

Semiparametric Extensions of Mixture Models

WEIXIN YAO, *University of California, Riverside, weixin.yao@ucr.edu*

Abstract:

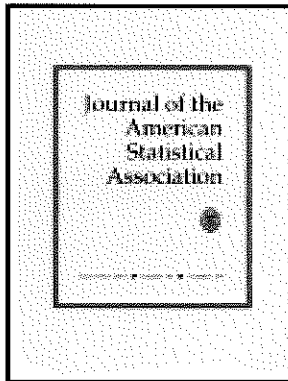
In this talk, several semiparametric extension of the traditional parametric finite mixture models are introduced. Mixture models are widely used when the population consists of several homogeneous subgroups. Currently most of mixture models considered are fully parametric. This talk will introduce three possible semiparametric extensions of traditional parametric mixture models and discuss their applications and estimation methods. The first extension considers the two component mixture model when one component is known and the other component is unknown. The second extension considers the mixture of linear regression model when the mixture proportions depend on predictors nonparametrically. The third extension considers a new class of mixture of single-index models, where the mixing proportions, mean functions, and variances are unknown but smooth functions of an index.

This article was downloaded by: [Kansas State University Libraries]

On: 12 November 2012, At: 08:10

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information:
<http://www.tandfonline.com/loi/uasa20>

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

Mian Huang^a & Weixin Yao^b

^a School of Statistics and Management, Shanghai University of Finance and Economics (SHUFE), Shanghai, 200433, P. R. China

^b Department of Statistics, , Kansas State University, Manhattan, Kansas, 66506

Accepted author version posted online: 14 May 2012. Version of record first published: 24 Jul 2012.

To cite this article: Mian Huang & Weixin Yao (2012): Mixture of Regression Models With Varying Mixing Proportions: A Semiparametric Approach, *Journal of the American Statistical Association*, 107:498, 711-724

To link to this article: <http://dx.doi.org/10.1080/01621459.2012.682541>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

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

Mian HUANG and Weixin YAO

In this article, 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 expectation-maximization-type (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.

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

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 C be a latent class variable with $P(C = c | \mathbf{x}) = \pi_c$ for $c = 1, 2, \dots, C$, where \mathbf{x} is a p -dimensional vector. Given $C = c$, suppose that the response y depends on \mathbf{x} in a linear way $y = \mathbf{x}^T \boldsymbol{\beta}_c + \epsilon_c$, where $\boldsymbol{\beta}_c = (\beta_{0c}, \beta_{1c}, \dots, \beta_{pc})^T$ and $\epsilon_c \sim N(0, \sigma_c^2)$. Then the conditional distribution of Y given \mathbf{x} can be written as

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

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

In this article, 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 \mathbf{x} or not. Consider the analysis of a CO₂-GDP (carbon dioxide-gross domestic product) dataset published by the World Resource Institute. As shown in Figure 3(a), the CO₂-GDP dataset contains two related variables of 171 countries in year 2005. The response variable is the CO₂ emission per capita in year 2005,

and the predictor is the GDP per capita in the same year, measured by the current U.S. 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 \mathbf{x} , which violates the constant proportion assumption of model (1.1).

The ideas that allow the proportions to depend on the covariates in a mixture model can be found in literature, for example, the hierarchical mixtures of experts model (Jordan and Jacobs 1994) in machine learning. Huang (2009) proposed a fully nonparametric mixture of regression models by assuming that 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 article, 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), 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 a kernel regression technique and a local likelihood method (Fan and Gijbels 1996). To achieve the optimal convergence rate for the global parameters $\boldsymbol{\beta}_c$'s and σ_c^2 's and the nonparametric

Mian Huang is Assistant Professor, School of Statistics and Management, Shanghai University of Finance and Economics (SHUFE), Shanghai, 200433, P. R. China (E-mail: huang.mian@shufe.edu.cn). Weixin Yao is the corresponding author and Assistant Professor, Department of Statistics, Kansas State University, Manhattan, Kansas 66506 (E-mail: wxyao@ksu.edu). Huang's research is partially supported by a National Science Foundation grant DMS 0348869, a funding through Projct 211 Phase 3 of SHUFE, and Shanghai Leading Academic Discipline Project, B803. The authors are grateful to the editor, the associate editor, and the referees for their insightful comments and suggestions, which greatly improved this article.

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 article makes the following major contributions to the literature:

- We show that mixture of regression models with varying mixing proportions are identifiable under certain conditions.
- 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 \sqrt{n} consistent estimates of β_c 's and σ_c^2 's have the same first-order asymptotic bias and variance as the kernel estimates with true values of β_c 's and σ_c^2 's.
- We develop a fast modified expectation-maximization (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 article 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 discussions. Technical conditions and proofs are given in Section 5.

2. ESTIMATION PROCEDURE AND ASYMPTOTIC PROPERTIES

2.1 The Semiparametric Mixture of Regressions

Suppose that $\{(\mathbf{X}_i, Y_i, Z_i), i = 1, \dots, n\}$ is a random sample from population (\mathbf{X}, Y, Z) . Throughout this article, X is p -dimensional and Y and Z are univariate. Let \mathcal{C} be a latent class variable, and assume that conditioning on \mathbf{x} , $Z = z$, \mathcal{C} has a discrete distribution $P(\mathcal{C} = c | \mathbf{x}, Z = z) = \pi_c(z)$ for $c = 1, 2, \dots, C - 1$. Here, Z can be a part of X . We assume that $\pi_c(z)$'s are smooth functions of z for $c = 1, 2, \dots, C$, and $\sum_{c=1}^C \pi_c(z) = 1$ for all z . Given $\mathcal{C} = c$, \mathbf{x} , and $Z = z$, Y follows a normal distribution with mean $\mathbf{x}^T \beta_c$ and variance σ_c^2 . In other words, conditioning on \mathbf{x} and $Z = z$, the response variable Y follows a finite mixture of normals

$$Y | \mathbf{x}, Z = z \sim \sum_{c=1}^C \pi_c(z) N(\mathbf{x}^T \beta_c, \sigma_c^2), \quad (2.1)$$

where $\mathbf{x} = (1, \mathbf{x}^T)^T$. 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 Titterton, Smith, and Makov (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} cannot be too small; see Hennig (2000) and Section 8.2.2 of Frühwirth-Schnatter (2006) for details. 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, $c = 1, \dots, C$, and (β_c, σ_c^2) , $c = 1, \dots, C$, are distinct pairs. In addition, assume that the domain \mathcal{X} of \mathbf{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 $\{(\mathbf{X}_i, Y_i, Z_i), i = 1, \dots, n\}$. That is,

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

where $\boldsymbol{\beta} = \{\beta_1^T, \dots, \beta_C^T\}^T$, $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \dots, \sigma_C^2\}^T$, and $\boldsymbol{\pi}(\cdot) = \{\pi_1(\cdot), \dots, \pi_{C-1}(\cdot)\}^T$. Since $\boldsymbol{\pi}(\cdot)$ consists of nonparametric functions, (2.2) is not yet ready for maximization. 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

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

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 $\tilde{\pi}_c(z) = \tilde{\pi}_c$, $\tilde{\beta}_c(z) = \tilde{\beta}_c$, 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 \sqrt{n} consistency. To improve the efficiency, the 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),

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

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$ be the solution of maximizing (2.4). Their \sqrt{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

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

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 $\hat{\sigma}^2$ 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 under-smoothing 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 $\hat{\sigma}^2$. We will show that the one-step estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are \sqrt{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

$$\begin{aligned} \rho(y|\mathbf{x}, \boldsymbol{\theta}) &= \sum_{c=1}^C \pi_c \phi(y|\mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2), \\ \ell(\boldsymbol{\theta}, \mathbf{x}, y) &= \log \rho(y|\mathbf{x}, \boldsymbol{\theta}), \\ q_{\boldsymbol{\theta}}\{\boldsymbol{\theta}, \mathbf{x}, y\} &= \frac{\partial \ell(\boldsymbol{\theta}, \mathbf{x}, y)}{\partial \boldsymbol{\theta}}, \\ q_{\boldsymbol{\theta}\boldsymbol{\theta}}\{\boldsymbol{\theta}, \mathbf{x}, y\} &= \frac{\partial^2 \ell(\boldsymbol{\theta}, \mathbf{x}, y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \end{aligned}$$

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

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

and

$$\Lambda(u|z) = \mathbb{E}[q_{\boldsymbol{\pi}}\{\boldsymbol{\theta}(z), \mathbf{X}, Y\} | Z = u],$$

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 $\boldsymbol{\psi}(\mathbf{x}, y, z)$ the vector that consists of the first $(C - 1)$ elements of $\mathcal{I}_{\boldsymbol{\theta}}^{-1}(z) \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}(z), \mathbf{x}, y)$.

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

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

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

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

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

Define

$$\kappa_i = \int u^i K(u) du \quad \text{and} \quad \nu_i = \int u^i K^2(u) du.$$

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

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

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

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

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

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

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

$c = 1, \dots, C.$

Let $\{u_1, \dots, u_N\}$ 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 $z \in \{u_j, j = 1, \dots, N\}$,

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

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

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

where $c = 1, \dots, C$, $w_{ic}^{(l+1)} = r_{ic}^{(l+1)} K_h(Z_i - z)$, $\mathbf{W}_c^{(l+1)} = \text{diag}\{w_{1c}^{(l+1)}, \dots, w_{nc}^{(l+1)}\}$, $\mathbf{y} = (Y_1, \dots, Y_n)^T$, and $\mathbf{S} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$. Furthermore, we update $\pi_c^{(l+1)}(Z_i)$, $\boldsymbol{\beta}_c^{(l+1)}(Z_i)$, and $\sigma_c^{2(l+1)}(Z_i)$, $i = 1, \dots, n$ by linearly interpolating $\pi_c^{(l+1)}(u_j)$, $\boldsymbol{\beta}_c^{(l+1)}(u_j)$, and $\sigma_c^{2(l+1)}(u_j)$, $j = 1, \dots, N$, respectively. In practice, if n is not very large, we may directly set the observed $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ to be the grid points. We also set grid points to be $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ when deriving the asymptotic ascent properties for the proposed algorithm.

In (2.7), for the 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, e.g., Celeux, Hurn, and Robert 2000; Stephens 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.

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

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

Then in the M-step, we update β_c 's and σ_c^2 's,

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

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

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

2.3.3 EM Algorithm for (2.5). Given $\hat{\beta}$ and $\hat{\sigma}$, we would maximize (2.5) to obtain the estimate $\hat{\pi}(z)$. Since $\hat{\beta}_c$ 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

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

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

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

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

2.3.4 A Computational Accelerating Scheme. To avoid extensive computation, many researchers prefer using a one-step estimate in semiparametric modeling, for example, a partially linear model (Hunsberger 1994; Severini and Staniswalis 1994), a generalized partially linear single-index model (Carroll et al. 1997), and a 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 avoided. Next, we discuss one approach to approximate the fully iterated estimation procedure with less computation.

In the E-step of l th cycle,

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

In the M-step, we simultaneously update β , σ , and $\pi(z)$ by

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

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

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

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

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 $nh \rightarrow \infty$ as $n \rightarrow \infty$ and $h \rightarrow 0$, we have

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

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), that is, $\ell_3(\pi^{(l+1)}(z)) \geq \ell_3(\pi^{(l)}(z))$, for all l , where $\ell_3(\cdot)$ is given in (2.5).
- (c) The iterations of (2.15)–(2.18) have the following property:

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

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 $\ell_1\{\theta(z)\}$ 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 $\{u_j, j = 1, \dots, N\}$ 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 $\pi(\cdot)$. In practice, data driven methods can be used for bandwidth selection, such as cross-validation (CV). Denote by \mathcal{D} the full dataset. We then partition \mathcal{D} into a training set \mathcal{R}_j and a test set \mathcal{T}_j , that is, $\mathcal{D} = \mathcal{T}_j \cup \mathcal{R}_j$ for $j = 1, \dots, J$. We use the training set \mathcal{R}_j to obtain the estimates $\{\hat{\pi}_c(\cdot), \hat{\sigma}_c^2, \hat{\beta}_c\}$. Then we can estimate $\pi_c(z)$ for the data points belonging to the corresponding test set. For $(\mathbf{x}_i, y_i, z_i) \in \mathcal{T}_j$,

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

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

$$CV = \sum_{j=1}^J \sum_{i \in \mathcal{T}_j} \log \left\{ \sum_{q=1}^C \hat{\pi}_q(z_i) \phi(y_i | \mathbf{x}_i^T \hat{\beta}_q, \hat{\sigma}_q^2) \right\}. \quad (3.1)$$

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 $nh^4 \rightarrow 0$ and $nh^2 \log(1/h) \rightarrow \infty$ to get the \sqrt{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:

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

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, Fan, and Li 2000) to construct the null distribution. Let $\{\bar{\pi}, \bar{\beta}, \bar{\sigma}^2\}$ be the maximum likelihood estimator (MLE) under null hypothesis. For given x_i , we can generate Y_i^* from the distribution $\sum_{c=1}^C \bar{\pi}_c N(\mathbf{x}_i^T \bar{\beta}_c, \bar{\sigma}_c^2)$. 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 null

parameters $\pi_c, c = 1, \dots, C - 1$, then the conditional bootstrap method is valid. Although a solid theoretical research is out of the scope in this article, we investigate the Wilk's phenomenon (Fan, Zhang, and Zhang 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 convenient way to conduct the likelihood ratio test for the above testing problem.

In addition, we use a bootstrap procedure to construct confidence intervals (CIs) for the parameters and pointwise CIs for the proportion functions. For given covariates, the response variable Y_i^* can be generated from the distribution $\sum_{c=1}^C \hat{\pi}_c(z_i) N(\mathbf{x}_i^T \hat{\beta}_c, \hat{\sigma}_c^2)$. We apply the proposed estimation procedure to each of the bootstrap samples, and further obtain the CIs. The bootstrap approach to construct CIs for nonparametric regression has been studied 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, p. 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 two-component mixture of regression model with varying mixing proportions:

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

where $m_1(\mathbf{x})$ and $m_2(\mathbf{x})$ are the regression functions for the first and second components, respectively. Therefore, in this example, $z = \mathbf{x}$, $\beta_1 = (4, -2)$, and $\beta_2 = (0, 3)$. The sample sizes $n = 200$ and 400 were conducted with 500 replicates. The predictor \mathbf{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} , and $2\hat{h}$, which correspond to the under-smoothing, appropriate smoothing, and over-smoothing, respectively. It was shown that the asymptotic distribution of the nonparametric 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" shows the results given

Table 1. The averages of MSEs of parameters and RASE $_{\pi}$ (the values are 100 times)

MSE	Bandwidth ($n = 200$)				Bandwidth ($n = 400$)			
	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	-

by the proposed method, while “M2” gives the results assuming that η were known.

Table 1 displays the mean square error (MSE) of regression parameter estimates and the average of RASE $_{\pi}$ over 500 simulations (the values are 100 times). 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 that 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, for example, RASE $_{\pi}$, 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 $\hat{\pi}(\cdot)$ works almost as well as if the true value of η were known and works better if it is not under-smoothing.

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

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

the true standard errors. To test the accuracy of 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 pointwise coverage probabilities for the proportion functions. For a set of grid points evenly distributed in the support of \mathbf{x} , Table 3 shows the results at the level of 95% for both “M1” and “M2.” For most 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 Tables 2 and 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 three 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 three densities were very close, plotted as 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 π . To validate the conditional bootstrap method, we select three typical samples generated from the three values of π_1 ’s. For each typical sample, we compute the conditional null distribution based on its 500 bootstrap samples. The resulting three 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 a sequence of local alternatives indexed by λ :

$$H_0 : \pi_1(\mathbf{x}) \equiv \pi_1 \quad \text{versus} \\ H_1 : \pi_1(\mathbf{x}) = 0.1 + 0.8\lambda \sin(\pi\mathbf{x})/\sqrt{nh},$$

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 sizes $n = 200, 400$. 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.

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 the World Resource Institute. We know that GDP is a measure of the size of a nation’s economy, and CO₂ is an important greenhouse gas that causes the greenhouse effect and may relate to global warming. Development with high GDP

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

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 10,000 and X by 10, so that they have comparable numerical scale.

For this dataset, we consider a two-component mixture of regression models with varying 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, respectively. 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 estimates of β along with its 95% 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, an increase 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 the United States, the United Kingdom, Canada, Australia, etc. The upper component has an estimated slope $\hat{\beta}_{21} = 1.021$. For countries within this component, an increase 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. Representative 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 CI are depicted in Figure 3(b).

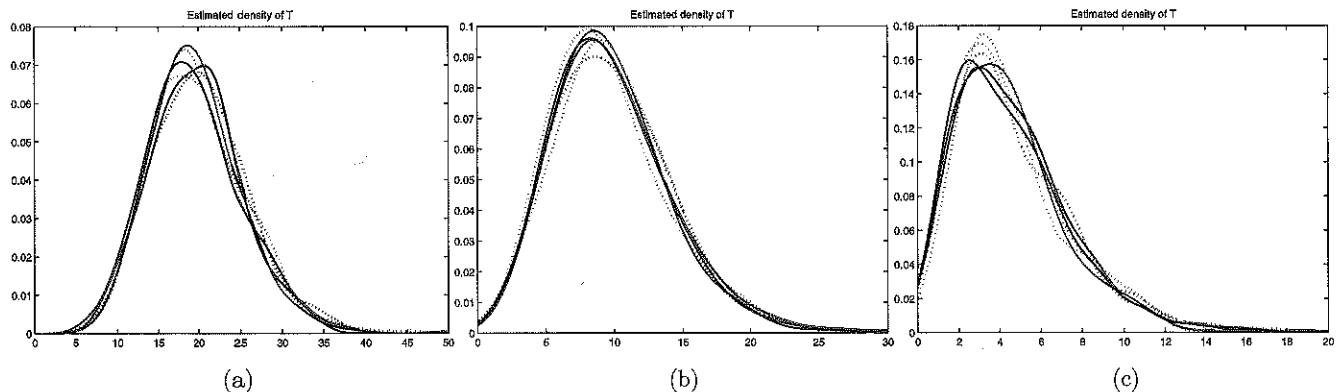


Figure 1. The estimated density of unconditional null distributions of T (solid lines), and the estimated density of conditional null distributions of T (dotted lines); the bandwidth is 0.04, 0.08, 0.16 in (a), (b), and (c), respectively. The online version of this figure is in color.

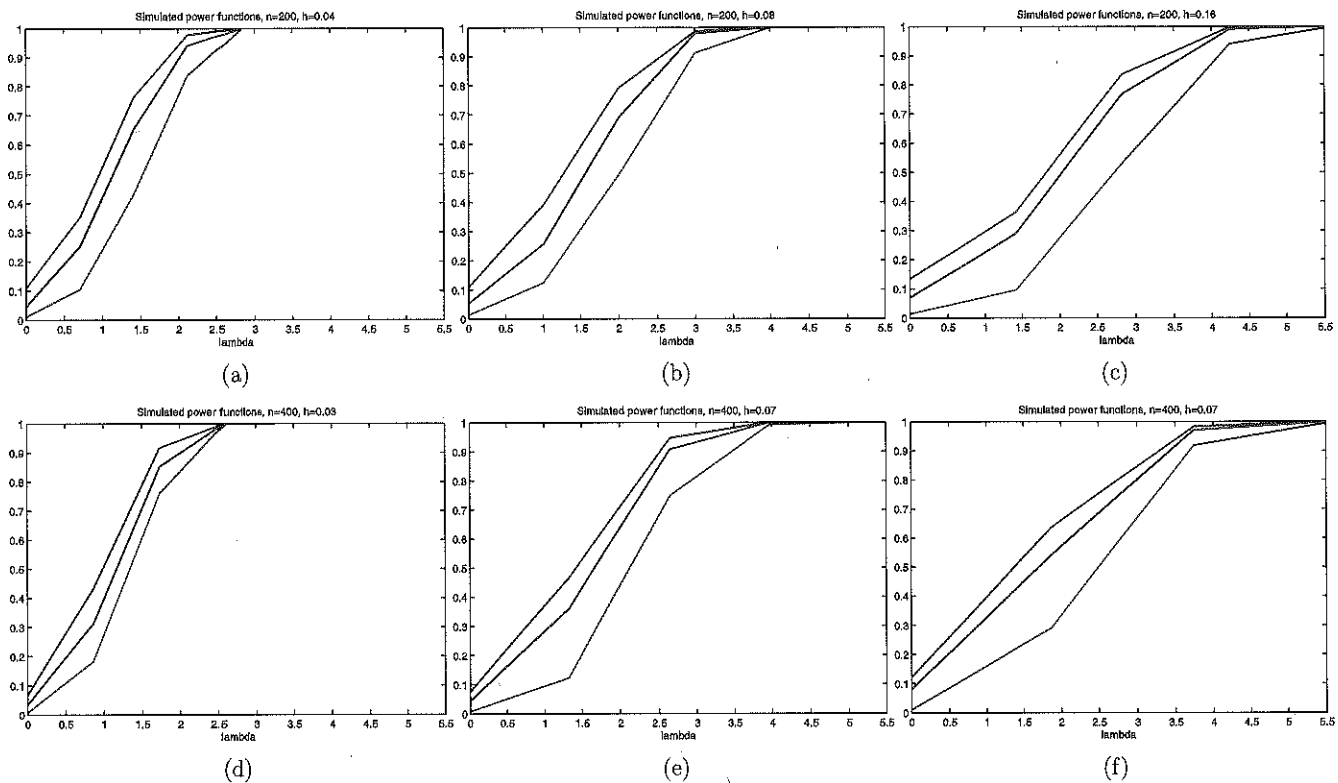


Figure 2. The power functions of the test against local alternatives; (a) $n = 200, h = 0.04$; (b) $n = 200, h = 0.08$; (c) $n = 200, h = 0.16$; (d) $n = 400, h = 0.03$; (e) $n = 400, h = 0.07$; (f) $n = 400, h = 0.14$. The online version of this figure is in color.

The result shows that as GDP per capita increases, the proportion of low-CO₂-emission countries 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 Akaike information criterion (AIC) and Bayesian information criterion (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, Chen, and Kalbfleisch (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

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

Table 4. Estimated parameters and confidence intervals

	Estimate	Bootstrap 95% CI	Estimate	Bootstrap 95% CI	Estimate	Bootstrap 95% 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)

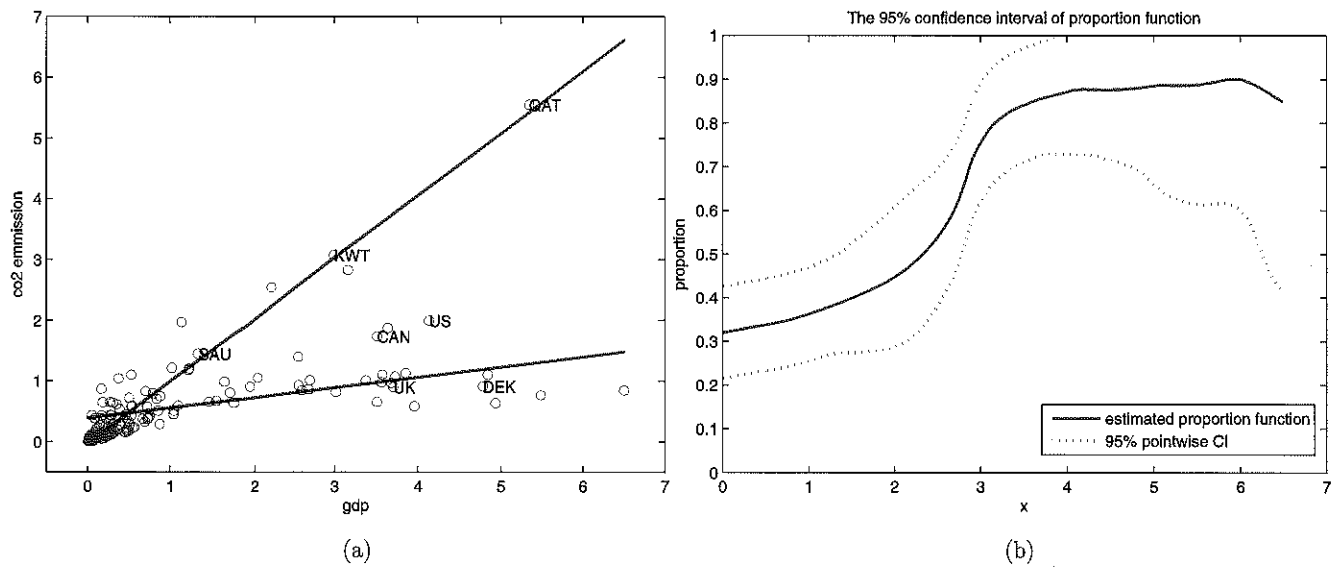


Figure 3. (a) The CO₂-GDP data, year 2005. y: CO₂ emission per capita; x: GDP per capita. (b) The estimated proportion function of the lower component and confidence interval. The online version of this figure is in color.

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

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

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

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

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 $(\boldsymbol{\gamma}_d, \delta_d^2), d = 1, \dots, D$.

From Lemma 1, for any given (x, z) such that both sets of parameters $(\mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2), c = 1, \dots, C$, and $(\mathbf{x}^T \boldsymbol{\gamma}_d, \delta_d^2), d = 1, \dots, D$, are distinct pairs, respectively, model (2.1) conditioning on $t = (x, z)$ is identifiable. Therefore, $C = D$ and there exists a permutation $\omega_t = \{\omega_t(1), \dots, \omega_t(C)\}$ of set $\{1, \dots, C\}$ depending on t , such that $\lambda_{\omega_t(c)}(z) = \pi_c(z), \mathbf{x}^T \boldsymbol{\gamma}_{\omega_t(c)} = \mathbf{x}^T \boldsymbol{\beta}_c, \delta_{\omega_t(c)}^2 = \sigma_c^2, c = 1, \dots, C$. Consider any permutation $\omega = \{\omega(1), \dots, \omega(C)\}$ such that

$$\mathbf{x}^T \boldsymbol{\gamma}_{\omega(c)} = \mathbf{x}^T \boldsymbol{\beta}_c, \delta_{\omega(c)}^2 = \sigma_c^2, \quad c = 1, \dots, C. \quad (5.1)$$

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 $\{1, 2, \dots, C\}$ and the domain \mathcal{X} of x contains an open set in \mathbb{R}^p , there must exist a permutation $\omega^* = \{\omega^*(1), \dots, \omega^*(C)\}$, such that (5.1) holds on a subset of \mathcal{X} with nonzero Lebesgue measure. Hence, $\boldsymbol{\gamma}_{\omega^*(c)} = \boldsymbol{\beta}_c, \delta_{\omega^*(c)}^2 = \sigma_c^2, c = 1, \dots, C$. Because that $(\boldsymbol{\beta}_c, \sigma_c^2), c = 1, \dots, C$ are distinct and $(\boldsymbol{\gamma}_c, \delta_c^2), c = 1, \dots, C$ 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 $\lambda_{\omega^*(c)}(z) = \pi_c(z), c = 1, \dots, C$ for any $z \in \mathcal{Z}$. If z is from $x, \lambda_{\omega^*(c)}(z) = \pi_c(z), c = 1, \dots, C$, 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 2–4. 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 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{(p+3)C-1})^T$ without changing the order of $\boldsymbol{\pi}, \boldsymbol{\sigma}^2$, and $\boldsymbol{\beta}$.

5.1 Regularity Conditions

- A. The sample $\{(\mathbf{X}_i, Y_i, Z_i), i = 1, \dots, n\}$ 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 $E\{M_{jkl}(X, Y, Z)\}$ is bounded for all j, k, l , and all X and Y .
- D. The unknown functions $\pi_c(z), c = 1, \dots, C - 1$, have continuous second derivative.
- E. The kernel density function $K(\cdot)$ is symmetric, continuous, and has a closed and bounded support.

- F. For $c = 1, \dots, C$, $\sigma_c^2 > 0$ and $\pi_c(z) > 0$ hold for all $z \in \mathcal{Z}$.
- G. The second derivative matrix $-\text{E}\{\partial^2 \ell(\theta(z), x, y) / \partial \theta \partial \theta^T \mid Z = z\}$ is positive definite, where $\theta(z) = (\pi^T(z), (\sigma^2)^T, \beta^T)^T$.
- H. $\text{E}(Z^{2r}) < \infty$ for some $\varepsilon < 1 - r^{-1}$, $n^{2\varepsilon-1}h \rightarrow \infty$.

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

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

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

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

Note that $\phi(y | \mathbf{x}^T \beta_c, \sigma_c^2)$ and $\phi(y | \mathbf{x}^T \beta_c, \sigma_c^2) (y - \mathbf{x}^T \beta_c)^k$ are bounded for any c and $k > 0$. Then we have

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

and

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

if \mathbf{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 $\{(\mathbf{X}_i, Y_i), i = 1, \dots, n\}$ be iid 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 $\text{E}|Y|^r < \infty$ 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

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

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

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

hood estimator of (2.3). Denote

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

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

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

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

where Δ_n is defined in (5.4), and

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

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

$$\begin{aligned} \ell_n^*(\theta^*) &= h \sum_{i=1}^n \{\ell(\theta(z) + \gamma_n \theta^*, X_i, Y_i) \\ &\quad - \ell(\theta(z), X_i, Y_i)\} K_h(Z_i - z), \end{aligned} \quad (5.2)$$

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

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

where

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

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

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

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

Since each element in Γ_n is sum of iid 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 that $\ell_n^*(\theta^*)$ is a concave function of θ^* for large n . Then by Condition (F), when n is large enough, $-\ell_n^*(\theta^*)$ is a convex function defined on a convex open set. Thus, by the convexity lemma (Pollard 1991),

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

holds uniformly for all $z \in \mathcal{Z}$ and θ^* 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 ; $\tilde{\theta}^*$ is a maximizer of (5.3). Then by lemma A.1 of Carroll et al. (1997), we have

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

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

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

By a expansion, we have

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

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

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

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

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

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

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

It follows that the order also holds for the right side of (5.12), that is,

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

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 $\hat{\eta}^* = \sqrt{n}(\hat{\eta} - \eta)$, where η is the true value. Further, define

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

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

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

By a Taylor expansion and some calculation,

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

where

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

For B_n , it can be shown that

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

Then by (5.14), we have

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

Next, we expand A_n as

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

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

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

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

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

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

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

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

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

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

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

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

By (5.15) and quadratic approximation lemma,

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

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

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

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

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

where $\psi(X, Y, Z)$ are the [1th, ..., (C - 1)th] elements of $\mathcal{I}_{\theta}^{-1}(Z) \times \{\partial \ell(\theta(Z), X, Y)/\partial \theta\}$. Further calculation shows that $E\{\omega(X, Y, Z)\} = 0$. So we have $E(A_n) = 0$. 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

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

where

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

It can be calculated that

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

where

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

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

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

Hence

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

where

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

We can show that

$$\text{var}(\Delta_n) = \mathcal{I}_{\pi}(z)f(z)\nu_0$$

and

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

where $\kappa_1 = \int u^1 K(u) du$, and $\nu_1 = \int u^1 K^2(u) du$. Then the result of Theorem 3 follows a standard argument.

Proof of Theorem 4. (a) We assume the unobserved data $(C_i, i = 1, \dots, n)$ are random samples from population \mathcal{C} , and the complete data $\{(X_i, Y_i, Z_i, C_i), i = 1, 2, \dots, n\}$ are random samples from (X, Y, Z, \mathcal{C}) . The conditional distribution of \mathcal{C} given X, Y , and θ is

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

For given $\theta^{(l)}(Z_i) = \{\pi^{(l)}(Z_i), \beta^{(l)}(Z_i), \sigma^{2(l)}(Z_i)\}$, we have $g\{c|X_i, Y_i, \theta^{(l)}(Z_i)\} = r_{ic}^{(l+1)}$, and $\sum_{c=1}^C r_{ic}^{(l+1)} = 1, i = 1, \dots, n$. Then

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

By (5.18), we also have

$$\begin{aligned} \log \left\{ \sum_{c=1}^C \pi_c \phi(Y_i|\mathbf{x}_i^T \beta_c, \sigma_c^2) \right\} \\ = \log \{ \pi_c \phi(Y_i|\mathbf{x}_i^T \beta_c, \sigma_c^2) \} - \log[g\{c|X_i, Y_i, \theta\}]. \end{aligned} \quad (5.20)$$

Thus, we have

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

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

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

It suffices to show that

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

in probability. Define

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

and

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

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

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

By a standard argument, we know

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

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

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

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

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

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)

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

Based on the Jensen's inequality, we have

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

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

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

Therefore, we only need to show

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

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

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

in probability. Hence,

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

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

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

where

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

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

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

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

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

This completes the proof of Theorem 4(c).

[Received March 2011. Revised December 2011.]

REFERENCES

Cai, Z., Fan, J., and Li, R. (2000), "Efficient Estimation and Inferences for Varying-Coefficient Models," *Journal of the American Statistical Association*, 95, 888-902. [715]
 Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477-489. [714]

- 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. [713]
- 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. [718]
- Chen, J., and Li, P. (2009), "Hypothesis Test for Normal Mixture Models: The EM Approach," *The Annals of Statistics*, 37, 2523–2542. [718]
- Claeskens, G., and Van Keilegom, I. (2003), "Bootstrap Confidence Bands for Regression Curves and Their Derivatives," *The Annals of Statistics*, 31(6), 1852–1884. [715]
- Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, New York: Cambridge University Press. [715]
- Eubank, R. L., and Speckman, P. L. (1993), "Confidence Bands in Nonparametric Regression," *Journal of the American Statistical Association*, 88, 1287–1301. [715]
- Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall. [711]
- Fan, J., and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli*, 11, 1031–1057. [720]
- Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193. [715]
- 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. [711]
- (2006), *Finite Mixture and Markov Switching Models*, New York: Springer. [711]
- Goldfeld, S. M., and Quandt, R. E. (1973), "A Markov Model for Switching Regression," *Journal of Econometrics*, 1, 3–15. [711]
- Green, P. J., and Richardson, S. (2002), "Hidden Markov Models and Disease Mapping," *Journal of the American Statistical Association*, 97, 1055–1070. [711]
- 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. [715]
- Härdle, W., and Marron, J. S. (1991), "Bootstrap Simultaneous Error Bars for Nonparametric Regression," *The Annals of Statistics*, 19(2), 778–796. [715]
- Hathaway, R. J. (1985), "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions," *The Annals of Statistics*, 13, 795–800. [712]
- Hennig, C. (2000), "Identifiability of Models for Clusterwise Linear Regression," *Journal of Classification*, 17, 273–296. [712]
- Huang, M. (2009), "Nonparametric Techniques in Mixture of Regression Models," Ph.D. Dissertation, The Pennsylvania State University. [711]
- Hunsberger, S. (1994), "Semiparametric Regression in Likelihood-Based Models," *Journal of the American Statistical Association*, 89, 1354–1365. [714]
- Hurn, M., Justel, A., and Robert, C. (2003), "Estimating Mixture of Regressions," *Journal of Computational and Graphical Statistics*, 12, 55–79. [711]
- Jordan, M. I., and Jacobs, R. A. (1994), "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Computation*, 6, 181–214. [711]
- Leroux, B. G. (1992), "Consistent Estimation of a Mixing Distribution," *The Annals of Statistics*, 20, 1350–1360. [718]
- Li, R., and Liang, H. (2008), "Variable Selection in Semiparametric Modeling," *The Annals of Statistics*, 36, 261–286. [714]
- McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley. [711]
- Neumann, M. H., and Polzehl, J. (1998), "Simultaneous Bootstrap Confidence Bands in Nonparametric Regression," *Journal of Nonparametric Statistics*, 9, 307–333. [715]
- Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199. [720]
- Severini, T. A., and Staniswalis, J. G. (1994), "Quasilikelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511. [714]
- Stephens, M. (2000), "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society, Series B*, 62, 795–809. [713]
- Titterton, D., Smith, A., and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley. [712]
- Wang, P., Puterman, M. L., Cockburn, I., and Le, N. (1996), "Mixed Poisson Regression Models With Covariate Dependent Rates," *Biometrics*, 52, 381–400. [711]
- 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. [711]
- Xia, Y. C. (1998), "Bias-Corrected Confidence Bands in Nonparametric Regression," *Journal of the American Statistical Association*, 60, 797–811. [715]
- Yao, W., and Lindsay, B. G. (2009), "Bayesian Mixture Labeling by Highest Posterior Density," *Journal of the American Statistical Association*, 104, 758–767. [713]
- Young, D. S., and Hunter, D. R. (2010), "Mixtures of Regressions With Predictor-Dependent Mixing Proportions," *Computational Statistics and Data Analysis*, 54, 2253–2266. [711]

Minimum profile Hellinger distance estimation for a semiparametric mixture model

Sijia XIANG¹, Weixin YAO^{2*} and Jingjing WU³

¹*School of Mathematics and Statistics, Zhejiang University of Finance and Economics, Hangzhou, Zhejiang 310018, P.R. China*

²*Department of Statistics, Kansas State University, Manhattan, KS 66506, U.S.A.*

³*Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4*

Key words and phrases: Minimum profile Hellinger distance; semiparametric EM algorithm; semiparametric mixture models.

MSC 2010: Primary 62G05; secondary 62G35

Abstract: In this paper, we propose a new effective estimator for a class of semiparametric mixture models where one component has known distribution with possibly unknown parameters while the other component density and the mixing proportion are unknown. Such semiparametric mixture models have been often used in multiple hypothesis testing and the sequential clustering algorithm. The proposed estimator is based on the minimum profile Hellinger distance (MPHD), and its theoretical properties are investigated. In addition, we use simulation studies to illustrate the finite sample performance of the MPHD estimator and compare it with some other existing approaches. The empirical studies demonstrate that the new method outperforms existing estimators when data are generated under contamination and works comparably to existing estimators when data are not contaminated. Applications to two real data sets are also provided to illustrate the effectiveness of the new methodology. *The Canadian Journal of Statistics* xx: 1–22; 2014 © 2014 Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–22; 2014 © 2014 Société statistique du Canada

1. INTRODUCTION

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

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

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

Bordes, Delmas, & Vandekerkhove (2006) studied a special case when ξ is assumed to be known, that is, the first component density is completely known and model (1) becomes

$$h(x) = \pi f_0(x) + (1 - \pi)f(x - \mu), \quad x \in \mathbb{R}. \quad (2)$$

Model (2) is motivated by multiple hypothesis testing to detect differentially expressed genes under two or more conditions in microarray data. For this purpose, we build a test statistic for each gene. The test statistics can be considered as coming from a mixture of two distributions: the

* Author to whom correspondence may be addressed.
E-mail: wxyao@ksu.edu

known distribution f_0 under null hypothesis, and the other distribution $f(\cdot - \mu)$, the unknown distribution of the test statistics under the alternative hypothesis. Please see Section 4 for such an application on multiple hypothesis testing.

Song, Nicolae, & Song (2010) studied another special case of model (1),

$$h(x) = \pi\phi_\sigma(x) + (1 - \pi)f(x), \quad x \in \mathbb{R}, \quad (3)$$

where ϕ_σ is a normal density with mean 0 and unknown standard deviation σ and $f(x)$ is an unknown density. Model (3) was motivated by a sequential clustering algorithm (Song & Nicolae, 2009), which works by finding a local centre of a cluster first, and then identifying whether an object belongs to that cluster or not. If we assume that the objects belonging to the cluster come from a normal distribution with known mean (such as zero) and unknown variance σ^2 and that the objects not belonging to the cluster come from an unknown distribution f , then identifying the points in the cluster is equivalent to estimating the mixing proportion in model (3).

Bordes, Delmas, & Vandekerkhove (2006) proposed to estimate model (2) based on symmetrization of the unknown distribution f and proved the consistency of their estimator. However, the asymptotic distribution of their estimator has not been provided. Song, Nicolae, & Song (2010) also proposed an EM-type estimator and a maximizing π -type estimator (inspired by the constraints imposed to achieve identifiability of the parameters and Swanepoel's approach (Swanepoel, 1999)) to estimate model (3) without providing any asymptotic properties.

In this article, we propose a new estimation procedure for the unified model (1) based on minimum profile Hellinger distance (MPHD) (Wu, Schick, & Karunamuni, 2011). We will investigate the theoretical properties of the proposed MPHD estimator for the semiparametric mixture model, such as existence, consistency and asymptotic normality. A simple and effective algorithm is also given to compute the proposed estimator. Using simulation studies, we illustrate the effectiveness of the MPHD estimator and compare it with the estimators suggested by Bordes, Delmas, & Vandekerkhove (2006) and Song, Nicolae, & Song (2010). Compared to the existing methods (Bordes, Delmas, & Vandekerkhove, 2006; Song, Nicolae, & Song, 2010), the new method can be applied to the more general model (1). In addition, the MPHD estimator works competitively under semiparametric model assumptions, while it is more robust than the existing methods when data are contaminated.

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

There are other well-known robust approaches within the mixture model-based clustering literature. García-Escudero, Gordaliza, & Matrán (2003) proposed exploratory graphical tools based on trimming for detecting main clusters in a given dataset, where the trimming is obtained by resorting to trimmed k -means methodology. García-Escudero et al. (2008) introduced a new method for performing clustering with the aim of fitting clusters with different scatters and weights. García-Escudero et al. (2010) reviewed different robust clustering approaches in the literature, emphasizing on methods based on trimming which try to discard most outlying data when carrying out the clustering process. A more recent work by Punzo & McNicholas (2013) introduced a family of 14 parsimonious mixtures of contaminated Gaussian distributions models within the general model-based classification framework.

The rest of the article is organized as follows. In Section 2, we introduce the proposed MPHD estimator and discuss its asymptotic properties. Section 3 presents simulation results for comparing the new estimation with some existing methods. Applications to two real data sets are also provided in Section 4 to illustrate the effectiveness of the proposed methodology. A discussion section ends the paper.

2. MPHD ESTIMATION

2.1. Introduction of MPHD Estimator

In this section, we develop a MPHD estimator for model (1). Let

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

where

$$\Theta = \{\theta = (\pi, \xi, \mu) : \pi \in (0, 1), \xi \in \mathbb{R}, \mu \in \mathbb{R}\},$$

$$\mathcal{F} = \{f : f \geq 0, \int f(x)dx = 1\}$$

be the functional space for the semiparametric model (1). In practice, the parameter space of ξ depends on its interpretation. For example, if ξ is the standard deviation of f_0 , then the parameter space of ξ will be \mathbb{R}^+ . For model (2), ξ is known and thus the parameter space of ξ is a singleton and, as a result, $\theta = (\pi, \mu)$.

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

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

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

$$\left\| h_{t, l}^{1/2} - \hat{h}_n^{1/2} \right\| \quad (4)$$

over all $t \in \Theta$ and $l \in \mathcal{F}$, where \hat{h}_n is an appropriate nonparametric density estimator of $h_{\theta, f}$. Note that the above objective function (4) contains both the parametric component t and the nonparametric component l . Here, we propose to use the profile idea to implement the calculation.

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

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

and then define the profile Hellinger distance as

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

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

$$T(g) = \arg \min_{t \in \Theta} d_{PH}(t, g) = \arg \min_{t \in \Theta} \left\| h_{t, f(t, g)}^{1/2} - g^{1/2} \right\|. \quad (5)$$

Given the sample X_1, X_2, \dots, X_n , one can construct an appropriate nonparametric density estimator of $h_{\theta, f}$, say \hat{h}_n , and then the proposed MPHD estimator of θ is given by $T(\hat{h}_n)$. In the

examples of Sections 3 and 4, we use the kernel density estimator for \hat{h}_n and the bandwidth h is chosen based on Botev, Grotowski, & Kroese (2010).

2.2. Algorithm

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

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

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

Similar to Wu, Schick, & Karunamuni (2011), we obtain that

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

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

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

$$\left\| [\pi^{(k+1)} f_0(\cdot; \xi^{(k+1)}) + (1 - \pi^{(k+1)}) f^{(k+1)}(\cdot - \mu^{(k+1)})]^{1/2} - \hat{h}_n^{1/2}(\cdot) \right\|. \quad (6)$$

Then go back to **Step 1**.

Each of the above two steps monotonically decreases the objective function (4) until convergence. In Step 1, if $f(\cdot)$ is assumed to be symmetric, then we can further symmetrize $f^{(k+1)}(\cdot)$ as

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

Note that there is no closed form for (6) in Step 2 and thus some numerical algorithms, such as the Newton–Raphson algorithm, are needed to minimize (6). In our examples, we used the “fminsearch” function in Matlab to find the minimizer numerically. “fminsearch” function uses the Nelder–Mead simplex algorithm as described in Lagarias et al. (1998).

2.3. Asymptotic Results

Note that θ and f in the semiparametric mixture model (1) are not generally identifiable without any assumptions for f . Bordes, Delmas, & Vandekerkhove (2006) showed that model (2) is not generally identifiable if we do not put any restrictions on the unknown density f , but identifiability can be achieved under some sufficient conditions. One of these conditions is that $f(\cdot)$ is symmetric about 0. Under these conditions, Bordes, Delmas, & Vandekerkhove (2006) proposed an elegant estimation procedure based on the symmetry of f . Song, Nicolae, & Song (2010) also addressed the non-identifiability problem and noticed that model (3) is not generally identifiable. However, due to the additional unknown parameter σ in the first component, Song, Nicolae, & Song (2010) mentioned that it is hard to find the conditions to avoid unidentifiability of model (3) and proposed using simulation studies to check the performance of the proposed estimators.

Please refer to Bordes, Delmas, & Vandekerkhove (2006) and Song, Nicolae, & Song (2010) for detailed discussions on the identifiability of model (1).

Next, we discuss some asymptotic properties of the proposed MPHD estimator. Here, for simplicity of explanation, we will only consider model (2) for which Bordes, Delmas, & Vandekerkhove (2006) has proved identifiability. However, we conjecture that all the results presented in this section also apply to the unified model (1) when it is identifiable. But this is beyond the scope of the article and requires more research to find the identifiable conditions for the general model (1).

The next theorem gives results on the existence and uniqueness of the proposed estimator, and the continuity of the functional defined in (5), which is in line with Theorem 1 of Beran (1977).

Theorem 1. *With T defined by (5), if model (2) is identifiable, then we have*

1. *For every $h_{\theta, f} \in \mathcal{H}$, there exists $T(h_{\theta, f}) \in \Theta$ satisfying (5);*
2. *$T(h_{\theta, f}) = \theta$ uniquely for any $\theta \in \Theta$;*
3. *$T(h_n) \rightarrow T(h_{\theta, f})$ for any sequences $\{h_n\}_{n \in \mathbb{N}}$ such that $\|h_n^{1/2} - h_{\theta, f}^{1/2}\| \rightarrow 0$ and $\sup_{t \in \Theta} \|h_{t, f(t, h_n)} - h_{t, f(t, h_{\theta, f})}\| \rightarrow 0$ as $n \rightarrow \infty$.*

Remark 1. *Without the global identifiability of model (2), the local identifiability of model (2) proved by Bordes, Delmas, & Vandekerkhove (2006) tells that there exists one solution that has the asymptotic properties presented in Theorem 1.*

Define a kernel density estimator based on X_1, X_2, \dots, X_n as

$$\hat{h}_n(x) = \frac{1}{nc_n s_n} \sum_{i=1}^n K\left(\frac{x - X_i}{c_n s_n}\right), \quad (7)$$

where $\{c_n\}$ is a sequence of constants (bandwidths) converging to zero at an appropriate rate and s_n is a robust scale statistic. Under further conditions on the kernel density estimator defined in (7), the consistency of the MPHD estimator is established in the next theorem.

Theorem 2. *Suppose that*

1. *The kernel function $K(\cdot)$ is absolutely continuous and bounded with compact support.*
2. *$\lim_{n \rightarrow \infty} c_n = 0$, $\lim_{n \rightarrow \infty} n^{1/2} c_n = \infty$.*
3. *The model (2) is identifiable and $h_{\theta, f}$ is uniformly continuous.*

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

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

$$s_{\theta + \alpha e, g}(x) = s_{\theta, g}(x) + \alpha e^T \dot{s}_{\theta, g}(x) + \alpha e^T u_{\alpha, g}(x), \quad (8)$$

$$\dot{s}_{\theta + \alpha e, g}(x) = \dot{s}_{\theta, g}(x) + \alpha \ddot{s}_{\theta, g}(x)e + \alpha v_{\alpha, g}(x)e, \quad (9)$$

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

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

Theorem 3. *Suppose that*

1. *Model (2) is identifiable.*
2. *The conditions in Theorem 2 hold.*
3. *The map $\theta \mapsto s_{\theta, g}$ satisfies (9) and (9) with continuous gradient vector $\dot{s}_{\theta, g}$ and continuous Hessian matrix $\ddot{s}_{\theta, g}$ in the sense that $\|\dot{s}_{\theta_n, g_n} - \dot{s}_{\theta, g}\| \rightarrow 0$ and $\|\ddot{s}_{\theta_n, g_n} - \ddot{s}_{\theta, g}\| \rightarrow 0$ whenever $\theta_n \rightarrow \theta$ and $\|g_n^{1/2} - g^{1/2}\| \rightarrow 0$ as $n \rightarrow \infty$.*
4. *$\langle \bar{s}_{\theta, h_{\theta, f}}, h_{\theta, f}^{1/2} \rangle$ is invertible.*

Then, with T defined in (5) for model (2), the asymptotic distribution of $n^{1/2}(T(\hat{h}_n) - T(h_{\theta, f}))$ is $N(0, \Sigma)$ with variance matrix Σ defined by

$$\Sigma = \langle \bar{s}_{\theta, h_{\theta, f}}, h_{\theta, f}^{1/2} \rangle^{-1} \langle \dot{s}_{\theta, h_{\theta, f}}, \dot{s}_{\theta, h_{\theta, f}}^T \rangle \langle \bar{s}_{\theta, h_{\theta, f}}, h_{\theta, f}^{1/2} \rangle^{-1}.$$

3. SIMULATION STUDIES

In this section, we investigate the finite sample performance of the proposed MPHD estimator and compare it to Maximizing- π type estimator (Song, Nicolae, & Song, 2010), EM-type estimator (Song, Nicolae, & Song, 2010) and Symmetrization estimator (Bordes, Delmas, & Vandekerkhove, 2006) under both models (2) and (3).

Model (3) that Song, Nicolae, & Song (2010) considered does not have a location parameter in the second component. However, we can equivalently replace $f(x)$ with $f(x - \mu)$, where $\mu \in \mathbb{R}$ is a location parameter. Throughout this section, we will consider this equivalent form of (3). Under this model, after we have $\hat{\pi}$ and $\hat{\sigma}$, we can simply estimate μ by

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

where

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

We first compare the performance of different estimators under model (2). Suppose (X_1, \dots, X_n) are generated from one of the following five cases:

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

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

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

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

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

Figure 1 shows the density plots of the five cases. Cases I, II and III are the models used by Song, Nicolae, & Song (2010) to show the performance of their Maximizing- π type and EM-type estimators. Case I represents the situation when two components are close, and Case II represents the situation when two components are apart. Cases IV and V are suggested by Bordes, Delmas, & Vandekerkhove (2006) to show the performance of their semiparametric EM algorithm. In

Accepted Proof

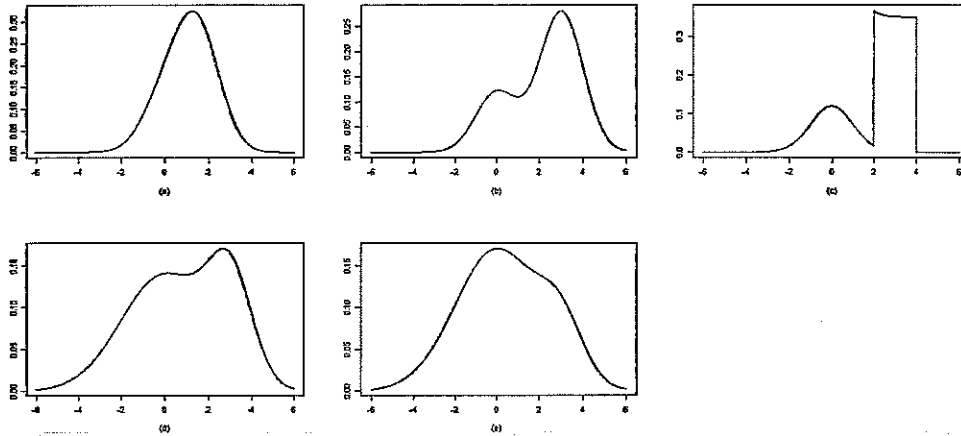


FIGURE 1: Density plots of: (a) Case I; (b) Case II; (c) Case III; (d) Case IV and (e) Case V.

addition, we also consider the corresponding contaminated models by adding 2% outliers from $U(10, 20)$ to the above five models.

Tables 1, 2 and 3 report the bias and MSE of the parameter estimates of (π, μ) for the four methods when $n = 100, n = 250$ and $n = 1,000$, respectively, based on 200 repetitions. Tables 4, 5 and 6 report the respective results for $n = 100, n = 250$ and $n = 1,000$ when the data are under 2% contamination from $U(10, 20)$. The best values are highlighted in bold. From the six tables, we can see that the MPHD estimator has better overall performance than the Maximizing- π type, the EM-type and the Symmetrization estimators, especially when sample size is large. When the sample is not contaminated by outliers, the MPHD estimator and the Symmetrization estimator are very competitive and perform better than other estimators. When the sample is contaminated by outliers, the MPHD estimator performs much better and therefore is more robust than the other three methods. We also observe that when the sample is contaminated by outliers, among the Maximizing- π type, the EM-type and the Symmetrization estimators, the EM-type estimator tends to give better mixing proportion estimates than the other two.

TABLE 1: Bias (MSE) of point estimates for model (2) over 200 repetitions with $n = 100$.

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

TABLE 2: Bias (MSE) of point estimates for model (2) over 200 repetitions with $n = 250$.

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

Next, we also evaluate how the MPHD estimator performs under model (3), where the variance σ^2 is assumed to be unknown, and compare it with other methods using the same five cases as in Tables 1–6.

Tables 7, 8 and 9 report the bias and MSE of the parameter estimates for $n = 100$, $n = 250$ and $n = 1,000$, respectively, when there are no contaminations. Based on these three tables, we can see that when there are no contaminations, the MPHD estimator and the Symmetrization estimator perform better than the Maximizing- π type estimator and the EM-type estimator. Tables 10, 11 and 12 report the results when models are under 2% contamination from $U(10, 20)$ for $n = 100$, $n = 250$ and $n = 1,000$, respectively. From these three tables, we can see that the MPHD estimator performs much better again than the other three methods.

TABLE 3: Bias (MSE) of point estimates for model (2) over 200 repetitions with $n = 1,000$.

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

TABLE 4: Bias (MSE) of point estimates for model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 100$.

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

To see the comparison and difference better, we also plot in Figures 2–4 the results reported in Tables 6 and 9. Figure 2 contains the MSE of point estimates of μ that are presented in Table 9 for model (3) (σ unknown) and Figures 3 and 4 contain the MSEs of point estimates of μ and π , respectively, that are presented in Table 6 for model (2) (σ known), under 2% contamination from $U(10, 20)$. From the three plots, we can see that all four estimators perform well in Cases II and III. The EM-type estimator performs poorly in Case I, and is the worst estimate of μ in Cases IV and V when data are contaminated. The Symmetrization estimator is sensitive to contamination, especially in Cases IV and V, no matter σ is known or not. Comparatively, the Maximizing- π type estimator is more robust, but it does not perform well in Cases IV and V when data are not under contamination. However, the MPHD estimator performs well in all cases.

TABLE 5: Bias (MSE) of point estimates for model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 250$.

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

TABLE 6: Bias (MSE) of point estimates for model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1,000$.

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

4. REAL DATA APPLICATION

Example 1 (Iris data). We illustrate the application of the new estimation procedure to the sequential clustering algorithm using the Iris data, which are perhaps one of the best known data sets in pattern recognition literature. Iris data were first introduced by Fisher (1936) and are referenced

TABLE 7: Bias (MSE) of point estimates for model (3) over 200 repetitions with $n = 100$.

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

TABLE 8: Bias (MSE) of point estimates for model (3) over 200 repetitions with $n = 250$.

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

TABLE 9: Bias (MSE) of point estimates for model (3) over 200 repetitions with $n = 1,000$.

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

TABLE 10: Bias (MSE) of point estimates for model (3), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 100$.

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

frequently to this day. These data contain four attributes: sepal length (in cm), sepal width (in cm), petal length (in cm) and petal width (in cm), and there are three classes of 50 instances each, where each class refers to a type of Iris plant. One class is linearly separable from the other two and the latter are not linearly separable from each other.

Assuming the class indicators are unknown, we want to recover the three clusters in the data. After applying the search algorithm for centres of clusters by Song, Nicolae, & Song (2010), observation 8 is selected as the centre of the first cluster. We adjust all observations by subtracting observation 8 from each observation. As discussed by Song, Nicolae, & Song (2010), the proportion of observations that belong to a cluster can be considered as the mixing proportion in the two-component semiparametric mixture model (3).

Principal component analysis shows that the first principal component accounts for 92.46% of the total variability, so it would seem that the Iris data tend to fall within a 1-dimensional subspace of the 4-dimensional sample space. Figure 5 is a histogram of the first principal component. From the histogram, we can see that the first cluster is separated from the rest of the data, with observation 8 (first principal component score equals -2.63) being the centre of it. The first principal component loading vector is $(0.36, -0.08, 0.86, 0.35)$, which implies that the petal length contains most of the information. We apply each of the four estimation methods discussed above to the first principal component. Note, however, that the leading principal components are not necessary to have better clustering information than other components. Some cautions are needed when using principal components in clustering applications.

Similar to Song, Nicolae, & Song (2010), in Table 13, we report the estimates of proportion based on the first principal component. Noting that the true proportion is $1/3$, we can see that the MPHD and the Symmetrization estimators perform better than the other two estimators.

TABLE 11: Bias (MSE) of point estimates for model (3), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 250$.

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

Example 2 (Breast cancer data). Next, we illustrate the application of the new estimation procedure to multiple hypothesis testing using the breast cancer data from Hedenfalk et al. (2001), who examined gene expressions in breast cancer tissues from women who were carriers of the hereditary BRCA1 or BRCA2 gene mutations, predisposing to breast cancer. The breast cancer data were downloaded from "http://research.nhgri.nih.gov/microarray/NEJM_Supplement/" and contains gene expression ratios derived from the fluorescent intensity (proportional to the gene expression level) from a tumour sample divided by the fluorescent intensity from a common reference sample (MCF-10A cell line). The ratios were normalized (or calibrated) such that the majority of the gene expression ratios from a pre-selected internal control gene set was around 1.0, but no log-transformation was used. The data set consists of 3,226 genes on $n_1 = 7$ BRCA1 arrays and $n_2 = 8$ BRCA2 arrays. If any gene had one or more measurement exceeding 20, then this gene was eliminated (Storey & Tibshirani, 2003). This left 3,170 genes. The p -values were calculated based on permutation tests (Storey & Tibshirani, 2003). We then transform the p -values via the probit transformation to z -score, given by $z_i = \Phi^{-1}(1 - p_i)$ (McLachlan & Wockner, 2010). Figure 6 displays the fitted densities, and Table 14 lists the parameter estimates of the four methods discussed in the article. MPHD estimator shows that among the 3170 genes examined, around 29% genes are differentially expressed between those tumour types, which is close to the 33% from Storey & Tibshirani (2003) and 32.5% from Langaas, Lindqvist, & Ferkingstad (2005).

Let

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

TABLE 12: Bias (MSE) of point estimates for model (3), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1,000$.

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

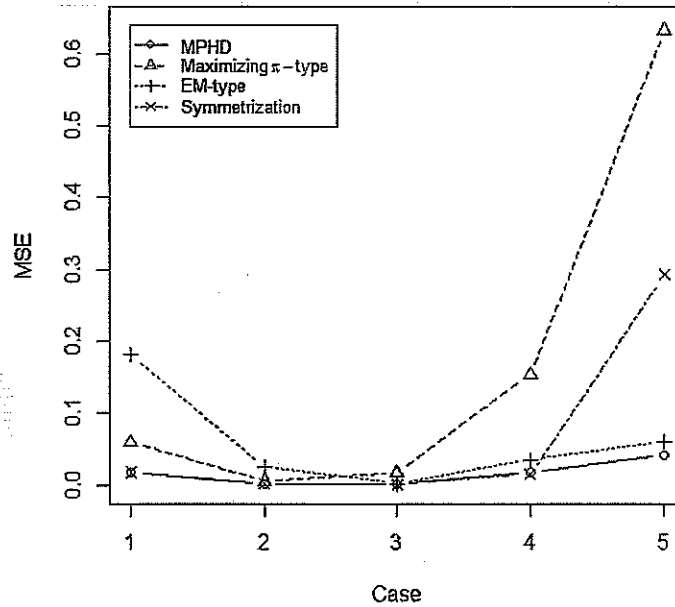


FIGURE 2: MSE of point estimates of μ of model (3), over 200 repetitions with $n = 1,000$.

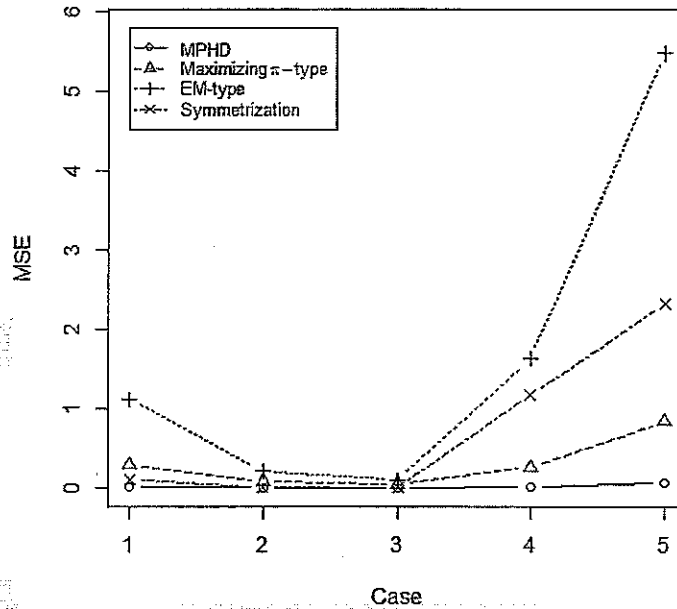


FIGURE 3: MSE of point estimates of μ of model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1,000$.

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

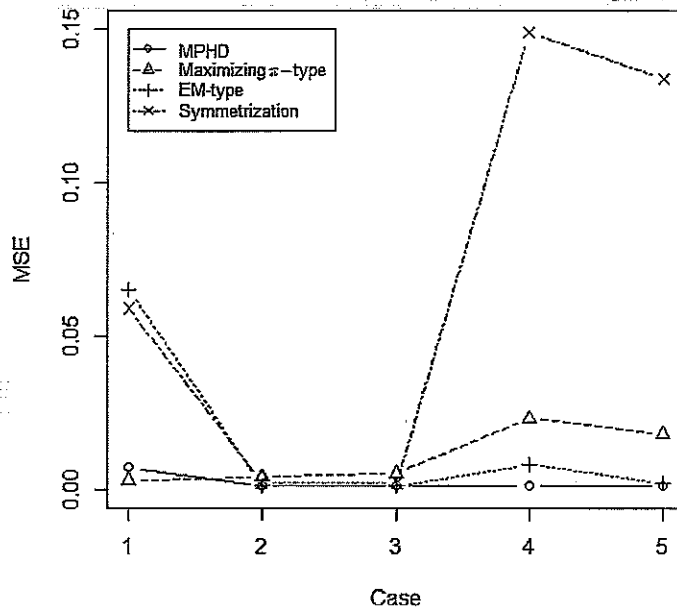


FIGURE 4: MSE of point estimates of π of model (2), under 2% contamination from $U(10, 20)$, over 200 repetitions with $n = 1,000$.

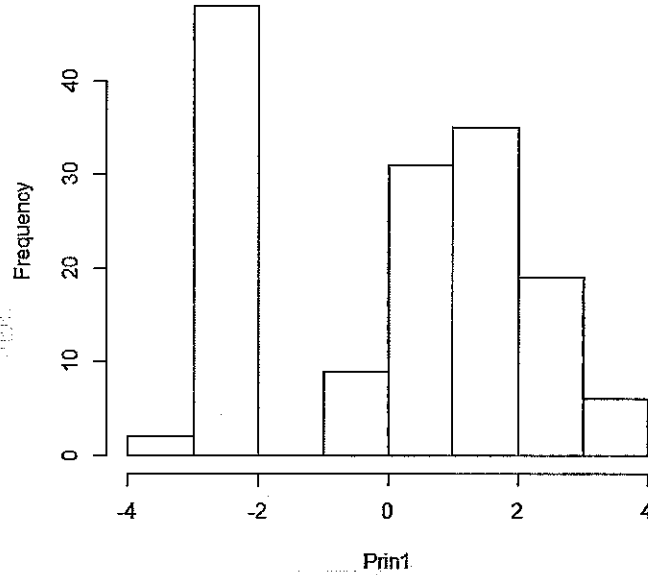


FIGURE 5: Histogram of the first principal component in the Iris data.

(2006), the FDR can be estimated by

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

where $N_r = \sum_i I_{[0, c_0]} \hat{\tau}_0(z_i)$ is the total number of found differentially expressed genes and $I_A(x)$ is the indicator function, which is one if $x \in A$ and is zero otherwise. Table 15 reports the number of selected differentially expressed genes (N_r) and the estimated false discovery rate (FDR) for different threshold c values based on MPHD estimate. For comparison, we also include the results of McLachlan & Wockner (2010), which assumes a two-component mixture of heterogeneous normals (MLE) for z_i s.

5. DISCUSSION

In this paper, we proposed a MPHD estimator for a class of semiparametric mixture models and investigated its existence, consistency and asymptotic normality. Simulation study shows that the

TABLE 13: Estimates of first principal component in Iris data.

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

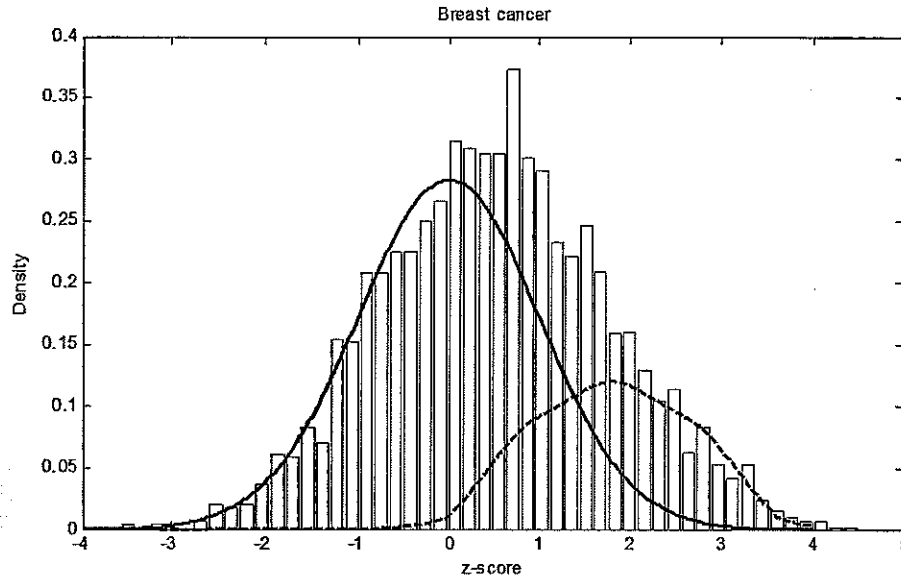


FIGURE 6: Breast cancer data: plot of fitted two-component mixture model with theoretical $N(0, 1)$ null and non-null component (weighted respectively by $\hat{\pi}$ and $(1 - \hat{\pi})$) imposed on histogram of z-score.

MPHD estimator outperforms existing estimators when data are under contamination, while it performs competitively to other estimators when there is no contamination.

We indicated two fields of application of the model. The first is microarray data analysis, which is the initial motivation of introducing model (2) (see Bordes, Delmas, & Vandekerkhove, 2006). The second is sequential clustering algorithm, which is the initial motivation of introducing model (3) (see Song, Nicolae, & Song, 2010). Two real data applications are also provided to illustrate the effectiveness of the proposed methodology.

In this article, we only considered the asymptotic results for model (2), since its identifiability property has been established by Bordes, Delmas, & Vandekerkhove (2006). When the first component of the general model (1) has normal distribution, empirical studies demonstrated the success of proposed MPHD estimator. We conjecture that the asymptotic results of MPHD also apply to the more general model (1) when it is identifiable. However, it requires further research to find sufficient conditions for the identifiability of model (1). In addition, more work remains to be done on the application of MPHD estimation in regression settings such as mixture of regression models.

TABLE 14: Parameter estimates for the breast cancer data.

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

TABLE 15: Estimated FDR for various levels of the threshold c applied to the posterior probability of nondifferentially expression for the breast cancer data.

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

APPENDIX

The proofs of Theorems 1, 2 and 3 are presented in this section.

Proof of Theorem 1. The method of proof is similar to that of Theorem 2.1 of Beran (1977).

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

$$\begin{aligned} |d^2(t_n) - d^2(t)| &= \left| \int (h_{t_n, f(t_n, h_{\theta, f})}^{1/2}(x) - h_{\theta, f}^{1/2}(x))^2 dx - \int (h_{t, f(t, h_{\theta, f})}^{1/2}(x) - h_{\theta, f}^{1/2}(x))^2 dx \right| \\ &= 2 \left| \int (h_{t_n, f(t_n, h_{\theta, f})}^{1/2}(x) - h_{t, f(t, h_{\theta, f})}^{1/2}(x)) h_{\theta, f}^{1/2}(x) dx \right| \\ &\leq 2 \left\| h_{t_n, f(t_n, h_{\theta, f})}^{1/2} - h_{t, f(t, h_{\theta, f})}^{1/2} \right\|. \end{aligned}$$

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

$$\begin{aligned} \left\| h_{t_n, f(t_n, h_{\theta, f})}^{1/2} - h_{t, f(t, h_{\theta, f})}^{1/2} \right\|^2 &= \int \left[h_{t_n, f(t_n, h_{\theta, f})}^{1/2}(x) - h_{t, f(t, h_{\theta, f})}^{1/2}(x) \right]^2 dx \\ &\leq \int \left| h_{t, f(t, h_{\theta, f})}(x) - h_{t_n, f(t_n, h_{\theta, f})}(x) \right| dx \\ &= 2 \int \left[h_{t, f(t, h_{\theta, f})}(x) - h_{t_n, f(t_n, h_{\theta, f})}(x) \right]^+ dx. \end{aligned}$$

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

(ii) By assumption, $h_{\theta, f}$ is identifiable. Immediately, we have $T(h_{\theta, f}) = \theta$ uniquely. ^{Q3}

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

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

Consequently,

$$\begin{aligned} \sup_{t \in \Theta} |d_n(t) - d(t)| &\leq \left\{ 2 \sup_{t \in \Theta} \int \left[h_{t,f(t,h_n)}^{1/2}(x) - h_{t,f(t,h_{\theta,f})}^{1/2}(x) \right]^2 dx \right. \\ &\quad \left. + 2 \int \left[h_n^{1/2}(x) - h_{\theta,f}^{1/2}(x) \right]^2 dx \right\}^{1/2}, \end{aligned} \tag{10}$$

and the right-hand side of (10) goes to zero as $n \rightarrow \infty$ by assumptions. Then with $\theta_0 = T(h_{\theta,f})$ and $\theta_n = T(h_n)$, we have $d_n(\theta_0) \rightarrow d(\theta_0)$ and $d_n(\theta_n) - d(\theta_n) \rightarrow 0$ as $n \rightarrow \infty$.

If $\theta_n \not\rightarrow \theta_0$, then there exists a subsequence $\{\theta_m\} \subseteq \{\theta_n\}$ such that $\theta_m \rightarrow \theta' \neq \theta_0$, implying that $\theta' \in \Theta$ and $d(\theta_m) \rightarrow d(\theta')$ by the continuity of d . From the above result, we have $d_m(\theta_m) - d_m(\theta_0) \rightarrow d(\theta') - d(\theta_0)$. By the definition of θ_m , $d_m(\theta_m) - d_m(\theta_0) \leq 0$, and therefore, $d(\theta') - d(\theta_0) \leq 0$. However, by the definition of θ_0 and the uniqueness of it, $d(\theta') > d(\theta_0)$. This is a contradiction, and therefore $\theta_n \rightarrow \theta_0$. ■

Proof of Theorem 2. Let H_n denote the empirical cdf of X_1, X_2, \dots, X_n , which are assumed i.i.d. with density $h_{\theta,f}$ and cdf H . Let

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

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

$$\begin{aligned} \sup_x |\hat{h}_n(x) - \tilde{h}_n(x)| &= \sup_x n^{-1/2} (c_n s_n)^{-1} \left| \int K((c_n s_n)^{-1}(x - y)) dB_n(y) \right| \\ &\leq n^{-1/2} (c_n s_n)^{-1} \sup_x |B_n(x)| \int |K'(x)| dx \xrightarrow{P} 0. \end{aligned} \tag{11}$$

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

$$\begin{aligned} \sup_x |\tilde{h}_n(x) - h_{\theta,f}(x)| &= \sup_x \left| \int K(t) h_{\theta,f}(x - c_n s_n t) dt - h_{\theta,f}(x) \right| \\ &= \sup_x \left| h_{\theta,f}(x - c_n s_n \xi) \int K(t) dt - h_{\theta,f}(x) \right|, \text{ with } \xi \in [a, b] \\ &\leq \sup_x \sup_{t \in [a,b]} |h_{\theta,f}(x - c_n s_n t) - h_{\theta,f}(x)| \xrightarrow{P} 0 \end{aligned} \tag{12}$$

From (11) and (12), we have

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

From an argument similar to the proof of Theorem 1, $\|\hat{h}_n^{1/2}(x) - h_{\theta, f}^{1/2}(x)\| \xrightarrow{P} 0$ and $\sup_{t \in \Theta} \|h_{x, f(t, \hat{h}_n)} - h_{t, f(t, h_{\theta, f})}\| \rightarrow 0$ as $n \rightarrow \infty$. By Theorem 1, $T(\hat{h}_n) \xrightarrow{P} T(h_{\theta, f})$ as $n \rightarrow \infty$. ■

Proof of Theorem 3. Let

$$D(\theta, g) = \int \dot{s}_{\theta, g}(x) g^{1/2}(x) dx = \langle \dot{s}_{\theta, g}, g^{1/2} \rangle,$$

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

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

Since the map $\theta \mapsto s_{\theta, g}$ satisfies (9) and (9), $D(\theta, g)$ is differentiable in θ with derivative

$$\dot{D}(\theta, g) = \langle \dot{s}_{\theta, g}, g^{1/2} \rangle$$

that is continuous in θ . Then,

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

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

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

Applying the algebraic identity

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

we have that

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

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

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

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

$$\Sigma = \langle \ddot{s}_{\theta, h_{\theta, f}}, h_{\theta, f}^{1/2} \rangle^{-1} \langle \dot{s}_{\theta, h_{\theta, f}}, \dot{s}_{\theta, h_{\theta, f}}^T \rangle \langle \ddot{s}_{\theta, h_{\theta, f}}, h_{\theta, f}^{1/2} \rangle^{-1}$$

ACKNOWLEDGEMENTS

The authors would like to thank the editors and two anonymous referees for their valuable comments and suggestions, which greatly improved this article.

BIBLIOGRAPHY

- Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate—A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.
- Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, 5, 445–463.
- Bordes, L., Delmas, C., & Vandekerckhove, P. (2006). Semiparametric estimation of a two-component mixture model where one-component is known. *Scandinavian Journal of Statistics*, 33, 733–752.
- Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, 2010, 2916–2957.
- Donoho, D. L. & Liu, D. C. (1988). The automatic robustness of minimum distance functionals. *Annals of Statistics*, 16(2), 552–586.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7, Part II, 179–188.
- García-Escudero, L. A., Gordaliza, A., & Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2), 434–449.
- García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Isacar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3), 1324–1345.

- García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Isacar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2), 89–109.
- Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, B., Borg, A., & Trent, J. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344, 539–548.
- Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1), 112–147.
- Langaas, M., Lindqvist, B. H., & Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 555–572.
- Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance estimation and related methods. *Annals of Statistics*, 22, 1081–1114.
- McLachlan, G. J., Bean, R. W., & Jones, L. B. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22, 1608–1615.
- McLachlan, G. J. & Wockner, L. (2010). Use of mixture models in multiple hypothesis testing with applications in bioinformatics. In *Classification as a Tool for Research: Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation*, Locarek-Junge, H. & Weihs, C., editors. Springer-Verlag, Heidelberg, pp. 177–184.
- Punzo, A. & McNicholas, P. D. (2013). Outlier detection via parsimonious mixtures of contaminated Gaussian distributions. ArXiv:1305.4669.
- Song, J. & Nicolae, D. L. (2009). A sequential clustering algorithm with applications to gene expression data. *Journal of the Korean Statistical Society*, 38, 175–184.
- Song, S., Nicolae, D. L., & Song, J. (2010). Estimating the mixing proportion in a semiparametric mixture model. *Computational Statistics and Data Analysis*, 54, 2276–2283.
- Storey, J. D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 3889–3894.
- Swanepoel, J. W. H. (1999). The limiting behavior of a modified maximal symmetric 2s-spacing with applications. *Annals of Statistics*, 27, 24–35.
- Wu, J., Schick, A., & Karunamuni, R. J. (2011). Profile Hellinger distance estimation. Technical Report.^{Q2}

Received 29 June 2013

Accepted 10 February 2014

Q1: Author: A running head short title was not supplied; please check if this one is suitable and, if not, please supply a short title of up to 45 characters that can be used instead.

Q2: Author: Please provide complete details.

Q3: Author: Please check the list numbering here.

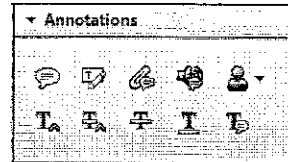
USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

Required software to e-Annotate PDFs: **Adobe Acrobat Professional** or **Adobe Reader (version 8.0 or above)**. (Note that this document uses screenshots from **Adobe Reader X**)
 The latest version of Acrobat Reader can be downloaded for free at: <http://get.adobe.com/reader/>

Once you have Acrobat Reader open on your computer, click on the Comment tab at the right of the toolbar:



This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the Annotations section, pictured opposite. We've picked out some of these tools below:



1. Replace (Ins) Tool – for replacing text.



Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it

- Highlight a word or sentence.
- Click on the Replace (Ins) icon in the Annotations section.
- Type the replacement text into the blue box that appears.

standard framework for the analysis of microeconomics. Nevertheless, it also led to the emergence of strategic behavior. The number of competitors is that the strategic behavior of the main components of the model are extremely important. We have seen the 'black box'...



2. Strikethrough (Del) Tool – for deleting text.



Strikes a red line through text that is to be deleted.

How to use it

- Highlight a word or sentence.
- Click on the Strikethrough (Del) icon in the Annotations section.

there is no room for extra profits and the number of firms are zero and the number of firms (net) values are not determined by Blanchard and Kiyotaki (1987), perfect competition in general equilibrium. The aggregate demand and supply in the classical framework assuming monopolistic competition and an exogenous number of firms...

3. Add note to text Tool – for highlighting a section to be changed to bold or italic.



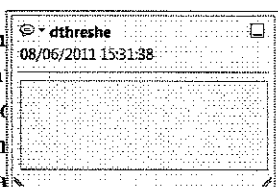
Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the Add note to text icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.

dynamic responses of mark-ups are consistent with the VAR evidence...

...with the demand...



4. Add sticky note Tool – for making notes at specific points in the text.

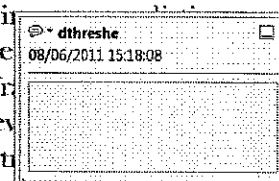


Marks a point in the proof where a comment needs to be highlighted.

How to use it


- Click on the Add sticky note icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.

...and supply shocks. Most of the... number of competitors and the impact is that the structure of the sector...



USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

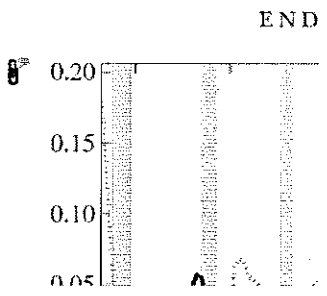
5. Attach File Tool – for inserting large amounts of text or replacement figures.

 Inserts an icon linking to the attached file in the appropriate place in the text.


How to use it

- Click on the Attach File icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.

E N D




6. Add stamp Tool – for approving a proof if no corrections are required.

 Inserts a selected stamp onto an appropriate place in the proof.

How to use it

- Click on the Add stamp icon in the Annotations section.
- Select the stamp you want to use. (The Approved stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).

...of the business cycle, starting with the
...on perfect competition, constant ret
...roduction. In the long-run, the
...es:
...h
...at
...otaki (1987), has introduced produc
...general equilibrium models with nomin
...and ...

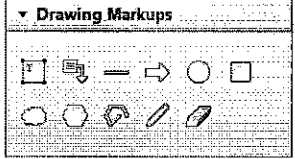
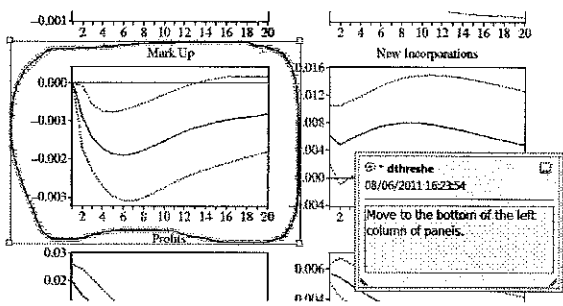


7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

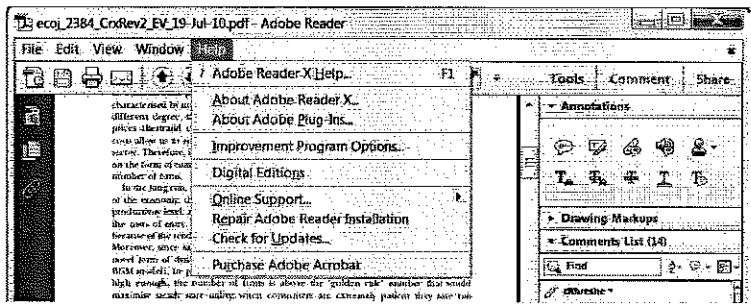
How to use it

- Click on one of the shapes in the Drawing Markups section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.

Move to the bottom of the left column of panels.

For further information on how to annotate proofs, click on the Help menu to reveal a list of further options:





Additional reprint and journal issue purchases

Should you wish to purchase additional copies of your article, please click on the link and follow the instructions provided:

<https://caesar.sheridan.com/reprints/redir.php?pub=10089&acro=CJS>

Corresponding authors are invited to inform their co-authors of the reprint options available.

Please note that regardless of the form in which they are acquired, reprints should not be resold, nor further disseminated in electronic form, nor deployed in part or in whole in any marketing, promotional or educational contexts without authorization from Wiley. Permissions requests should be directed to mailto: permissionsus@wiley.com

For information about 'Pay-Per-View and Article Select' click on the following link: <http://wileyonlinelibrary.com/ppv>