^{(l+1)}(x) = \sum_{i=1}^{n} w_{ic}^{(l+1)}(x) Y_i \Big/ \sum_{i=1}^{n} w_{ic}^{(l+1)}(x), \tag{3.8}$$

$$\sigma_c^{2(l+1)}(x) = \frac{\sum_{i=1}^{n} w_{ic}^{(l+1)}(x)\left\{Y_i - m_c^{(l+1)}(x)\right\}^2}{\sum_{i=1}^{n} w_{ic}^{(l+1)}(x)}, \tag{3.9}$$

where $w_{ic}^{(l+1)}(x) = r_{ic}^{(l+1)} K_h(X_i - x)$. Furthermore, we update $\pi_c^{(l+1)}(X_i)$, $m_c^{(l+1)}(X_i)$, and $\sigma_c^{2(l+1)}(X_i)$, $i = 1, \ldots, n$, by linearly interpolating $\pi_c^{(l+1)}(u_j)$, $m_c^{(l+1)}(u_j)$, and $\sigma_c^{2(l+1)}(u_j)$, $j = 1, \ldots, N$, respectively. With initial values of $\pi_c(\cdot)$, $m_c(\cdot)$, and $\sigma^2(\cdot)$, the proposed estimation procedure is summarized in the following algorithm.

An EM algorithm:

Initial Value: Conduct a mixture of polynomial regressions with constant proportions and variances, and obtain the estimates of mean functions $\bar{m}_c(x)$, and parameters $\bar{\sigma}_c^2, \bar{\pi}_c$. Set the initial values $m_c^{(1)}(x) = \bar{m}_c(x)$, $\sigma^{2(1)}(x) = \bar{\sigma}_c^2$, and $\pi_c^{(1)}(x) = \bar{\pi}_c$.

E-step: Use Equation (3.3) to calculate $r_{ic}^{(l)}$ for $i = 1, \ldots, n$, and $c = 1, \ldots, C$.

M-step: For $c = 1, \ldots, C$ and $j = 1, \ldots, N$, evaluate $\pi_c^{(l+1)}(u_j)$ in (3.7), $m_c^{(l+1)}(u_j)$ in (3.8) and $\sigma_c^{2(l+1)}(u_j)$ in (3.9). Further obtain $\pi_c^{(l+1)}(X_i)$, $m_c^{(l+1)}(X_i)$ and $\sigma_c^{2(l+1)}(X_i)$ using linear interpolation.

Iteratively update the E-step and the M-step with $l = 2, 3, \ldots$, until the algorithm converges.

It is well known that an ordinary EM algorithm for parametric models possesses an ascent property, which is a desired property. The modified EM algorithm can be regarded as a generalization of the EM algorithm from parametric models to nonparametric ones. Thus, it is of interest to study whether the modified EM algorithm still preserves the ascent property.

Let $\boldsymbol{\theta}^{(l)}(\cdot) = \{\boldsymbol{\pi}^{(l)}(\cdot), \boldsymbol{\sigma}^{2(l)}(\cdot), \mathbf{m}^{(l)}(\cdot)\}$ be the estimated functions in the lth cycle of the proposed EM algorithm. We rewrite the local log-likelihood function (3.2) as

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, Y_i) K_h(X_i - x). \tag{3.10}$$

Theorem 3. For any given point x in the interval of $\mathcal{X}$, suppose that $\boldsymbol{\theta}^{(l)}(\cdot)$ has a continuous first derivative, $h \to 0$, and $nh \to \infty$ as $n \to \infty$. It follows that

$$\liminf_{n\to\infty}\; n^{-1}\left[\ell_n\left\{\boldsymbol{\theta}^{(l+1)}(x)\right\} - \ell_n\left\{\boldsymbol{\theta}^{(l)}(x)\right\}\right] \geq 0 \tag{3.11}$$

in probability.

4. MODEL SELECTION

Model selection for model (2.1) includes selection of the number of components C and the bandwidth h. For any given C, bandwidth selection is relatively easy as compared to the selection of C, and data-driven methods such as CV can be applied. However, selection of C in mixture models is challenging, and the situation becomes worse in the complicated setting (2.1).

We first investigate the methods for determining C in literature that can be adapted to our model.

4.1 Selection of the Number of Components

Two main approaches to determining the number of components are based on the likelihood function. One approach is to carry out a likelihood ratio test. Note that the regularity conditions for deriving a likelihood ratio test do not hold with order testing problem in mixture models, because the true parameter values of a smaller mixture model do not necessarily lie in the interior of the parameter space of the larger model. Many efforts have been made to overcome the difficulty, for example, the modified likelihood ratio test (Chen, Chen, and Kalbfleisch 2001), and the EM test (Li and Chen 2010). However, these procedures are developed for parametric mixture distributions and difficult to implement in more complex models such as model (2.1).

In this article, we focus on another likelihood-based approach: the information criterion approach. In general, an information criterion has the form

$$-2\mathcal{L} + \lambda \times df, \tag{4.1}$$

where $\mathcal{L}$ is the maximum log-likelihood of a specific model, and df is the degree of freedom, which accounts for the model complexity. The first term is a measurement of goodness of fit, and the second term is a penalty for model complexity. Two popular criteria, AIC and BIC, are obtained by setting $\lambda = 2$ and $\lambda = \log(n)$, respectively. For comparison, we need to estimate the mixture model under a misspecified C. When C is less than the true number of components, model bias could be large, which may lead to a smaller value of the log-likelihood. When C is greater than the true number of components, overfitting of the number of components occurs in model (2.1). As an illustration, the mixture likelihood of a model with $C-1$ distinct components is equal to the likelihood of a C-component model, in which either one component has zero proportion or two components are identical. This implies that the larger model is nonidentifiable, and the likelihood function is irregular. Nevertheless, information approaches have been investigated and applied in mixture models, as the first term of Equation (4.1) relates only to the likelihood. In practice, we may employ the EM algorithm for estimation and benefit from its ascent property. Feng and McCulloch (1996) showed that in parametric mixture density setting, the maximum likelihood estimator converges to a point belonging to a set of nonidentifiable parameter values that characterize the true density. Leroux (1992) proved that in the context of finite mixture distribution, using AIC and BIC would not underestimate the true number of components asymptotically. There are encouraging results for applications of BIC in mixture models, while AIC tends to overestimate the true number of components. Detailed reviews of applications of information criteria in finite mixture models are summarized in the literature by Frühwirth-Schnatter (2006) and McLachlan and Peel (2000).

For model (2.1), we first estimate the unknown functions and evaluate the likelihood in Equation (3.1). To implement the information criteria, we need to assess the model complexity. Here we consider the degree of freedom derived by Fan, Zhang, and Zhang (2001), which is originally developed for testing hypotheses on nonparametric functions. The degree of freedom of a one-dimensional varying coefficient function is

$$df = \tau_K h^{-1}|\Omega| \left\{ K(0) - \frac{1}{2}\int K^2(t)dt \right\},$$

where Ω is the support of the varying-coefficient covariate, and

$$\tau_K = \frac{K(0) - \frac{1}{2}\int K^2(t)dt}{\int \left\{K(t) - \frac{1}{2}K * K(t)\right\}^2 dt}.$$

This definition is analogous to the number of parameters in piecewise constant approximation. Fan, Zhang, and Zhang (2001) remarked that in the local polynomial fitting the result holds if we replace K by its equivalent kernel. As in model (2.1), the degree of freedom is $(3C-1) \times df$, with Ω replaced by the support of x. Note that the degree of freedom depends on both C and bandwidth in our model setting. We propose applying the information criteria under a wide range of bandwidths, and comparing the minimum scores of models with different number of components.

4.2 Bandwidth Selection

Bandwidth selection is a fundamental issue in nonparametric smoothing. For given C, we propose a multifold CV method to choose the bandwidth. Denote by $\mathcal{D}$ the full dataset. Then we randomly partition $\mathcal{D}$ into a training set $\mathcal{R}_j$, and a test set $\mathcal{T}_j$, $j = 1, \ldots, J$. Based on the data in training set $\mathcal{R}_j$ we obtain estimates $\{\hat{m}_c(\cdot), \hat{\sigma}_c^2(\cdot), \hat{\pi}_c(\cdot)\}$, and evaluate $m_c(\cdot)$, $\sigma_c^2(\cdot)$, and $\pi_c(\cdot)$ for the data in the corresponding testing set.

Then we calculate the probability of membership in test set $\mathcal{T}_j$. For $(x_l, y_l) \in \mathcal{T}_j$, $c = 1, \ldots, C$,

$$\hat{r}_{lc} = \frac{\hat{\pi}_c(x_l)\phi\left\{y_l | \hat{m}_c(x_l), \hat{\sigma}_c^2(x_l)\right\}}{\sum_{q=1}^C \hat{\pi}_q(x_l)\phi\left\{y_l | \hat{m}_q(x_l), \hat{\sigma}_q^2(x_l)\right\}}.$$

Now we can implement the regular CV criterion in the proposed model, that is,

$$\mathrm{CV} = \sum_{j=1}^J \sum_{l \in \mathcal{T}_j} (y_l - \hat{y}_l)^2, \tag{4.2}$$

where $\hat{y}_l = \sum_{c=1}^C \hat{r}_{lc}\, \hat{m}_c(x_l)$ is the predicted value for y_l in the test set $\mathcal{T}_j$. We select the bandwidth that minimizes CV.

Remark. The selection of C and h might affect each other. In practice, we would suggest first choosing C by minimizing the information criterion score over both a set of possible values for C and a wide range values for h (i.e., a set of two-dimensional grid points). After determining C, we choose h by minimizing the CV score defined in Equation (4.2).

5. SIMULATION AND APPLICATION

In this section, we address some practical implementation issues such as standard error estimation for our model. To assess the performance of the estimators of the unknown regression functions $m_c(x)$, we consider the square root of the average squared errors (RASE) of estimators for the unknown functions.

Table 1. Frequencies of selected C's by BIC

$n = 200$	$h = 0.06$	$h = 0.10$	$h = 0.14$	$h = 0.18$	$h = 0.22$	$\min_h \text{BIC}_h$
$C = 1$	986	520	89	22	18	17
$C = 2$	1	465	907	972	979	961
$C = 3$	9	7	1	3	0	11
$C = 4$	3	4	0	0	0	3
$C = 5$	1	4	3	3	3	8
$n = 400$	$h = 0.05$	$h = 0.08$	$h = 0.11$	$h = 0.14$	$h = 0.17$	$\min_h \text{BIC}_h$
$C = 1$	877	8	0	0	0	0
$C = 2$	120	990	998	998	1000	997
$C = 3$	3	0	0	0	0	0
$C = 4$	0	0	0	1	0	1
$C = 5$	0	2	2	1	0	2
$n = 800$	$h = 0.04$	$h = 0.06$	$h = 0.08$	$h = 0.10$	$h = 0.12$	$\min_h \text{BIC}_h$
$C = 1$	5	0	0	0	0	0
$C = 2$	995	1000	999	999	999	998
$C = 3$	0	0	0	0	0	0
$C = 4$	0	0	0	0	0	0
$C = 5$	0	0	1	1	1	2

For the mean functions,

$$\text{RASE}_m^2 = N^{-1} \sum_{c=1}^{C} \sum_{j=1}^{N} \{\hat{m}_c(u_j) - m_c(u_j)\}^2,$$

where $\{u_j, j = 1, \ldots, N\}$ are the grid points taken evenly in the range of covariate x. Similarly, we can define RASE for variance functions $\sigma_c^2(x)$s and proportion functions $\pi_c(x)$s, denoted by RASE_σ and RASE_π, respectively.

We use a bootstrap procedure to estimate the standard errors, and construct pointwise confidence intervals for the unknown functions. For given x_i, we can generate bootstrapped data Y_i^* from the distribution $\sum_{c=1}^{C} \hat{\pi}_c(x_i) N\{\hat{m}_c(x_i), \hat{\sigma}_c^2(x_i)\}$. By applying our estimation procedure for each of the bootstrap samples, we obtain the standard errors and confidence intervals.

5.1 Simulation Study

Example 1. In this example, we conduct a simulation for a two-component nonparametric mixture of regression model with

$$\begin{aligned} &\pi_1(x) = \exp(0.5x)/\{1 + \exp(0.5x)\}, \\ &\qquad\qquad\qquad\qquad \text{and} \quad \pi_2(x) = 1 - \pi_1(x), \\ &m_1(x) = 3 - \sin(2\pi x), \quad \text{and} \quad m_2(x) = \cos(3\pi x), \\ &\sigma_1(x) = 0.6\exp(0.5x), \quad \text{and} \quad \sigma_2(x) = 0.5\exp(-0.2x). \end{aligned}$$

We generate the predictor x from one-dimensional uniform distribution on [0, 1], and set the number of grid points $N = 100$. The Epanechnikov kernel is used in our simulation. It is well known that the EM algorithm may be trapped by local maximizers and thus is sensitive to initial values. To obtain good initial values, we first fit a mixture of polynomial regression models, and obtain the estimates of mean functions $\bar{m}_c(x)$, and parameters $\bar{\sigma}_c^2, \bar{\pi}_c$. The order of polynomial regression for each mean function is set to be 5. Then we set the initial values $m_c^{(1)}(x) = \bar{m}_c(x), \sigma^{2(1)}(x) = \bar{\sigma}_c^2$, and $\pi_c^{(1)}(x) = \bar{\pi}_c$. Based on our limited simulation experience, our procedure with initial values given by an overfitting (high polynomial order) mixture of polynomial regression model performs almost as well as those with true values as initial values, and order 5 works well in the simulation setting.

We first test the performance of information criterion (4.1) in selecting the number of components under BIC. For each dataset, we fit the nonparametric mixture of regression models with 1, 2, 3, 4, and 5 components under five different bandwidths, and then compare the information scores. The sets of bandwidths cover the cases of undersmoothing, appropriate smoothing, and oversmoothing (see the next paragraph and bandwidths in Table 2). For a given bandwidth h, we report the frequencies of selected C's over 1000 simulations in Table 1, from which it can be seen that undersmoothing may yield an underestimated C. For each simulated dataset, we should select C by minimizing the BIC score over the five C's and the five bandwidths. The frequencies of such selected C's over 1000 simulations are depicted in the last column of Table 1, from which the proportions of BIC choosing the correct model (i.e., two-component model) in the 1000 simulations are 96.1%, 99.7%, and 99.8% for $n = 200$, 400, and 800, respectively. This result shows that Equation (4.1) using BIC works reasonably

Table 2. Mean and standard deviation of RASEs

n	h	RASE_m	RASE_{σ^2}	RASE_π
200	0.067	0.328 (0.067)	0.560 (0.071)	0.113 (0.023)
	0.10	0.315 (0.074)	0.506 (0.073)	0.099 (0.026)
	0.15	0.388 (0.092)	0.455 (0.080)	0.097 (0.033)
400	0.053	0.252 (0.042)	0.501 (0.053)	0.090 (0.017)
	0.08	0.234 (0.049)	0.461 (0.057)	0.077 (0.018)
	0.12	0.283 (0.064)	0.427 (0.062)	0.079 (0.026)
800	0.04	0.195 (0.028)	0.463 (0.040)	0.073 (0.012)
	0.06	0.174 (0.032)	0.436 (0.044)	0.062 (0.013)
	0.09	0.195 (0.046)	0.414 (0.046)	0.059 (0.018)

Table 3. Standard error via bootstrap ($n = 200$, $h = 0.10$)

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$m_1(\cdot)$	SD	0.265	0.195	0.186	0.189	0.207	0.218	0.222	0.220	0.212
	SE	0.199	0.172	0.164	0.171	0.185	0.202	0.204	0.201	0.201
	Std	0.048	0.040	0.035	0.034	0.037	0.044	0.044	0.040	0.040
$m_2(\cdot)$	SD	0.197	0.183	0.134	0.146	0.169	0.143	0.139	0.164	0.148
	SE	0.160	0.152	0.134	0.132	0.141	0.131	0.123	0.141	0.137
	Std	0.047	0.037	0.035	0.034	0.033	0.032	0.031	0.034	0.038
$\sigma_1(\cdot)$	SD	0.213	0.179	0.176	0.176	0.213	0.261	0.282	0.260	0.286
	SE	0.181	0.159	0.151	0.165	0.196	0.236	0.257	0.256	0.249
	Std	0.062	0.060	0.055	0.057	0.070	0.085	0.090	0.089	0.091
$\sigma_2(\cdot)$	SD	0.091	0.105	0.098	0.094	0.085	0.078	0.071	0.082	0.083
	SE	0.078	0.094	0.098	0.088	0.079	0.071	0.067	0.071	0.079
	Std	0.034	0.039	0.045	0.041	0.032	0.029	0.026	0.030	0.044
$\pi_1(\cdot)$	SD	0.119	0.095	0.084	0.086	0.089	0.090	0.090	0.088	0.084
	SE	0.102	0.091	0.086	0.085	0.086	0.087	0.086	0.084	0.084
	Std	0.015	0.010	0.010	0.009	0.009	0.010	0.010	0.009	0.010

well in our simulation setting. In the following simulation, we assume that the number of components C is known.

To demonstrate that the proposed procedure works quite well for a wide range of bandwidths, we use the following strategy rather than CV to determine the bandwidth for each generated sample in our simulation study. For a given sample size, we generate several simulated datasets, and then choose a bandwidth by CV for each generated dataset. The optimal bandwidth is taken to be the average of these selected bandwidths with rounding. Then we consider three different representative bandwidths of undersmoothing, appropriate smoothing, and oversmoothing cases: 2/3 of the optimal bandwidth, the optimal bandwidth, and 1.5 times the optimal bandwidth. We conduct 500 simulations with sample sizes $n = 200, 400$, and 800, respectively. Table 2 displays the mean and standard deviation of RASEs over the 500 simulations. From Table 2, we see that the performance of the proposed procedure is not sensitive to a wide range of bandwidths.

We next test the accuracy of the standard error estimation via a bootstrap method. Tables 3, 4, and 5 summarize the performance of the standard errors of the functional estimates at $x = 0.1, 0.2, \ldots, 0.9$. The standard deviation of 500 estimates, denoted by SD, can be viewed as the true standard errors. We then calculate the sample average and standard deviation of the 500 estimated standard errors using bootstrap, denoted by SE and Std in Tables 3, 4, and 5. The result shows that although underestimations are present in many cases, the proposed bootstrap procedure works reasonably well because the difference between the true value and the estimate is less than twice of the standard error of the estimate.

Now we illustrate the performance of the proposed procedure using a typical simulated sample with $n = 400$. This typical sample is selected to be the one with median of RASE$_m$ in the 500 simulation samples. Figure 2(a) shows the scatterplot of the typical sample data and the true mean functions. Before we conduct analysis using the nonparametric mixture of regression model for this dataset, we first determine the number of components C using the information approach developed in Section 4. For this typical dataset, we fit the data using model (2.1) with one, two, three, and four components under a set of bandwidths $\{0.05, 0.08, 0.11, 0.14, 0.17\}$, and then calculate their corresponding BIC scores. The smallest BIC score yields a

Table 4. Standard error via bootstrap ($n = 400$, $h = 0.08$)

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$m_1(\cdot)$	SD	0.193	0.153	0.134	0.156	0.155	0.178	0.167	0.169	0.176
	SE	0.160	0.136	0.130	0.136	0.146	0.159	0.161	0.160	0.164
	Std	0.032	0.027	0.023	0.022	0.025	0.030	0.029	0.028	0.027
$m_2(\cdot)$	SD	0.149	0.124	0.113	0.106	0.120	0.107	0.104	0.111	0.106
	SE	0.122	0.116	0.105	0.102	0.107	0.102	0.097	0.105	0.101
	Std	0.029	0.025	0.023	0.022	0.022	0.020	0.019	0.021	0.020
$\sigma_1(\cdot)$	SD	0.168	0.138	0.131	0.140	0.159	0.194	0.226	0.225	0.213
	SE	0.149	0.127	0.124	0.135	0.162	0.192	0.204	0.207	0.214
	Std	0.041	0.040	0.036	0.037	0.047	0.055	0.062	0.064	0.057
$\sigma_2(\cdot)$	SD	0.073	0.081	0.080	0.069	0.069	0.068	0.061	0.056	0.058
	SE	0.066	0.077	0.077	0.069	0.064	0.060	0.056	0.057	0.057
	Std	0.023	0.028	0.029	0.024	0.022	0.021	0.018	0.018	0.022
$\pi_1(\cdot)$	SD	0.092	0.075	0.071	0.068	0.074	0.069	0.069	0.069	0.070
	SE	0.083	0.071	0.068	0.068	0.068	0.070	0.069	0.067	0.067
	Std	0.011	0.007	0.006	0.007	0.006	0.007	0.006	0.007	0.007

Table 5. Standard error via bootstrap ($n = 800$, $h = 0.06$)

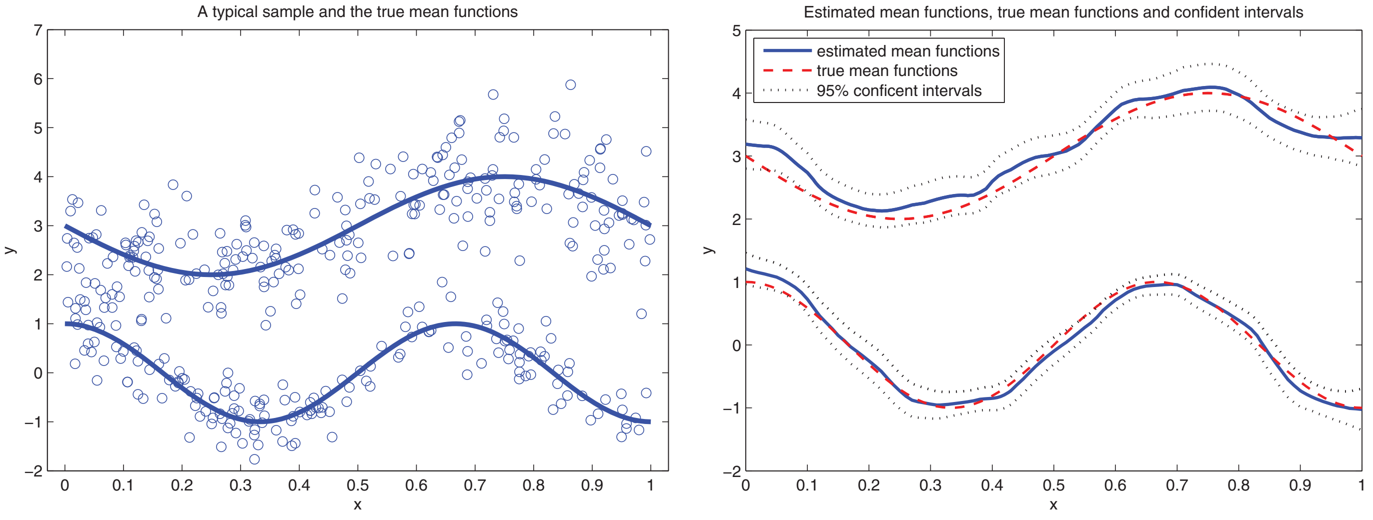
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$m_1(\cdot)$	SD	0.153	0.114	0.110	0.113	0.121	0.136	0.137	0.132	0.131
	SE	0.130	0.110	0.107	0.111	0.120	0.130	0.131	0.131	0.134
	Std	0.023	0.018	0.015	0.016	0.018	0.021	0.021	0.019	0.017
$m_2(\cdot)$	SD	0.116	0.093	0.083	0.086	0.092	0.085	0.076	0.086	0.082
	SE	0.098	0.092	0.081	0.081	0.085	0.082	0.079	0.082	0.080
	Std	0.019	0.017	0.013	0.013	0.015	0.014	0.013	0.013	0.013
$\sigma_1(\cdot)$	SD	0.132	0.112	0.104	0.116	0.138	0.167	0.184	0.172	0.176
	SE	0.119	0.104	0.105	0.113	0.134	0.160	0.168	0.171	0.174
	Std	0.030	0.028	0.024	0.026	0.032	0.039	0.043	0.039	0.036
$\sigma_2(\cdot)$	SD	0.072	0.069	0.056	0.052	0.060	0.049	0.051	0.046	0.046
	SE	0.059	0.063	0.058	0.054	0.054	0.050	0.048	0.046	0.046
	Std	0.017	0.019	0.015	0.014	0.016	0.013	0.013	0.012	0.013
$\pi_1(\cdot)$	SD	0.075	0.061	0.055	0.054	0.054	0.060	0.054	0.057	0.055
	SE	0.068	0.058	0.056	0.056	0.056	0.056	0.056	0.055	0.055
	Std	0.008	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005

two-component model. Thus, a two-component model is selected. We use the proposed CV criterion to select a bandwidth. The CV bandwidth selector yields the bandwidth 0.065. The resulting estimate of mean functions along with their pointwise confidence intervals are depicted in Figure 2(b), from which we can see that the true mean functions lies within the confidence interval for most points. This implies that the proposed estimation procedure performs quite well with moderate sample sizes.

Example 2. The goal is to compare model (2.1) with the mixture of linear regression models with varying proportions (Equation (2.2)), and the classical mixture of regression models. In this example, 500 random samples are generated from the three different scenarios: (a) model (2.1) with the same settings in Example 1; (b) model (2.2) with

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

which is designed in example 1 by Huang and Yao (2012); and (c) classical mixture of regression models with $\pi_1 = \pi_2 = 0.5$, while mean and variance functions are the same as those in scenario (b). The predictor x is taken from uniform distribution on [0, 1], and the grid points are evenly distributed with $N = 100$. For each sample we fit the data by model (2.1), model (2.2) with $z = x$, and the classical mixture of regression models. In estimation, we assumed that the number of components is fixed, and used Epanechnikov kernel function for smoothing. CV method is used to select bandwidths for model (2.2). We choose the bandwidths using procedure similar to the one used for Table 2 rather than using CV for each simulation, and only report the best bandwidth. The mean and standard deviation of RASEs for three models are recorded in Table 6. The lines beginning with "M1," "M2," and "M3" give the results of model (2.1), model (2.2), and classical mixture of regression models, respectively. In scenario (a), model (2.1) performs significantly better than other two models as expected. In scenarios (b) and (c), one may see how much efficiency is lost if the nonparametric approach is



(a)

(b)

Figure 2. (a) A typical sample of simulated data ($n = 400$), and the plot of true mean functions; (b) The estimated mean functions (solid lines), true mean functions (dashed lines), and 95% pointwise confidence intervals (dotted lines) with $n = 400$, $h = 0.065$.

Table 6. Mean and standard deviation of RASEs

	RASE_m	RASE_{σ^2}	RASE_π	RASE_m	RASE_{σ^2}	RASE_π
	Scenario (a), $n = 200$ ($h_{M_1} = 0.10, h_{M_2} = 0.10$)			$n = 400$ ($h_{M_1} = 0.08, h_{M_2} = 0.08$)		
M1	0.315 (0.074)	0.506 (0.073)	0.099 (0.027)	0.234 (0.049)	0.461 (0.057)	0.077 (0.018)
M2	3.461 (1.494)	0.831 (0.470)	0.239 (0.080)	3.687 (1.407)	0.819 (0.395)	0.241 (0.072)
M3	3.406 (1.517)	0.719 (0.439)	0.132 (0.075)	3.634 (1.438)	0.628 (0.247)	0.120 (0.053)
	Scenario (b), $n = 200$ ($h_{M_1} = 0.13, h_{M_2} = 0.08$)			$n = 400$ ($h_{M_1} = 0.12, h_{M_2} = 0.07$)		
M1	0.248 (0.085)	0.163 (0.043)	0.138 (0.047)	0.199 (0.064)	0.168 (0.032)	0.120 (0.038)
M2	0.083 (0.033)	0.150 (0.025)	0.107 (0.030)	0.060 (0.024)	0.153 (0.019)	0.082 (0.023)
M3	0.113 (0.042)	0.155 (0.027)	0.254 (0.005)	0.102 (0.032)	0.157 (0.020)	0.253 (0.004)
	Scenario (c), $n = 200$ ($h_{M_1} = 0.14, h_{M_2} = 0.11$)			$n = 400$ ($h_{M_1} = 0.13, h_{M_2} = 0.08$)		
M1	0.187 (0.085)	0.160 (0.029)	0.098 (0.040)	0.155 (0.077)	0.164 (0.021)	0.075 (0.032)
M2	0.079 (0.033)	0.153 (0.023)	0.100 (0.030)	0.057 (0.021)	0.155 (0.017)	0.084 (0.023)
M3	0.075 (0.031)	0.154 (0.023)	0.033 (0.023)	0.055 (0.020)	0.156 (0.016)	0.022 (0.018)

used for the mean, variance, and proportion functions. For the case $n = 200$ of scenario (b), the RASEs of model (2.1) have around two times increment for the mean function, and 29% increment for the proportion function, as compared to those of the model (2.2). Similar results are found for the case $n = 400$.

5.2 Application

In this section, we illustrate the proposed model and estimation procedure by an analysis of a real dataset, which contains the monthly change of S&P/Case-Shiller HPI and monthly growth rate of United States gross domestic product (GDP) from January 1990 to December 2002. Note that the observations are time series data, and may not be independent as assumed by model (2.1). We discussed this problem in Section 6. It is known that HPI is a measure of a nation's average housing price in repeat sales, and the S&P/Case-Shiller HPI uses a modified weighted method that may adjust for the quality of the houses; GDP is a 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. The housing sector plays an important role in the national economy, and the house price and GDP are interrelated. It is of interest to investigate the impact of GDP growth rate on HPI change. Hence, we set HPI change to be the response variable, and the GDP growth rate to be the predictor. The scatterplot of this dataset is depicted in Figure 3(a). As expected, the impact of GDP growth rate on HPI change may have different patterns in different macroeconomic cycles, which provides a connection to a mixture framework. In the analysis, we do not specify the cycle identities which the observations belong to, and treat the underlying cycle as a latent variable. Then we analyze the data by model (2.1) via the proposed estimation procedure.

We first determine the number of component C using the information approach. For a set of bandwidths $\{0.08, 0.11, 0.14, 0.17, 0.20\}$, the dataset is fitted with model (2.1) with one, two, three, and four components, respectively. Then we calculate and compare their BIC scores. For each of the five bandwidths, the minimum BIC score is achieved at $C = 2$, hence a two-component model is selected. This result suggests that likely there are two economic cycles from 1990 to 2002 as reflected by the relation between HPI change and GDP growth. We next select the bandwidth for the two-component model. An optimal bandwidth is selected at 0.11 by a five-fold CV selector described in Equation (4.2). With this selected bandwidth, we fit the data with a two-component nonparametric mixture of regression models. We choose $N = 100$ grid points evenly from the range of the predictor. The estimated mean functions, proportion functions, and variance functions with their 95% pointwise confidence intervals are shown in Figure 3(b), (c), and (d), respectively. We further depict the hard-clustering result in Figure 3(b), which is obtained by assigning component identities according to the largest r_{ic}, $c = 1, 2$. Together with the original data with actual calendar dates, it can be seen that the circle points from the lower cluster are mainly from January 1990 to September 1997. The triangle points in the upper cluster are mainly from October 1997 to December 2002, during which the economy experienced an Internet boom and bust. From the result, we observe that in the first cycle (lower component), GDP growth has a positive impact on HPI change; in the second cycle (upper component), HPI change tends to be lower when GDP growth is in the middle, as compared to the situations of both high and low GDP growth.

6. DISCUSSION

In this article, we proposed a class of nonparametric finite mixture of regression models, which allows the mixing proportions, the mean functions, and the variance functions all to be nonparametric functions of the covariate. We showed that the proposed models are identifiable under mild conditions. There are some structures not satisfying condition (ii) of Theorem 1, for example, consider the mean functions of a two-component structure as shown in Figure 1(b). If the variance functions of the two components are the same, then there are two solutions of mean functions: (i) $m_1(x)$ and $m_2(x)$ are tangent to each other; (ii) $m_1(x)$ is a monotone decreasing mean, and $m_2(x)$ a monotone increasing mean. It is unclear which paths the mean functions will follow without knowing the second derivatives of the mean functions at the "cross." Condition (ii) excludes such case, and the identifiability issue for model (2.1) deserves further research.

For the proposed nonparametric finite mixture of regression models, we focus on estimation when x is an interior point in the range of covariate. It is certainly of interest to study the boundary

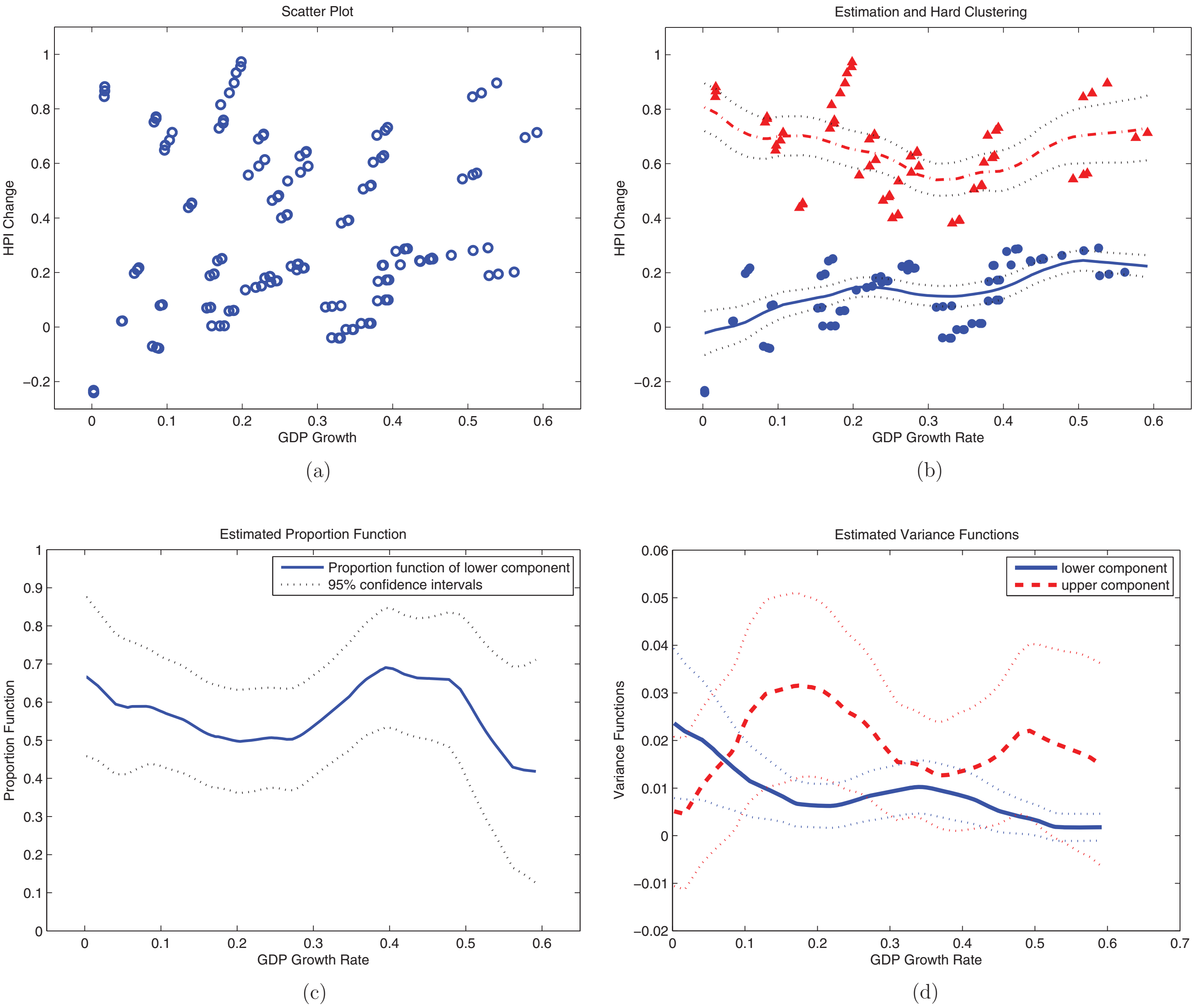


Figure 3. (a) Scatterplot of U.S. house price index data. (b) Estimated mean functions with 95% confidence intervals and a hard-clustering result. (c) Estimated mixing proportion function. (d) Estimated variance functions.

effect of the proposed procedure. The boundary effect has been studied in the article by Cheng, Fan, and Marron (1997) for the nonparametric regression model. It is interesting to conduct hypothesis test on whether the mixing proportions are constants, on whether some of the mean functions are constants or of specific parametric forms, and on whether the variance functions are parametric ones. These issues need further investigations.

In the real data analysis, we were analyzing time series data by model (2.1), and this may not fit the general setting of the article where the data are assumed to be independent. We have examined the autocorrelation of series of HPI change and GDP growth, and the sample autocorrelation functions show that they are not independent. An ADF test shows that both HPI and GDP series are trend stationary. For this type of time series data, we conjecture that under certain regularity conditions, the $\sqrt{nh}$ convergence rate of the estimators still holds, although the asymptotic bias and variance terms may have different forms. Extension of our methodology to time series data is an interesting problem and deserves further research.

APPENDIX: TECHNICAL CONDITIONS AND PROOFS

Proof of Theorem 1. Let us consider the subset of $\mathbb{R}$

$$S = \left\{x_k : \left(m_i(x_k), \sigma_i^2(x_k)\right) = \left(m_j(x_k), \sigma_j^2(x_k)\right) \text{ for some } i \neq j\right\},$$

where some parameter curves intersect. Since any two parameter curves are transversal, any point in S is an isolated point. This implies that set $S \subset \mathbb{R}$ has no limit point and contains at most countably many points. Therefore, without loss of generality, we assume that $x_k < x_{k+1}$ and $(x_k, x_{k+1}) \cap S = \emptyset$, $k = 0, \pm 1, \pm 2, \ldots$.

Assume that model (2.1) admits another representation

$$Y|X = x \sim \sum_{d=1}^{D} \lambda_d(x) N\left(\nu_d(x), \delta_d^2(x)\right),$$

where $\lambda_d(x) > 0$, $d = 1, \ldots, D$.

We know that the finite mixture of normal distributions is identifiable (see Titterington et al. 1985, p. 38, example 3.1.4). Hence, for any given $x \notin S$, model (2.1) is identifiable. It follows that $C = D$, and there exists a permutation $\omega_x = \{\omega_x(1), \ldots, \omega_x(C)\}$ of set $\{1, \ldots, C\}$

depending on x, such that

$$\lambda_{\omega_x(c)}(x) = \pi_c(x),\ \ \nu_{\omega_x(c)}(x) = m_c(x),\ \ \delta^2_{\omega_x(c)}(x) = \sigma_c^2(x),\quad c = 1, \dots, C. \tag{A.1}$$

Since all the parameter curves $(m_i(x), \sigma_i^2(x))$ are continuous, and no pair of parameter curves intersect on any interval (x_k, x_{k+1}), the permutation ω_x must be constant on (x_k, x_{k+1}). On the other hand, for any $x_k \in S$, consider a small neighborhood $(x_k - u, x_k + u)$ such that $(x_k - u, x_k + u) \subset (x_{k-1}, x_{k+1})$. Since any pair of parameter curves are transversal, they have different derivatives at x_k if they intersect at x_k, hence the permutation must be constant on the neighborhood $(x_k - u, x_k + u)$ since equation (A.1) implies the identity of derivatives of parameter curves on either side of x_k. Therefore, there exists a permutation $\omega = \{\omega(1), \dots, \omega(C)\}$ of set $\{1, \dots, C\}$ that is independent of x such that

$$\lambda_{\omega(c)}(x) = \pi_c(x),\ \ \nu_{\omega(c)}(x) = m_c(x),\ \ \delta^2_{\omega(c)}(x) = \sigma_c^2(x),\ \ c = 1, \dots, C.$$

This completes the proof of identifiability.

Now we outline the key steps for proofs of Theorems 2 and 3. Note that $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, \boldsymbol{\sigma}^{2T}, \mathbf{m}^T)^T$ is a $(3C - 1) \times 1$ vector. Whenever necessary, we rewrite $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{3C-1})^T$ without changing the order of $\boldsymbol{\pi}$, $\boldsymbol{\sigma}^2$, and $\mathbf{m}$. Otherwise, we will use the same notation as defined in Section 2.

Regularity Conditions

A. The sample $\{(X_i, Y_i),\ i = 1, \dots, n\}$ is independent and identically distributed from its population (X, Y). The support for X, denoted by $\mathcal{X}$, is a compact subset of $\mathbb{R}^1$.
B. The marginal density function $f(x)$ of X is twice continuously differentiable and positive for all $x \in \mathcal{X}$.
C. There exists a function $M(y)$, with $\mathrm{E}\{M(Y)\} < \infty$, such that for all y, and all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}(x)$, $|\partial^3 \ell(\boldsymbol{\theta}, y)/\partial\theta_j\partial\theta_k\partial\theta_l| < M(y)$.
D. The unknown functions $\boldsymbol{\theta}(x)$ have continuous second derivatives; moreover, for $c = 1, \dots, C$, $\sigma_c^2(x) > 0$, and $\pi_c(x) > 0$ hold for all $x \in \mathcal{X}$.
E. The kernel function $K(\cdot)$ has a bounded support, and satisfies that

$$\int K(u)du = 1,\quad \int uK(u)du = 0,\quad \int u^2K(u)du < \infty,$$
$$\int K^2(u)du < \infty,\quad \int |K^3(u)|du < \infty.$$

F. The following conditions hold for all i and j:

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

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

All these are mild conditions and have been used in the literature of local-likelihood estimation and mixture models. The following lemma will be used in the proof of Theorem 2.

Lemma 1. Under Conditions A, C, and F, for any interior point x of $\mathcal{X}$, it holds that

$$\mathrm{E}[q_1\{\boldsymbol{\theta}(X), Y\}|X = x] = 0, \tag{A.2}$$
$$\mathrm{E}[q_2\{\boldsymbol{\theta}(X), Y\}|X = x] = -\mathrm{E}[q_1\{\boldsymbol{\theta}(X), Y\}q_1^T\{\boldsymbol{\theta}(X), Y\}|X = x]. \tag{A.3}$$

Proof. Conditioning $X = x$, Y follows a finite mixture of normals. Thus, by some calculations, Equation (A.2) holds. Furthermore, Equation (A.3) follows from regularity conditions C, F and the arguments on page 39 of the literature by McLachlan and Peel (2000) together. This completes the proof of the lemma.

We refer to Equations (A.2) and (A.3) as the local Barlett's first and second identities, respectively. Equation (A.2) implies that $\Lambda(x|x) = 0$.

Proof of Theorem 2. For $c = 1, \dots, C$, denote

$$m_c^* = \sqrt{nh}\{m_c - m_c(x)\},$$
$$\sigma_c^{2*} = \sqrt{nh}\left\{\sigma_c^2 - \sigma_c^2(x)\right\}.$$

For $c = 1, \dots, C - 1$, denote

$$\pi_c^* = \sqrt{nh}\{\pi_c - \pi_c(x)\}.$$

Let $\mathbf{m}^* = (m_1^*, \dots, m_C^*)^T$, $\boldsymbol{\sigma}^{2*} = (\sigma_1^{2*} \dots, \sigma_C^{2*})^T$, and $\boldsymbol{\pi}^* = (\pi_1^*, \dots, \pi_{C-1}^*)^T$. Denote $\boldsymbol{\theta}^* = (\boldsymbol{\pi}^{*T}, \boldsymbol{\sigma}^{2*T}, \mathbf{m}^{*T})^T$. Recall that

$$\ell(\boldsymbol{\theta}(x), y) = \log \eta\{y|\boldsymbol{\theta}(x)\} = \log\left\{\sum_{c=1}^C \pi_c(x)\phi\left\{y|m_c(x), \sigma_c^2(x)\right\}\right\}.$$

Let

$$\ell(\boldsymbol{\theta}(x) + \gamma_n\boldsymbol{\theta}^*, y) = \log\left\{\sum_{c=1}^C(\pi_c(x) + \gamma_n\pi_c^*) \times \phi\left(y|m_c(x) + \gamma_n m_c^*, \sigma_c^2(x) + \gamma_n\sigma_c^{2*}\right)\right\}.$$

Thus, if $\{\tilde{\pi}, \tilde{\boldsymbol{\sigma}}^2, \tilde{\mathbf{m}}\}$ maximizes Equation (3.2), then $\tilde{\boldsymbol{\theta}}^*$ maximizes

$$\ell_n^*(\boldsymbol{\theta}^*) = h\sum_{i=1}^n\{\ell(\boldsymbol{\theta}(x) + \gamma_n\boldsymbol{\theta}^*, Y_i) - \ell(\boldsymbol{\theta}(x), Y_i)\}K_h(X_i - x). \tag{A.4}$$

By Taylor expansion,

$$\ell_n^*(\boldsymbol{\theta}^*) = \Delta_n\boldsymbol{\theta}^* + \frac{1}{2}\boldsymbol{\theta}^{*T}\Gamma_n\boldsymbol{\theta}^* + \frac{h\gamma_n^3}{6}\sum_{i=1}^n R(\boldsymbol{\theta}(x), \tilde{\xi}_i), \tag{A.5}$$

where $\tilde{\xi}_i = t_i\gamma_n\boldsymbol{\theta}^*$ for some $t_i \in (0, 1)$, and

$$\Delta_n = \sqrt{\frac{h}{n}}\sum_{i=1}^n q_1\{\boldsymbol{\theta}(x), Y_i\}K_h(X_i - x),$$
$$\Gamma_n = \frac{1}{n}\sum_{i=1}^n q_2\{\boldsymbol{\theta}(x), Y_i\}K_h(X_i - x),$$
$$R(\boldsymbol{\theta}(x), \tilde{\xi}_i) = \sum_{j,k,l}\frac{\partial^3\ell(\boldsymbol{\theta}(x) + \tilde{\xi}_i, Y_i)}{\partial\theta_j\partial\theta_k\partial\theta_l}K_h(X_i - x)\theta_j^*\theta_k^*\theta_l^*.$$

Denote by $\Gamma_n(i, j)$ the (i, j)th element of Γ_n. By condition E, it can be shown that

$$\mathrm{E}\Gamma_n(i, j) = \int_Y\int_X \frac{\partial^2\ell(\boldsymbol{\theta}(x), y)}{\partial\theta_i\partial\theta_j}\eta\{y|\boldsymbol{\theta}(u)\}f(u)K_h(u - x)dudy$$
$$= f(x)\int_Y \frac{\partial^2\ell(\boldsymbol{\theta}(x), y)}{\partial\theta_i\partial\theta_j}\eta\{y|\boldsymbol{\theta}(x)\}dy + o(1).$$

Therefore, $\mathrm{E}\Gamma_n = -f(x)\mathcal{I}(x) + o(1)$. $\mathrm{var}\{\Gamma_n(i, j)\}$ is dominated by the term

$$\frac{1}{n}\int_Y\int_X\left\{\frac{\partial^2\ell(\boldsymbol{\theta}(x), y)}{\partial\theta_i\partial\theta_j}\right\}^2\eta\{y|\boldsymbol{\theta}(u)\}f(u)K_h^2(u - x)dudy,$$

which can be shown to have the order $O\{(nh)^{-1}\}$ under condition F. Therefore, we have

$$\Gamma_n = -f(x)\mathcal{I}(x) + o_P(1).$$

By condition C, the expectation of the absolute value of the last term of Equation (A.5) is bounded by

$$O\left(\gamma_n \mathrm{E}\max_{j,k,l}\left|\frac{\partial^3 \ell(\boldsymbol{\theta}(x)+\tilde{\xi}, Y)}{\partial\theta_j\partial\theta_k\partial\theta_l}K_h(X_i-x)\right|\right)=O(\gamma_n). \quad \text{(A.6)}$$

Thus, the last term of Equation (A.5) is of order $O_P(\gamma_n)$. Therefore, we have

$$\ell_n^*(\boldsymbol{\theta}^*) = \Delta_n\boldsymbol{\theta}^* - \frac{1}{2}f(x)\boldsymbol{\theta}^{*T}\mathcal{I}(x)\boldsymbol{\theta}^* + o_P(1). \quad \text{(A.7)}$$

Using the quadratic approximation lemma (see, e.g., p. 210 of the literature by Fan and Gijbels 1996), we have

$$\hat{\boldsymbol{\theta}}^* = f(x)^{-1}\mathcal{I}(x)^{-1}\Delta_n + o_P(1). \quad \text{(A.8)}$$

To establish the asymptotic normality, it remains to calculate the mean and variance of Δ_n, and verify the Lyapounov condition. Note that

$$\begin{aligned}\mathrm{E}(\Delta_n) &= \sqrt{nh}\int_Y\int_X q_1\{\boldsymbol{\theta}(x), y\}\eta\{y|\boldsymbol{\theta}(u)\}f(u)K_h(u-x)du dy\\ &= \sqrt{nh}\int_X \Lambda(u|x)f(u)K_h(u-x)du.\end{aligned}$$

Under conditions C, D, and F, $\Lambda(u|x)$ has a continuous second derivative. Thus, using the fact $\Lambda(x|x)=0$ by Lemma 1 and standard arguments in kernel regression, it follows that

$$\mathrm{E}(\Delta_n) = \sqrt{nh}f(x)\left\{\frac{f'(x)\Lambda_u'(x|x)}{f(x)}+\frac{1}{2}\Lambda_u''(x|x)\right\}\kappa_2 h^2\{1+o(1)\}.$$

For the covariance term of Δ_n, we have

$$\mathrm{cov}(\Delta_n) = h\mathrm{E}\left\{q_1\{\boldsymbol{\theta}(x), Y\}q_1^T\{\boldsymbol{\theta}(x), Y\}K_h^2(X-x)\right\} + o(1),$$

where its (i, j)th element is

$$\begin{aligned}&h\int_Y\int_X \frac{\partial\ell(\boldsymbol{\theta}(x), y)}{\partial\theta_i}\frac{\partial\ell(\boldsymbol{\theta}(x), y)}{\partial\theta_j}K_h^2(u-x)f(u)\eta\{y|\boldsymbol{\theta}(u)\}du dy\\ &\xrightarrow{P} f(x)\nu_0\int_Y \frac{\partial\ell(\boldsymbol{\theta}(x), y)}{\partial\theta_i}\frac{\partial\ell(\boldsymbol{\theta}(x), y)}{\partial\theta_j}\eta\{y|\boldsymbol{\theta}(x)\}dy\\ &= -f(x)\nu_0\int_Y\frac{\partial^2\ell(\boldsymbol{\theta}(x), y)}{\partial\theta_i\partial\theta_j}\eta\{y|\boldsymbol{\theta}(x)\}dy.\end{aligned}$$

The last step holds due to Equation (A.3). Thus, $\mathrm{cov}(\Delta_n) = f(x)\mathcal{I}(x)\nu_0 + o(1)$. To establish the asymptotic normality for Δ_n, it is necessary to show that for any unit vector $\mathbf{d}$,

$$\{\mathbf{d}^T\mathrm{cov}(\Delta_n)\mathbf{d}\}^{-1/2}\mathbf{d}^T\{\Delta_n - \mathrm{E}(\Delta_n)\} \xrightarrow{D} N(0, 1). \quad \text{(A.9)}$$

Since $\mathrm{cov}(\Delta_n) = O(1)$, it follows that $\{\mathbf{d}^T\mathrm{cov}(\Delta_n)\mathbf{d}\}^{-1} = O(1)$. Let $\lambda_i = \mathbf{d}^T q_1\{\boldsymbol{\theta}(x), Y_i\}K_h(X_i-x)$, then $\mathbf{d}^T\Delta_n = h\gamma_n\sum_{i=1}^n\lambda_i$. Therefore, it is sufficient to show that $nh^3\gamma_n^3\mathrm{E}(|\lambda_i|^3) = o(1)$. By condition F and arguments similar to Equation (A.6), it can be shown that $nh^3\gamma_n^3\mathrm{E}(|\lambda_i|^3) = O(\gamma_n) = o(1)$, and thus the Lyapounov condition holds for Equation (A.9). By Equation (A.8) and the Slutsky theorem, we have

$$\sqrt{nh}\{\gamma_n\tilde{\boldsymbol{\theta}}^* - \mathcal{B}(x) + o(h^2)\} \xrightarrow{D} N\left\{0, \nu_0 f^{-1}(x)\mathcal{I}^{-1}(x)\right\}. \quad \text{(A.10)}$$

Proof of Theorem 3. We assume the unobserved data $(\mathcal{C}_i, i = 1, \ldots, n)$ are a random sample from population $\mathcal{C}$, and the complete data $\{(X_i, Y_i, \mathcal{C}_i), i = 1, 2, \ldots, n\}$ are a random sample from $(X, Y, \mathcal{C})$. Let $h\{y, c|\boldsymbol{\theta}(x)\}$ be the joint distribution of $(Y, \mathcal{C})$ given $X = x$, and $f(x)$ be the marginal density of X. Conditioning on $X = x$, Y follows a distribution $\eta\{y|\boldsymbol{\theta}(x)\}$. The local log-likelihood function (3.2) can be rewritten as

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n\log\{\eta(Y_i|\boldsymbol{\theta})\}K_h(X_i-x). \quad \text{(A.11)}$$

The conditional probability of $\mathcal{C} = c$ given y and $\boldsymbol{\theta}$ is

$$g\{c|y, \boldsymbol{\theta}\} = h(y, c|\boldsymbol{\theta})/\eta(y|\boldsymbol{\theta}) = \frac{\pi_c\phi\left(y|m_c, \sigma_c^2\right)}{\sum_{c=1}^C\pi_c\phi\left(y|m_c, \sigma_c^2\right)}. \quad \text{(A.12)}$$

For given $\boldsymbol{\theta}^{(l)}(X_i)$, $i = 1, \ldots, n$, it is clear that $\int g\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\}dc = 1$. Then we have

$$\begin{aligned}\ell_n(\boldsymbol{\theta}) &= \sum_{i=1}^n\log\{\eta(Y_i|\boldsymbol{\theta})\}\left\{\int g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\right\}dc\right\}K_h(X_i-x)\\ &= \sum_{i=1}^n\left\{\int\log\{\eta(Y_i|\boldsymbol{\theta})\}g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\right\}dc\right\}K_h(X_i-x).\end{aligned} \quad \text{(A.13)}$$

By Equation (A.12), we also have

$$\log\{\eta(Y_i|\boldsymbol{\theta})\} = \log\{h(Y_i, c|\boldsymbol{\theta})\} - \log\{g(c|Y_i, \boldsymbol{\theta})\}. \quad \text{(A.14)}$$

Thus,

$$\begin{aligned}\ell_n(\boldsymbol{\theta}) =& \sum_{i=1}^n\left\{\int\log\{h(Y_i, c|\boldsymbol{\theta})\}g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\right\}dc\right\}K_h(X_i-x)\\ &- \sum_{i=1}^n\left\{\int\log\{g(c|Y_i, \boldsymbol{\theta})\}g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\right\}dc\right\}K_h(X_i-x),\end{aligned} \quad \text{(A.15)}$$

where $\boldsymbol{\theta}^{(l)}(X_i)$ is the estimate of $\boldsymbol{\theta}(X_i)$ at the lth iteration. Taking expectation leads to calculating Equation (3.3). In the M-step, we update $\boldsymbol{\theta}^{(l+1)}(x)$ such that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\left\{\int\log\{h(Y_i, c|\boldsymbol{\theta}^{(l+1)}(x))\}g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\right\}dc\right\}K_h(X_i-x))\\ &\geq \frac{1}{n}\sum_{i=1}^n\left\{\int\log\{h(Y_i, c|\boldsymbol{\theta}^{(l)}(x))\}g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\right\}dc\right\}K_h(X_i-x)).\end{aligned}$$

It suffices to show that

$$\begin{aligned}&\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\left[\int\log\left\{\frac{g\left\{c|Y_i, \boldsymbol{\theta}^{(l+1)}(x)\right\}}{g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(x)\right\}}\right\}g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\right\}dc\right]\\ &\times K_h(X_i-x) \leq 0\end{aligned} \quad \text{(A.16)}$$

in probability. Let

$$L_{n1} = \frac{1}{n}\sum_{i=1}^n\varphi(Y_i, X_i)K_h(X_i-x),$$

where

$$\varphi(Y_i, X_i) = \int\log\left\{\frac{g\left\{c|Y_i, \boldsymbol{\theta}^{(l+1)}(x)\right\}}{g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(x)\right\}}\right\}g\left\{c|Y_i, \boldsymbol{\theta}^{(l)}(X_i)\right\}dc.$$

By conditions A and D, we have $g\{c|Y, \boldsymbol{\theta}^{(l)}(X)\} > a > 0$ for some small value a, and $\mathrm{E}\{\varphi(Y, X)^2\} < \infty$. Then by condition E and theorem A by Mack and Silverman (1982), we have

$$\sup_J|L_{n1} - f(x)\mathrm{E}\varphi(Y, x)| = o_P(1),$$

where J is a compact interval on which the density of X is bounded below from 0. The proof follows from

$$\begin{aligned}\mathrm{E}\varphi(Y, x) &= \mathrm{E}\left[\int\log\left\{\frac{g\left\{\mathcal{C}|Y, \boldsymbol{\theta}^{(l+1)}(x)\right\}}{g\left\{\mathcal{C}|Y, \boldsymbol{\theta}^{(l)}(x)\right\}}\right\}g\left\{c|Y, \boldsymbol{\theta}^{(l)}(x)\right\}dc\right]\\ &\leq \mathrm{E}\left(\log\left[\int\left\{\frac{g\left\{\mathcal{C}|Y, \boldsymbol{\theta}^{(l+1)}(x)\right\}}{g\left\{\mathcal{C}|Y, \boldsymbol{\theta}^{(l)}(x)\right\}}\right\}g\left\{c|Y, \boldsymbol{\theta}^{(l)}(x)\right\}dc\right]\right)=0.\end{aligned}$$

[Received March 2012. Revised November 2012.]

REFERENCES

Aerts, M., and Claeskens, G. (1997), "Local Polynomial Estimation in Multiparameter Likelihood Models," *Journal of the American Statistical Association*, 92, 1536–1545. [929]

Chen, H., Chen, J., and Kalbfleisch, J. (2001), "A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models," *Journal of the Royal Statistical Society*, Series B, 63, 19–29. [933]

Cheng, M., Fan, J., and Marron, J. (1997), "On Automatic Boundary Corrections," *The Annals of Statistics*, 25, 1691–1708. [938]

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [929]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications* (Vol. 66), London: Chapman & Hall/CRC. [929,930,940]

Fan, J., Heckman, N., and Wand, M. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, 90, 141–150. [929]

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193. [933]

Feng, Z., and McCulloch, C. (1996), "Using Bootstrap Likelihood Ratios in Finite Mixture Models," *Journal of the Royal Statistical Society*, Series B, 58, 609–617. [933]

Frühwirth-Schnatter, S. (2001), "Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models," *Journal of the American Statistical Association*, 96, 194–209. [929]

——— (2006), *Finite Mixture and Markov Switching Models*, New York: Springer Verlag. [929,930,933]

Goldfeld, S., and Quandt, R. (1973), "A Markov Model for Switching Regressions," *Journal of Econometrics*, 1, 3–15. [929,930]

Green, P., and Richardson, S. (2002), "Hidden Markov Models and Disease Mapping," *Journal of the American Statistical Association*, 97, 1055–1070. [929]

Hennig, C. (2000), "Identifiability of Models for Clusterwise Linear Regression," *Journal of Classification*, 17, 273–296. [930]

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. [930,936]

Hurn, M., Justel, A., and Robert, C. (2003), "Estimating Mixtures of Regressions," *Journal of Computational and Graphical Statistics*, 12, 55–79. [929]

Leroux, B. (1992), "Consistent Estimation of a Mixing Distribution," *The Annals of Statistics*, 20, 1350–1360. [933]

Li, P., and Chen, J. (2010), "Testing the Order of a Finite Mixture," *Journal of the American Statistical Association*, 105, 1084–1092. [933]

Lindsay, B. (1995), *Mixture Models: Theory, Geometry and Applications*, Hayward, CA: Institute of Mathematical Statistics. [929]

Mack, Y., and Silverman, B. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Probability Theory and Related Fields*, 61, 405–415. [940]

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models* (Vol. 299), New York: Wiley-Interscience. [933,939]

Rossi, P., Allenby, G., and McCulloch, R. (2005), *Bayesian Statistics and Marketing*, Chichester: Wiley. [929]

Stephens, M. (2000), "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society*, Series B, 62, 795–809. [929]

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567. [929]

Titterington, D., Smith, A., and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions* (Vol. 38), New York: Wiley. [930,938]

Wang, P., Puterman, M., Cockburn, I., and Le, N. (1996), "Mixed Poisson Regression Models With Covariate Dependent Rates," *Biometrics*, 52, 381–400. [929]

Wedel, M., and DeSarbo, W. (1993), "A Latent Class Binomial Logit Methodology for the Analysis of Paired Comparison Choice Data," *Decision Sciences*, 24, 1157–1170. [929]

Yao, W., and Lindsay, B. (2009), "Bayesian Mixture Labeling by Highest Posterior Density," *Journal of the American Statistical Association*, 104, 758–767. [929]