Journal of Multivariate Analysis 137 (2015) 17–31

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Adaptive estimation for varying coefficient models

Yixin Chen [a], Qin Wang [c], Weixin Yao [b,*]

[a] Department of Biostatistics and Programming, Genzyme Corporation (A Sanofi Company), Cambridge, MA 02142, United States

[b] Department of Statistics, University of California, Riverside, CA 92521, United States

[c] Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA 23284, United States

ARTICLE INFO

Article history:
Received 18 July 2014
Available online 7 February 2015

AMS subject classifications:
62G08
62G05

Keywords:
Adaptive estimation
EM algorithm
Kernel smoothing
Local maximum likelihood
Varying coefficient models

ABSTRACT

In this article, a novel adaptive estimation is proposed for varying coefficient models. Unlike the traditional least squares based methods, the proposed approach can adapt to different error distributions. An efficient EM algorithm is provided to implement the proposed estimation. The asymptotic properties of the resulting estimator are established. Both simulation studies and real data examples are used to illustrate the finite sample performance of the new estimation procedure. The numerical results show that the gain of the new procedure over the least squares estimation can be quite substantial for non-Gaussian errors.

1. Introduction

Since the introduction in [5,17], varying coefficient models have gained considerable attention due to their flexibility and good interpretability. They are useful extensions of the classical linear models and have been widely used to explore the dynamic pattern in many scientific areas, such as finance, economics, epidemiology, ecology, etc. By allowing coefficients to vary over the so-called index variable, the modeling bias can be significantly reduced and the 'curse of dimensionality' can be avoided [14]. In recent years, varying coefficient models have experienced rapid developments in both theory and methodology, see, for example, [34,19,12,13,3,11,31,32], etc. We refer to readers to Fan and Zhang [14] for a nice and comprehensive survey.

Let $y \in \mathcal{R}^1$ be the response, $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathcal{R}^d$ be the covariate vector, and $u \in \mathcal{R}^1$ is the index variable. The varying coefficient model is defined as

$$y = \sum_{j=1}^{d} g_j(u)x_j + \epsilon, \tag{1.1}$$

where $\{g_1(u), \ldots, g_d(u)\}^T$ are unknown smooth coefficient functions. Throughout this article, we assume the random error ϵ to be independent of $(u, \mathbf{x})$, with mean 0 and a finite second-order moment σ^2. By setting $x_1 \equiv 1$, it allows a varying intercept in the model.

* Corresponding author.
E-mail addresses: Yixin.Chen@genzyme.com (Y. Chen), qwang3@vcu.edu (Q. Wang), weixin.yao@ucr.edu (W. Yao).

http://dx.doi.org/10.1016/j.jmva.2015.01.017
0047-259X/

Hastie and Tibshirani [17], Hoover et al. [19], Chiang et al. [4] and Eubank et al. [7] proposed using smoothing spline to estimate coefficient functions. Polynomial spline was used in [21,22,20]. Wu et al. [34], Hoover et al. [19], Fan and Zhang [12] and Kauermann and Tutz [23] adopted kernel smoothing to estimate coefficient functions. Fan and Zhang [13] further studied a two-step estimation procedure to deal with the situation where the coefficient functions admit different degrees of smoothness. Recently, Wang and Xia [32] proposed a shrinkage estimation procedure to select important nonparametric components. Wang et al. [31] developed a highly robust and efficient procedure based on local ranks. Nevertheless, most existing methods used least squares type criteria in estimation, which corresponds to the local likelihood when the error ϵ is distributed as a normal random variable. However, in the absence of normality, the traditional least squares based estimators will lose some efficiency.

In this article, we propose a novel adaptive kernel estimation procedure for varying coefficient models. It combines the kernel density estimation and the local maximum likelihood estimation so that the new estimator can adapt to different error distributions. The new estimator is "adaptive" and "efficient" in the sense that it is asymptotically equivalent to the infeasible local likelihood estimator [27,9], which requires the knowledge of the error distribution. An efficient EM algorithm is proposed to implement the adaptive estimation. We demonstrate through a simulation study that the new estimate is more efficient than the existing least squares based kernel estimate when the error distribution deviates from normal. In addition, when the error is exactly normal, the new method is broadly comparable to the existing kernel approach. We further illustrate the effectiveness of the proposed adaptive estimation method through two real data examples.

The rest of the article is organized as follows. In Section 2, we introduce the new adaptive estimation for the varying coefficient models and the EM algorithm. In Section 3, we compare our proposed approach with the traditional least squares based estimation for five different error densities through a simulation study and then apply the new method to two real data examples. We conclude this article with a brief discussion in Section 4. All technical conditions and proofs are relegated to the Appendix.

2. New adaptive estimation

2.1. Introduction to the new method

Suppose that $\{\mathbf{x}_i, u_i, y_i,\ i = 1, \ldots, n\}$ is a random sample from model (1.1). For u in a neighborhood of u_0, we can approximate the varying coefficient functions locally as

$$g_j(u) \approx g_j(u_0) + g_j'(u_0)(u - u_0) \equiv b_j + c_j(u - u_0), \quad \text{for } j = 1, \ldots, d. \tag{2.1}$$

The traditional local linear estimation of (1.1) is to minimize

$$\sum_{i=1}^{n} K_h(u_i - u_0) \left[y_i - \sum_{j=1}^{d} \{b_j + c_j(u_i - u_0)\} x_{ij} \right]^2, \tag{2.2}$$

with respect to $(b_1, \ldots, b_d)$ and $(c_1, \ldots, c_d)$ for a given kernel density $K(\cdot)$ and a bandwidth h, where $K_h(t) = h^{-1}K(t/h)$. It is well known that the choice of kernel function is not critical in terms of estimation efficiency. Throughout this article, a Gaussian kernel will be used for $K(\cdot)$. Due to the least squares in (2.2), the resulting estimate may lose some efficiency when the error distribution is not normal. Therefore, it is desirable to develop an estimation procedure which can adapt to different error distributions.

Let $f(\epsilon)$ be the density function of ϵ. If $f(\epsilon)$ were known, it would be natural to estimate the local parameters in (2.1) by maximizing the following local log-likelihood function

$$\sum_{i=1}^{n} K_h(u_i - u_0) \log f \left[y_i - \sum_{j=1}^{d} \{b_j + c_j(u_i - u_0)\} x_{ij} \right]. \tag{2.3}$$

However, in practice, $f(\epsilon)$ is generally unknown but can be replaced by a kernel density estimate based on the initial estimated residual $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$,

$$\tilde{f}(\epsilon_i) = \frac{1}{n} \sum_{j \neq i}^{n} K_{h_0}\left(\epsilon_i - \tilde{\epsilon}_j\right), \quad \text{for } i, j = 1, 2, \ldots, n \tag{2.4}$$

where $\tilde{\epsilon}_i = y_i - \sum_{j=1}^{d} \tilde{g}_j(u_i) x_{ij}$ and $\tilde{g}_j(\cdot)$ can be estimated by least squares (or L_1 norm, i.e., median regression) based local linear estimate (2.2). Here we use leave-one-out kernel density estimate for $f(\epsilon_i)$ to remove the estimation bias. Let $\boldsymbol{\theta} = (b_1, \ldots, b_d, c_1, \ldots, c_d)^T$. Then our proposed adaptive local linear estimate for the local parameter $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}), \tag{2.5}$$

where

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} K_h(u_i - u_0) \log\left(\frac{1}{n}\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l + c_l(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right]\right). \tag{2.6}$$

The idea of adaptiveness can be traced back to Beran [1] and Stone [29], where the adaptive estimation was proposed for location models. Later, this idea was extended to regression, time series and other models, see [2,25,28,26,6,18,39,38]. Linton and Xiao [24] proposed an elegant adaptive nonparametric regression estimator by maximizing the local likelihood function. In fact, the adaptive method proposed in [24] can be seen as a special case of ours when $d = 1$ in (1.1). Recently, Wang and Yao [33] and Yao and Zhao [37] extended the idea of adaptive estimation to sufficient dimension reduction and linear regression, respectively.

2.2. Computation: an EM algorithm

Unlike least squares criterion, (2.5) does not have an explicit solution due to the summation inside the log function, which is similar to the mixture structure. In this section, we propose an EM algorithm to maximize it by extending the generalized modal EM algorithm proposed in [36].

Let $\boldsymbol{\theta}^{(0)}$ be an initial parameter estimate, such as the least squares (or L_1 norm, i.e., median regression) based local linear estimate. We can update the parameter estimate according to the following algorithm.

Algorithm 2.1. At $(k+1)$th step, we calculate the following E and M steps:

E-Step: Calculate the classification probabilities $p_{ij}^{(k+1)}$,

$$\begin{aligned} p_{ij}^{(k+1)} &= \frac{K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l^{(k)} + c_l^{(k)}(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right]}{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l^{(k)} + c_l^{(k)}(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right]} \\ &\propto K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l^{(k)} + c_l^{(k)}(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right], \quad 1 \le j \neq i \le n. \end{aligned} \tag{2.7}$$

M-Step: Update $\boldsymbol{\theta}^{(k+1)}$,

$$\begin{aligned} \boldsymbol{\theta}^{(k+1)} &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j\neq i}\left\{p_{ij}^{(k+1)} K_h(u_i - u_0)\log\left(K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l + c_l(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right]\right)\right\} \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j\neq i}\left\{p_{ij}^{(k+1)} K_h(u_i - u_0)\left[y_i - \tilde{\epsilon}_j - \mathbf{z}_i^T\boldsymbol{\theta}\right]^2\right\}, \\ &= \left(\sum_{i=1}^{n}\sum_{j\neq i} p_{ij}^{(k+1)} K_h(u_i - u_0)\mathbf{z}_i\mathbf{z}_i^T\right)^{-1} \sum_{i=1}^{n}\sum_{j\neq i} p_{ij}^{(k+1)} K_h(u_i - u_0)(y_i - \tilde{\epsilon}_j)\mathbf{z}_i \end{aligned} \tag{2.8}$$

where $\mathbf{z}_i = \{\mathbf{x}_i^T, \mathbf{x}_i^T(u_i - u_0)\}^T$ and the second equation follows the use of Gaussian kernel for density estimation.

The above EM algorithm monotonically increases the estimated local log-likelihood (2.6) after each iteration, as shown in the following proposition. Its proof is given in the Appendix.

Proposition 2.1. *Each iteration of the above E and M steps will monotonically increase the local log-likelihood (2.6), i.e.,*

$$Q(\boldsymbol{\theta}^{(k+1)}) \geqslant Q(\boldsymbol{\theta}^{(k)}),$$

for all k, where $Q(\cdot)$ is defined as in (2.6).

2.3. Asymptotic result

We now establish the consistency and derive the asymptotic distribution of the proposed adaptive local linear estimator of $\boldsymbol{\theta}$. Define $\mu_k = \int u^k K(u)\mathrm{d}u$ and $\nu_k = \int u^k K^2(u)\mathrm{d}u$. Let $\mathbf{H} = \mathrm{diag}(1, h) \otimes \mathbf{I}_d$ with $\otimes$ denoting the Kronecker product and $\mathbf{I}_d$ being the $d \times d$ identity matrix. Let $q(\cdot)$ denote the marginal density of u, and

$$\Gamma_{jk}(u_i) = \mathrm{E}(x_{ij}x_{ik}|u_i) \quad \text{for } 1 \le j, k \le d,\ i = 1, \ldots, n, \tag{2.9}$$

$$\boldsymbol{\Gamma}(u_0) = \left\{\Gamma_{jk}(u_0)\right\}_{1\leqslant j,k\leqslant d}. \tag{2.10}$$

Theorem 2.1. *Under the regularity conditions in the Appendix, with probability approaching* 1, *there exists a consistent local maximizer* $\hat{\boldsymbol{\theta}} = (\hat{b}_1, \ldots, \hat{b}_d, \hat{c}_1, \ldots, \hat{c}_d)^T$ *of* (2.6) *such that*

$$\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_p\{(nh)^{-1/2} + h^2\}.$$

Based on Theorem 2.1, we can know that the proposed adaptive local linear estimator of $\boldsymbol{\theta}$ is consistent and its proof is provided in the Appendix. Next, we provide the asymptotic distribution of the proposed estimator.

Theorem 2.2. *Suppose that the regularity conditions in the Appendix hold. Then* $\hat{\boldsymbol{\theta}}$, *given in Theorem* 2.1, *has the following asymptotic distribution*

$$\sqrt{nh}\left\{\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \mathbf{S}^{-1}\frac{h^2}{2}\sum_{j=1}^{d} g_j''(u_0)\boldsymbol{\psi}_j(1 + o_p(1))\right\} \xrightarrow{D} N(\mathbf{0}_{2d}, [E\{\rho'(\epsilon_i)^2\}]^{-1}q(u_0)^{-1}\mathbf{S}^{-1}\boldsymbol{\Lambda}\mathbf{S}^{-1}),$$

where $\mathbf{0}_{2d}$ *is a* $2d \times 1$ *vector with each entry being* 0, $\rho(\cdot) = \log f(\cdot)$, $\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \otimes \boldsymbol{\Gamma}(u_0)$, $\boldsymbol{\Lambda} = \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \boldsymbol{\Gamma}(u_0)$, $\boldsymbol{\psi}_j = \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \otimes \left(\Gamma_{jk}(u_0)\right)_{1\le k\le d}^T$, *and* $\boldsymbol{\Gamma}(u_0)$ *is given by* (2.10).

A sketch of the proof of the above theorems is provided in the Appendix. As shown in [24], one important property of the proposed adaptive estimate is that it achieves the same asymptotic efficiency as if the error density were known. Therefore, estimating f by kernel density estimation will not affect the asymptotic distribution of the resulting estimator of $\boldsymbol{\theta}$. As Linton and Xiao [24] pointed out that such a new estimation method can "*do as well as the corresponding estimator one would compute if one knew the error density*". However it is not possible to achieve the lower bound here [8]. Any specific estimator can be bettered for some specific model setting.

Note that the least squares based local linear estimate [40], by minimizing (2.2), has the same asymptotic bias as the new method but slightly different asymptotic variance, which replaces $[\mathrm{E}\{\rho'(\epsilon_i)^2\}]^{-1}$ by $\sigma^2 = \mathrm{E}(\epsilon^2)$. Based on Cauchy–Schwarz inequality, we have

$$\mathrm{E}\{\rho'(\epsilon_i)^2\}\mathrm{E}(\epsilon^2) \ge [\mathrm{E}\{\epsilon\rho'(\epsilon)\}]^2 = 1$$

and the equality holds if and only if $\rho'(\epsilon) \propto \epsilon$, i.e., $f(\epsilon)$ is a normal density. Therefore, $[\mathrm{E}\{\rho'(\epsilon_i)^2\}]^{-1} \le \sigma^2$ and the asymptotic variance of the new estimator is no larger than that of least squares based local linear estimate for any error density $f(\epsilon)$.

3. Examples

3.1. Simulation study

In this section, we conduct a simulation study to compare the proposed adaptive estimation (Adapt) with the traditional least squares based kernel estimation (LS) for varying coefficient models. The following five error distributions of ϵ were considered in our numerical experiment:

1. $N(0, 1)$;
2. t_3;
3. $0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$;
4. $0.3N(-1.4, 1) + 0.7N(0.6, 0.4^2)$;
5. $0.9N(0, 1) + 0.1N(0, 10^2)$.

The standard normal distribution serves as a baseline in our comparison. The second one is a t-distribution with 3 degrees of freedom. The third density is bimodal and the fourth one is left skewed. The last one is a contaminated normal mixture distribution, where 10% of the data from $N(0, 10^2)$ are most likely to be outliers.

For each of the above error distributions, we consider the following two models:

Model 1: $y = g_1(u) + g_2(u)x_1 + g_3(u)x_2 + \epsilon$, where $g_1(u) = \exp(2u - 1)$, $g_2(u) = 8u(1 - u)$, and $g_3(u) = 2\sin^2(2\pi u)$.
Model 2: $y = g_1(u) + g_2(u)x_1 + g_3(u)x_2 + \epsilon$, where $g_1(u) = \sin(2\pi u)$, $g_2(u) = (2u-1)^2 + 0.5$, and $g_3(u) = \exp(2u-1) - 1$.

In both models, x_1 and x_2 follow a standard normal distribution with correlation coefficient $\gamma = 1/\sqrt{2}$. The index variable u is a uniform random variable on [0, 1], and is independent of (x_1, x_2). There are two bandwidths in the estimation, h in the local log-likelihood and h_0 in the kernel density estimation. An asymptotic optimal h can be found by minimizing the asymptotic mean squared errors provide in Theorem 2.2 and can be estimated by a plug-in estimator which replaces the unknown quantities in Theorem 2.2 by their estimates. In our examples, the bandwidth h is chosen by leave-one-out cross-validation with more details in [12], and $h_0 = h/\log(n)$ following Linton and Xiao [24]. The performance of estimator $\hat{g}(\cdot)$ is assessed via the square root of the average squared errors (RASE; Cai et al. [3]; Wang et al. [31]),

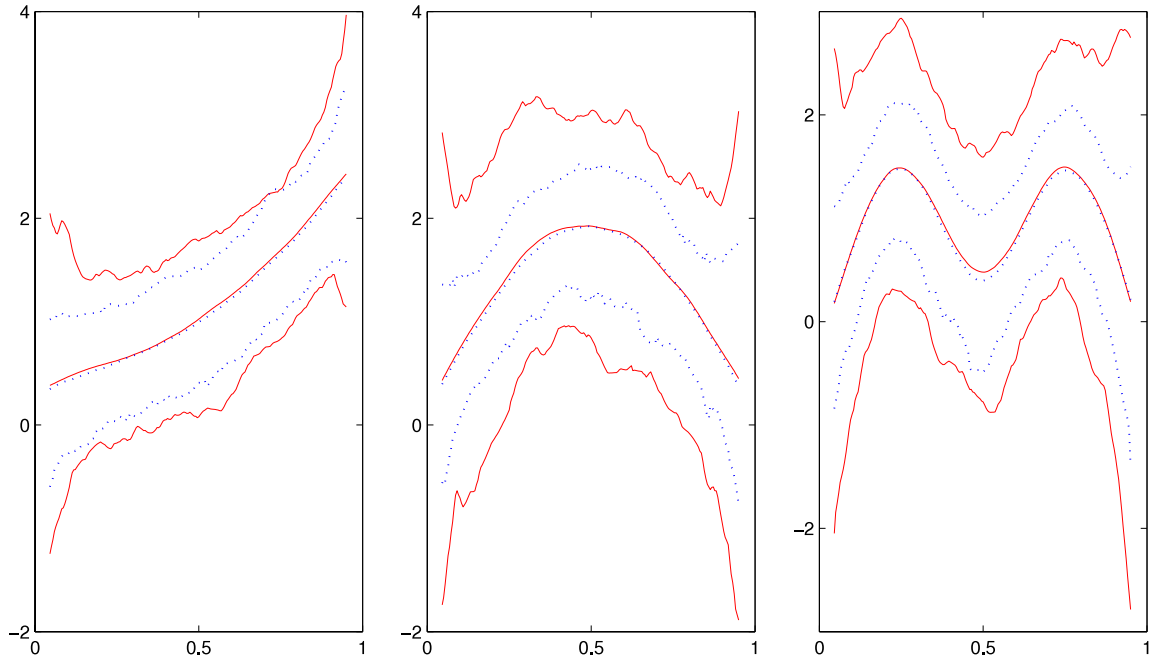
$$\mathrm{RASE}^2 = \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{3}[\hat{g}_j(u_k) - g_j(u_k)]^2, \tag{3.1}$$

Table 1
Model 1 estimation accuracy comparison—RASE and its standard error in brackets.

ϵ	$n = 200$		$n = 400$	
	LS	Adapt	LS	Adapt
1	0.483(0.079)	0.439(0.081)	0.366(0.053)	0.324(0.053)
2	0.671(0.167)	0.601(0.139)	0.493(0.111)	0.422(0.086)
3	0.500(0.083)	0.401(0.077)	0.379(0.061)	0.277(0.048)
4	0.508(0.088)	0.376(0.082)	0.383(0.062)	0.262(0.045)
5	1.188(0.411)	0.720(0.220)	0.871(0.227)	0.459(0.098)

Table 2
Model 2 estimation accuracy comparison—RASE and its standard error in brackets.

ϵ	$n = 200$		$n = 400$	
	LS	Adapt	LS	Adapt
1	0.362(0.077)	0.380(0.074)	0.263(0.051)	0.275(0.049)
2	0.618(0.301)	0.566(0.201)	0.431(0.129)	0.384(0.076)
3	0.412(0.091)	0.351(0.080)	0.290(0.059)	0.215(0.041)
4	0.407(0.102)	0.319(0.089)	0.291(0.061)	0.207(0.051)
5	1.133(0.397)	0.669(0.224)	0.828(0.224)	0.436(0.101)



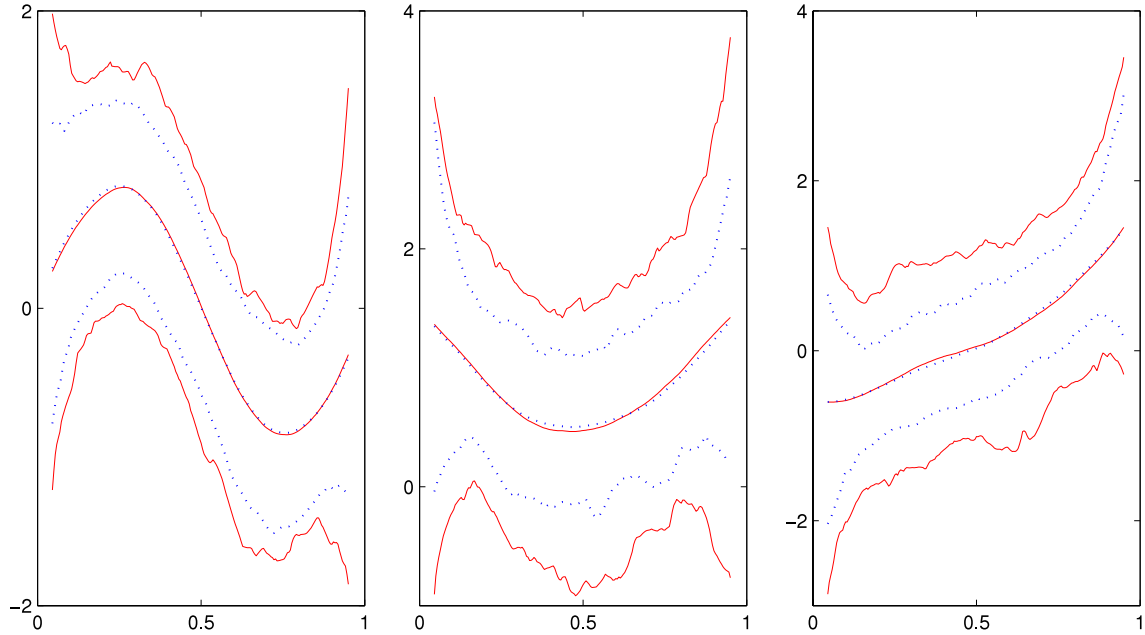
(a) $g_1(u)$. (b) $g_2(u)$. (c) $g_3(u)$.

Fig. 1. Estimated coefficient functions with 95% pointwise confidence intervals (blue dotted line for Adapt and red solid line for LS) for model 1.

where u_k, $k = 1, \ldots, N$, are the equally spaced grid points at which the functions $g_j(\cdot)$ were evaluated. We conduct two sets of simulations with sample size $n = 200$ and 400 respectively, each with 200 data replications.

The simulation results are summarized in Tables 1 and 2. We can clearly see that the proposed adaptive estimation outperforms the least squares method when the error is non-normal. The gain in estimation efficiency can be quite substantial even for moderate sample sizes. When the error follows exactly normal distribution, our approach is still broadly comparable with the least squares based method.

Figs. 1 and 2 plot the estimated coefficient functions and the 95% pointwise confidence intervals based on a typical sample when $n = 200$ and the error distribution is the contaminated normal mixture (Case 5). Due to the complex forms of the asymptotic standard errors of the coefficient functions, similar to Wang, Kai and Li [31], we adopt the bootstrap method to calculate the 95% pointwise confidence intervals. As expected, the adaptive estimation method provides narrower confidence intervals than the least squares based method, since the adaptive method provides more accurate estimate than the least squares estimate when the error is not normal.



(a) $g_1(u)$. (b) $g_2(u)$. (c) $g_3(u)$.

Fig. 2. Estimated coefficient functions with 95% pointwise confidence intervals (blue dotted line for Adapt and red solid line for LS) for model 2.

3.2. Real-data applications

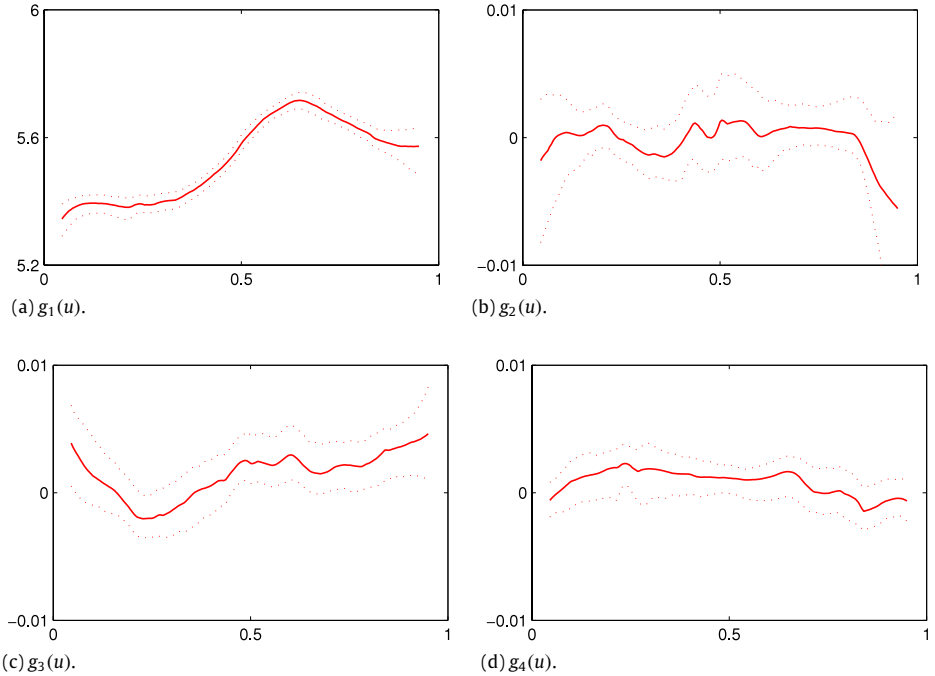
Example 1 (*Hong Kong Environmental Data*)**.** We now illustrate the adaptive estimation method via an application to an environmental data set. The data were collected daily in Hong Kong from January 1, 1994, to December 31, 1995 and have been analyzed by Fan and Zhang [12], Cai et al. [3], Xia et al. [35] and Fan and Zhang [14]. In this data set, a collection of daily measurements of pollutants and other environmental factors are included. Following Fan and Zhang [12], we consider three pollutants: sulfur dioxide x_2 (in $\mu g/m^3$), nitrogen dioxide x_3 (in $\mu g/m^3$), and respirable suspended particulates x_4 (in $\mu g/m^3$) (this variable is named as 'dust' in [12,14,3]). The response variable y is the logarithm of the number of daily hospital admissions. We set $x_1 = 1$ as the intercept term and let u denote time which is scaled to the interval [0, 1]. As in the previous analyses, all three predictors are centered. The following varying coefficient model is considered to investigate the relationship between y and the levels of pollutants x_2, x_3, and x_4.

$$y = g_1(u) + g_2(u)x_2 + g_3(u)x_3 + g_4(u)x_4 + \epsilon.$$

We set aside 50 observations as the test set. The bandwidth h, selected by leave-one-out cross-validation, is around 0.146. The estimated coefficient functions together with 95% pointwise confidence intervals are depicted in Fig. 3. We also compare the median squared prediction errors, MSPE $=$ Median$\{(y_j - \hat{y}_j)^2, j = 1, \ldots, k\}$, from our adaptive approach and the traditional least squares estimation, where $k = 50$ and $\hat{y}_j = \hat{g}_1(u_j) + \hat{g}_2(u_j)x_{j2} + \hat{g}_3(u_j)x_{j3} + \hat{g}_4(u_j)x_{j4}$. The MSPE from our adaptive approach is 0.0183, compared to 0.0178 from the LS estimation.

In Fig. 5(a), we give the residual QQ-plot for Hong Kong environmental data. From the plot, we can see that the residual is very close to normal, which explains why the MSPE of the adaptive approach is close to the MSPE of the LS estimation.

Example 2 (*Boston Housing Data*)**.** The Boston Housing Data (corrected version in [16]), which has been analyzed by Fan and Huang [11], Wang and Xia [32] and Sun et al. [30], is publicly available in the R package *mlbench*, (http://cran.r-project.org/). This data set includes the median value of owner-occupied homes in 506 US census tracts of the Boston area in 1970 and several variables that might explain the variation of housing values. Following Fan and Huang [11] and Wang and Xia [32], we considered seven independent variables: CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), TAX (full-value property-tax rate per $10,000), NOX (nitric oxides concentration parts per 10 million), PTRATIO (pupil–teacher ratio by town), AGE (proportion of owner-occupied units built prior to 1940), and LSTAT (lower status of the population). The response variable is CMEDV (corrected median value of owner-occupied homes in USD 1000's). We denote the covariates CRIM, RM, TAX, NOX, PTRATIO and AGE to be $x_2, x_3, \ldots, x_7$, respectively. Let $x_1 = 1$ be the intercept term and $u = \sqrt{\text{LSTAT}}$ be the index variable. By doing so, we can fit different regression models at different lower status population percentage [11]. Following Fan and Huang [11] we use the square root transformation on the index variable LSTAT to make



(a) $g_1(u)$. (b) $g_2(u)$.

(c) $g_3(u)$. (d) $g_4(u)$.

Fig. 3. Estimated coefficient functions (solid curves) with 95% pointwise confidence intervals (dotted curves) for Hong Kong environmental data.

the data symmetrically distributed. The following varying coefficient model was fit to the data,

$$y_i = g_1(u_i) + \sum_{j=2}^{7} g_j(u_i)x_{ij} + \epsilon_i.$$

Similar to the analysis in the previous example, we set aside 50 observations for checking prediction errors. The bandwidth h was selected by leave-one-out cross-validation, which is around 0.294. The estimated coefficient functions are depicted in Fig. 4. From the plot, we can see that the coefficient functions of x_2 (CRIM) and x_3 (RM) vary over time. The coefficient functions of x_4 (TAX), x_5 (NOX), and x_7 (AGE) are very close to zero and the coefficient function of x_6 (PTRATIO) shows no significant trend. These discoveries are consistent with those from [11,32]. In terms of the median squared prediction error (MSPE), the MSPE from our adaptive approach is 0.0484, compared to 0.0604 from the LS estimation.

In Fig. 5(b), the QQ-plot of residuals from the above fit showed a clear deviation from normality, which explains why the MSPE from the adaptive approach is much smaller than the MSPE from the LS estimation.

4. Discussion

In this article, we proposed an adaptive estimation for varying coefficient models. The new estimation procedure can adapt to different errors and thus provide a more efficient estimate than the traditional least squares based estimate. Simulation studies and two real data applications confirmed our theoretical findings.

It will be interesting to know whether we can also perform some adaptive hypothesis tests for the coefficient functions using the estimated error density. For example, we might be interested in testing some parametric assumptions, such as constant or zero, for the coefficient functions. It requires more research about whether the Wilks phenomenon for generalized likelihood ratio statistic proposed by Fan et al. [15] still holds for the proposed adaptive varying coefficient models.

The idea of the proposed adaptive estimator might also be generalized to many other models, such as varying coefficient partial linear models and nonparametric additive models. In addition, by combining this adaptive idea with shrinkage estimation, we can develop adaptive variable selection procedures. Such study is under way.

Zhang and Lee [40] investigated variable bandwidth selection for varying coefficient models and studied asymptotic properties of the resulting estimators and the bandwidth. It is our interest to extend their variable bandwidth selection method and the corresponding asymptotic properties to our adaptive estimation procedure.

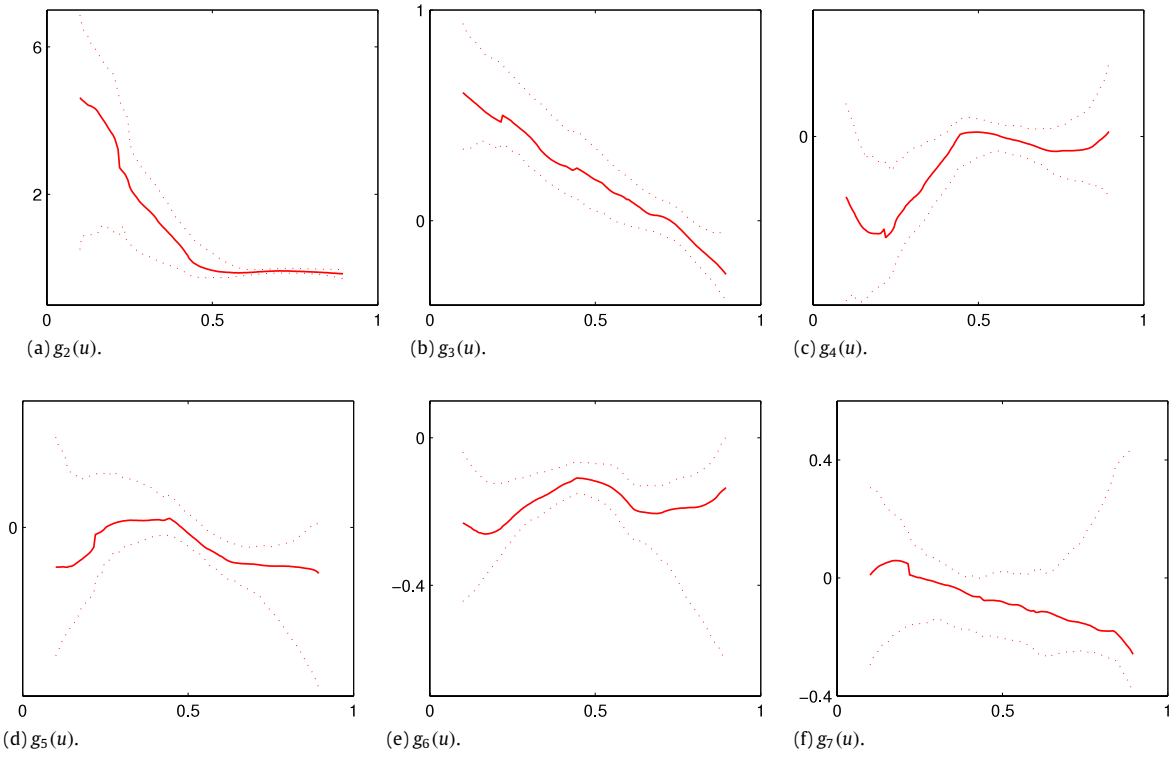


Fig. 4. Estimated coefficient functions (solid curves) with 95% pointwise confidence intervals (dotted curves) for Boston housing data.

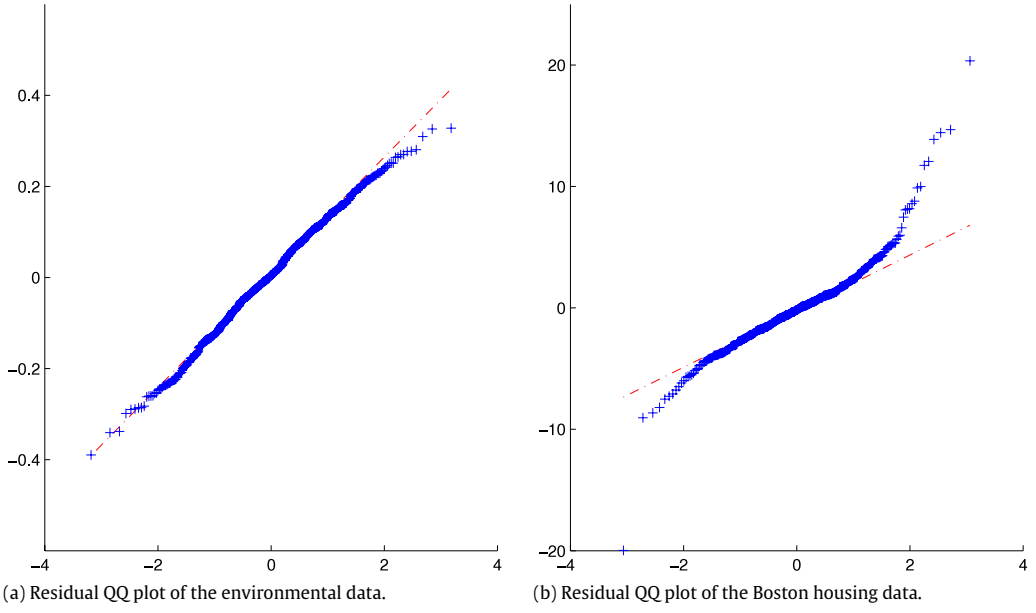


Fig. 5. Residual QQ-plot for two data examples: (a) Hong Kong environmental data; (b) Boston housing data.

As one referee pointed out that we could also extend the idea of Yuan and De Gooijer [39] to derive an adaptive estimate for varying coefficient model. Let $\epsilon_i(\boldsymbol{\theta}) = y_i - \sum_{l=1}^{d}[b_l + c_l(u_i - u_0)]$, and

$$f_n(\epsilon_i(\boldsymbol{\theta})) = \frac{1}{n-1}\sum_{j\neq i} K_h(r(\epsilon_i(\boldsymbol{\theta})) - r(\epsilon_j(\boldsymbol{\theta}))).$$

Based on Yuan and De Gooijer [39], we can estimate $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} K_h(u_i - u_0) \log f_n(\epsilon_i(\boldsymbol{\theta})).$$

Here, $r(\cdot)$ is some monotone nonlinear function that is used to avoid the cancellation of the intercept terms b_ls in $f_n(\epsilon_i(\boldsymbol{\theta}))$. One advantage of the above method is that it does not require an initial estimate. However, compared to the proposed estimate in this paper, the asymptotic variance of the above estimator depends on the choice of $r(\cdot)$ and generally does not reach the Cramér–Rao lower bound for a nonlinear function $r(\cdot)$. In addition, the computation of the above estimator is also more expensive due to the nonlinear function $r(\cdot)$.

Acknowledgments

The authors thank the editor, the associate editor, and reviewers for their constructive comments that have led to a dramatic improvement of the earlier version of this article. Yao's research is supported by NSF grant DMS-1461677.

Appendix

We first list the regularity conditions used in our proof.

Conditions. 1. $K(\cdot)$ is bounded, symmetric, and has bounded support and bounded derivative;
2. $\{\mathbf{x}_i\}_i$, $\{u_i\}_i$, $\{\epsilon_i\}_i$ are independent and identically distributed and $\{\epsilon_i\}_i$ is independent of $\{\mathbf{x}_i\}_i$ and $\{u_i\}_i$, where $\{\mathbf{x}_i\}_i$ means $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, same for notations $\{u_i\}_i$ and $\{\epsilon_i\}_i$. Additionally, the predictor $\mathbf{x}$ has a bounded support;
3. The probability distribution function $f(\cdot)$ of ϵ has bounded continuous derivatives up to order 4. Let $\rho(\epsilon) = \log f(\epsilon)$. Assume $\mathrm{E}[\rho'(\epsilon_i)] = 0$, $\mathrm{E}[\rho''(\epsilon_i)] < \infty$, $\mathrm{E}[\rho'(\epsilon_i)^2] < \infty$ and $\rho'''(\cdot)$ is bounded;
4. The marginal density of u has a continuous second derivative in some neighborhood of u_0 and $q(u_0) \neq 0$;
5. $h \to 0$, $nh \to \infty$ as $n \to \infty$ and $h_0 = h/\log(n)$;
6. $g_j(\cdot)$ has bounded, continuous 3rd derivatives for $1 \leq j \leq d$.

These conditions are adopted from [12,24]. They are not the weakest possible conditions. For instance, we can relax the bounded support assumption of $K(\cdot)$. All the asymptotic results still hold if we put a restriction on the tail of $K(\cdot)$. For example, $\limsup_{t\to\infty} |K(t)t^5| < \infty$ [10]. The independence of $\{\mathbf{x}_i\}_i$ and $\{\epsilon_i\}_i$ can be relaxed based on the discussion of Section 4 of Linton and Xiao [24].

Proof of Proposition 2.1. Note that

$$\begin{aligned}
Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^{n} K_h(u_i - u_0) \log \left\{ \frac{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k+1)} + c_l^{(k+1)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right\} \\
&= \sum_{i=1}^{n} K_h(u_i - u_0) \log \sum_{j\neq i} \left(\frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right) \\
&\quad \times \left(\frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k+1)} + c_l^{(k+1)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right) \\
&= \sum_{i=1}^{n} K_h(u_i - u_0) \log \left\{ \sum_{j\neq i} p_{ij}^{(k+1)} \frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k+1)} + c_l^{(k+1)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right\},
\end{aligned}$$

where

$$p_{ij}^{(k+1)} = \frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \} x_{il} - \tilde{\epsilon}_j \right]}{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d} \{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \} x_{il} - \tilde{\epsilon}_j \right]}.$$

From Jensen's inequality, we have

$$Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \geqslant \sum_{i=1}^{n} K_h(u_i - u_0) \sum_{j\neq i} p_{ij}^{(k+1)} \log \left\{ \frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k+1)} + c_l^{(k+1)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right\}.$$

Based on the property of M-step of (2.8), we have $Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \geq 0$. □

Proof of Theorem 2.1. Note that the estimator $\hat{\boldsymbol{\theta}}$ is the maximizer of the following objective function

$$\arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} K_h(u_i - u_0) \log \tilde{f}\left[y_i - \sum_{l=1}^{d} \{b_l + c_l(u_i - u_0)\} x_{il} \right], \tag{4.1}$$

where

$$\tilde{f}(\epsilon_i) = \frac{1}{n} \sum_{j\neq i} K_{h_0}\left(\epsilon_i - \tilde{\epsilon}_j\right)$$

is the kernel density estimate of $f(\cdot)$, and $\tilde{\epsilon}_i$ is the residual based on the least squares local linear estimate. By the adaptive nonparametric regression result of Linton and Xiao [24], the asymptotic result of $\hat{\boldsymbol{\theta}}$ in (4.1) is the same whether the true density $f(\cdot)$ is used or not. Therefore, we will mainly show the existence and asymptotic distribution of $\hat{\boldsymbol{\theta}}$ assuming $f(\cdot)$ is known.

We will first prove that with probability approaching 1, there exists a consistent local maximizer $\hat{\boldsymbol{\theta}} = (\hat{b}_1, \ldots, \hat{b}_d, \hat{c}_1, \ldots, \hat{c}_d)^T$ of (2.6) such that

$$\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_p\{(nh)^{-1/2} + h^2\}.$$

Then we establish the asymptotic distributions for such consistent estimate.

Denote $\boldsymbol{\theta}^* = \mathbf{H}\boldsymbol{\theta}$, $\mathbf{x}_i^* = (x_{i1}, x_{i2}, \ldots, x_{id}, (\frac{u_i - u_0}{h})x_{i1}, \ldots, (\frac{u_i - u_0}{h})x_{id})^T$, $K_i = K_h(u_i - u_0)$, $R(u_i, \mathbf{x}_i) = \sum_{j=1}^{d} g_j(u_i)x_{ij} - \sum_{j=1}^{d}[b_j + c_j(u_i - u_0)]x_{ij}$, and $a_n = (nh)^{-1/2} + h^2$. Let $\rho(\cdot) = \log f(\cdot)$, we have the objective function

$$L(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} K_i \rho(y_i - \boldsymbol{\theta}^{*T} \mathbf{x}_i^*) = L(\boldsymbol{\theta}^*).$$

It is sufficient to show that for any given $\eta > 0$, there exists a large constant c such that

$$P\left\{ \sup_{\|\mu\|=c} L(\boldsymbol{\theta}^* + a_n\mu) < L(\boldsymbol{\theta}^*) \right\} \geq 1 - \eta,$$

where μ has the same dimension as $\boldsymbol{\theta}$, a_n is the convergence rate. By using Taylor expansion, it follows that

$$\begin{aligned} L(\boldsymbol{\theta}^* + a_n\mu) - L(\boldsymbol{\theta}^*) &= \frac{1}{n} \sum_{i=1}^{n} K_i \{\rho(\epsilon_i + R(u_i, \mathbf{x}_i) - a_n\mu^T \mathbf{x}_i^*) - \rho(\epsilon_i + R(u_i, \mathbf{x}_i))\} \\ &= -\frac{1}{n} \sum_{i=1}^{n} K_i \rho'(\epsilon_i + R(u_i, \mathbf{x}_i)) a_n \mu^T \mathbf{x}_i^* + \frac{1}{2n} \sum_{i=1}^{n} K_i \rho''(\epsilon_i + R(u_i, \mathbf{x}_i)) a_n^2 (\mu^T \mathbf{x}_i^*)^2 \\ &\quad - \frac{1}{6n} \sum_{i=1}^{n} K_i \rho'''(z_i) a_n^3 (\mu^T \mathbf{x}_i^*)^3 \\ &\triangleq I_1 + I_2 + I_3, \end{aligned}$$

where z_i is a value between $\epsilon_i + R(u_i, \mathbf{x}_i) - a_n\mu^T\mathbf{x}_i^*$ and $\epsilon_i + R(u_i, \mathbf{x}_i)$. For $I_1 = -\frac{1}{n}\sum_{i=1}^{n} K_i \rho'(\epsilon_i + R(u_i, \mathbf{x}_i)) a_n \mu^T \mathbf{x}_i^*$, $\mathrm{E}(I_1) = -\mathrm{E}\left[K_i \rho'(\epsilon_i + R(u_i, \mathbf{x}_i)) a_n \mu^T \mathbf{x}_i^*\right]$. We have,

$$\rho'(\epsilon_i + R(u_i, \mathbf{x}_i)) \approx \rho'(\epsilon_i) + \rho''(\epsilon_i) R(u_i, \mathbf{x}_i) + \frac{1}{2}\rho'''(\epsilon_i) R^2(u_i, \mathbf{x}_i).$$

Based on the assumption that ϵ is independent of u and $\mathbf{x}$, and $\mathrm{E}[\rho'(\epsilon_i)] = 0$, we have

$$\mathrm{E}(I_1) \approx -a_n \mathrm{E}\left\{ K_i \left[\rho''(\epsilon_i) R(u_i, \mathbf{x}_i) + \frac{1}{2}\rho'''(\epsilon_i) R^2(u_i, \mathbf{x}_i) \right] \mu^T \mathbf{x}_i^* \right\}.$$

Since

$$\begin{aligned} R(u_i, \mathbf{x}_i) &= \sum_{j=1}^{d} g_j(u_i)x_{ij} - \sum_{j=1}^{d}[b_j + c_j(u_i - u_0)]x_{ij} \\ &= \sum_{j=1}^{d}\left[\sum_{m=2}^{\infty}\frac{1}{m!}g_j^{(m)}(u_0)(u_i - u_0)^m\right]x_{ij} \\ &= O_p(h^2), \end{aligned}$$

then $\frac{1}{2}\rho'''(\epsilon_i)R^2(u_i, \mathbf{x}_i) = [O_p(h^2)]^2 = O_p(h^4)$, which is a smaller order than $\rho''(\epsilon_i)R(u_i, \mathbf{x}_i)$. Thus,

$$\mathrm{E}(I_1) \approx -a_n\mathrm{E}\left\{K_i\rho''(\epsilon_i)R(u_i, \mathbf{x}_i)\mu^T\mathbf{x}_i^*\right\} = -a_n\mathrm{E}\left[\rho''(\epsilon_i)\right]\mathrm{E}\left[K_iR(u_i, \mathbf{x}_i)\mu^T\mathbf{x}_i^*\right].$$

Let $\delta_1 = \mathrm{E}\left\{\rho''(\epsilon_i)\right\}$, then

$$\mathrm{E}(I_1) \approx -a_n\delta_1\mathrm{E}\left[K_iR(u_i, \mathbf{x}_i)\mu^T\mathbf{x}_i^*\right] = -a_n\delta_1\mathrm{E}\left\{\mathrm{E}\left\{R(u_i, \mathbf{x}_i)\mu^T\mathbf{x}_i^*|u_i\right\}K_i\right\}.$$

By $\mu^T\mathbf{x}_i^* \le \|\mu\| \cdot \left\|\mathbf{x}_i^*\right\| = c\left\|\mathbf{x}_i^*\right\|$, we have $\mathrm{E}(I_1) = O(a_nch^2)$.

$$\mathrm{var}(I_1) = \frac{1}{n}\mathrm{var}\left\{K_i\rho'(\epsilon_i + R(u_i, \mathbf{x}_i))a_n\mu^T\mathbf{x}_i^*\right\} = \frac{1}{n}\{\mathrm{E}(A^2) - [\mathrm{E}(A)]^2\},$$

where $A = K_i\rho'(\epsilon_i + R(u_i, \mathbf{x}_i))a_n\mu^T\mathbf{x}_i^*$. Let $\delta_2 = \mathrm{E}\left\{\rho'(\epsilon_i)^2\right\}$, then

$$\begin{aligned} \mathrm{E}(A^2) &= \mathrm{E}\left\{K_i^2\rho'(\epsilon_i + R(u_i, \mathbf{x}_i))^2a_n^2(\mu^T\mathbf{x}_i^*)^2\right\} \\ &\approx a_n^2\mathrm{E}\left\{K_i^2\rho'(\epsilon_i)^2(\mu^T\mathbf{x}_i^*)^2\right\} \\ &= a_n^2\delta_2\mathrm{E}\left\{\mathrm{E}\left\{(\mu^T\mathbf{x}_i^*)^2|u_i\right\}K_i^2\right\} \\ &= O\left(a_n^2c^2\frac{1}{h}\right). \end{aligned}$$

Note that $[\mathrm{E}(A)]^2 = \left[O(a_nch^2)\right]^2 \ll \mathrm{E}(A^2)$, then $\mathrm{var}(I_1) \approx \frac{1}{n}\mathrm{E}(A^2) = O\left(a_n^2c^2\frac{1}{nh}\right)$. Hence, $I_1 = \mathrm{E}(I_1) + O_p(\sqrt{\mathrm{var}(I_1)}) = O_p(a_nch^2) + O_p\left(\sqrt{a_n^2c^2\frac{1}{nh}}\right) = O_p(ca_n^2)$. For $I_2 = \frac{1}{2n}\sum_{i=1}^{n}K_i\rho''(\epsilon_i + R(u_i, \mathbf{x}_i))a_n^2(\mu^T\mathbf{x}_i^*)^2$,

$$\begin{aligned} \mathrm{E}(I_2) &= \frac{1}{2}a_n^2\mathrm{E}\left\{K_i\rho''(\epsilon_i + R(u_i, \mathbf{x}_i))(\mu^T\mathbf{x}_i^*)^2\right\} \\ &= \frac{1}{2}a_n^2\mathrm{E}\left\{\rho''(\epsilon_i)K_i(\mu^T\mathbf{x}_i^*)^2\right\}(1 + o(1)) \\ &= \frac{1}{2}a_n^2\delta_1\mathrm{E}\left\{\mathrm{E}\left\{\mu^T\mathbf{x}_i^*\mathbf{x}_i^{*T}\mu|u_i\right\}K_i\right\}(1 + o(1)) \\ &= \frac{1}{2}a_n^2\delta_1\mu^T\mathrm{E}\left\{\mathrm{E}\left\{\mathbf{x}_i^*\mathbf{x}_i^{*T}|u_i\right\}K_i\right\}\mu(1 + o(1)). \end{aligned}$$

Note that $\mathbf{x}_i^*\mathbf{x}_i^{*T} = \left(x_{ij}x_{ik}\left(\frac{u_i - u_0}{h}\right)^l\right)_{1\le j,k\le d, l=0,1,2}$ and $\Gamma_{jk}(u_i) = \mathrm{E}(x_{ij}x_{ik}|u_i)$ for $1 \le j, k \le d$, then

$$\begin{aligned} \mathrm{E}\left\{\mathrm{E}\left\{x_{ij}x_{ik}\left(\frac{u_i - u_0}{h}\right)^l\middle|u_i\right\}K_i\right\} &= \mathrm{E}\left\{\mathrm{E}(x_{ij}x_{ik}|u_i)\left(\frac{u_i - u_0}{h}\right)^lK_i\right\} \\ &= \mathrm{E}\left\{\Gamma_{jk}(u_i)\left(\frac{u_i - u_0}{h}\right)^lK_i\right\}. \end{aligned}$$

By using Taylor expansion, we obtain

$$\begin{aligned} \mathrm{E}\left\{\mathrm{E}\left\{x_{ij}x_{ik}\left(\frac{u_i - u_0}{h}\right)^l\middle|u_i\right\}K_i\right\} &= \frac{1}{h}\int\Gamma_{jk}(u_i)\left(\frac{u_i - u_0}{h}\right)^lK\left(\frac{u_i - u_0}{h}\right)q(u_i)\mathrm{d}u_i \\ &= q(u_0)\Gamma_{jk}(u_0)\int t^lK(t)\mathrm{d}t(1 + o(1)). \end{aligned}$$

So we have

$$\mathrm{E}(I_2) = \frac{1}{2}a_n^2\delta_1 q(u_0)\mu^T\mathbf{S}\mu(1+o(1)),$$

where $\mathbf{S} = \begin{pmatrix}1 & 0\\ 0 & \mu_2\end{pmatrix} \otimes \mathbf{\Gamma}(u_0)$ is a $2d \times 2d$ matrix. Thus,

$$\mathrm{E}(I_2) = O(a_n^2\delta_1 q(u_0)\mu^T\mathbf{S}\mu)$$

and

$$\begin{aligned}\mathrm{var}(I_2) &= \frac{a_n^4}{4n}\mathrm{var}\left[\rho''(\epsilon_i + R(u_i,\mathbf{x}_i))K_i(\mu^T\mathbf{x}_i^*)^2\right]\\ &= \frac{a_n^4}{4n}\left\{\mathrm{E}(B^2) - [\mathrm{E}(B)]^2\right\},\end{aligned}$$

where $B = \rho''(\epsilon_i + R(u_i,\mathbf{x}_i))K_i(\mu^T\mathbf{x}_i^*)^2$. Let $\delta_3 = \mathrm{E}\left(\rho''(\epsilon_i)^2\right)$, then

$$\begin{aligned}\mathrm{E}(B^2) &= \mathrm{E}\left\{\rho''(\epsilon_i + R(u_i,\mathbf{x}_i))^2K_i^2(\mu^T\mathbf{x}_i^*)^4\right\}\\ &\approx \mathrm{E}\left\{\rho''(\epsilon_i)^2K_i^2(\mu^T\mathbf{x}_i^*)^4\right\}\\ &= \delta_3\mathrm{E}\left\{K_i^2(\mu^T\mathbf{x}_i^*)^4\right\}\\ &= O\left(\frac{1}{h}\right).\end{aligned}$$

Note that $[\mathrm{E}(B)]^2 = [O(1)]^2 = O(1) \ll \mathrm{E}(B^2)$, so $\mathrm{var}(I_2) = O\left(\frac{a_n^4}{nh}\right)$. Based on the result $I_2 = \mathrm{E}(I_2) + O_p(\sqrt{\mathrm{var}(I_2)})$ and the assumption $nh \to \infty$, it follows that

$$I_2 = a_n^2\delta_1 q(u_0)\mu^T\mathbf{S}\mu(1+o_p(1)).$$

Similarly, $I_3 = -\frac{1}{6n}\sum_{i=1}^n K_i\rho'''(z_i)a_n^3(\mu^T\mathbf{x}_i^*)^3 = O_p(a_n^3)$.

Assume $\delta_1 < 0$. Noticing that $\mathbf{S}$ is a positive matrix, $\|\mu\| = c$, we can choose c large enough such that I_2 dominates both I_1 and I_3 with probability at least $1-\eta$. Thus $P\left\{\sup_{\|\mu\|=c} L(\boldsymbol{\theta}^* + a_n\mu) < L(\boldsymbol{\theta}^*)\right\} \geq 1-\eta$. Hence with probability approaching 1, there exists a local maximizer $\hat{\boldsymbol{\theta}}^*$ such that $\left\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\right\| \leq a_n c$, where $a_n = (nh)^{-1/2} + h^2$. Based on the definition of $\boldsymbol{\theta}^*$, we can get, with probability approaching 1, $H(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_p((nh)^{-1/2} + h^2)$. □

Proof of Theorem 2.2. Now we provide the asymptotic distribution for such consistent estimate. Since $\hat{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$, then $L'(\hat{\boldsymbol{\theta}}) = 0$. By Taylor expansion,

$$0 = L'(\hat{\boldsymbol{\theta}}) = L'(\boldsymbol{\theta}_0) + L''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \frac{1}{2}L'''(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^2,$$

where $\tilde{\boldsymbol{\theta}}$ is a value between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Then $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = -[L''(\boldsymbol{\theta}_0)]^{-1}L'(\boldsymbol{\theta}_0)(1+o_p(1))$. Since $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n K_i\rho(y_i - \boldsymbol{\theta}^{*T}\mathbf{x}_i^*)$ and $y_i - \boldsymbol{\theta}^{*T}\mathbf{x}_i^* = \epsilon_i + R(u_i,\mathbf{x}_i)$, then $L''(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n K_i\rho''(\epsilon_i + R(u_i,\mathbf{x}_i))\mathbf{x}_i^*\mathbf{x}_i^{*T}$. We have the following expectation,

$$\begin{aligned}\mathrm{E}[L''(\boldsymbol{\theta}^*)] &= \mathrm{E}\left\{\rho''(\epsilon_i + R(u_i,\mathbf{x}_i))K_i\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\}\\ &\approx \mathrm{E}\left\{\rho''(\epsilon_i)K_i\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\}\\ &= \delta_1\mathrm{E}\left\{\mathrm{E}\left\{\mathbf{x}_i^*\mathbf{x}_i^{*T}|u_i\right\}K_i\right\}\\ &= \delta_1 q(u_0)\mathbf{S}(1+o(1)).\end{aligned}$$

Throughout this article, we consider the element-wise variance of a matrix,

$$\mathrm{var}[L''(\boldsymbol{\theta}^*)] = \frac{1}{n}\mathrm{var}\left\{K_i\rho''(\epsilon_i + R(u_i,\mathbf{x}_i))\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\} = O\left(\frac{1}{nh}\right).$$

Based on the result $L''(\boldsymbol{\theta}^*) = \mathrm{E}[L''(\boldsymbol{\theta}^*)] + O_p(\sqrt{\mathrm{var}[L''(\boldsymbol{\theta}^*)]})$ and the assumption $nh \to \infty$, it follows that

$$L''(\boldsymbol{\theta}^*) = \delta_1 q(u_0)\mathbf{S}(1+o_p(1)).$$

For $L'(\boldsymbol{\theta}^*)$, we can divide it into two parts.

$$L'(\boldsymbol{\theta}^*) = -\frac{1}{n}\sum_{i=1}^n K_i\rho'(\epsilon_i + R(u_i,\mathbf{x}_i))\mathbf{x}_i^*$$

$$\approx -\frac{1}{n}\sum_{i=1}^{n} K_i\rho'(\epsilon_i)\mathbf{x}_i^* - \frac{1}{n}\sum_{i=1}^{n} K_i\rho''(\epsilon_i)R(u_i,\mathbf{x}_i)\mathbf{x}_i^*$$
$$\triangleq -\mathbf{w}_n - \boldsymbol{\nu}_n.$$

The asymptotic result is determined by $\mathbf{w}_n$. In order to find the order of $\boldsymbol{\nu}_n$, we compute the following things.

$$\mathrm{E}(\boldsymbol{\nu}_n) = \mathrm{E}\left[K_i\rho''(\epsilon_i)R(u_i,\mathbf{x}_i)\mathbf{x}_i^*\right] = \delta_1\mathrm{E}\left\{\mathrm{E}\left\{R(u_i,\mathbf{x}_i)\mathbf{x}_i^*|u_i\right\}K_i\right\}.$$

Since $g_j'''(\cdot)$ is bounded, then we have

$$R(u_i,\mathbf{x}_i) = \sum_{j=1}^{d}\left\{\sum_{m=2}^{\infty}\frac{1}{m!}g_j^{(m)}(u_0)(u_i-u_0)^m\right\}x_{ij} = \sum_{j=1}^{d}\frac{1}{2}g_j''(u_0)(u_i-u_0)^2x_{ij}(1+o_p(1)).$$

By $\mathbf{x}_i^* = (x_{i1},\ldots,x_{id},(\frac{u_i-u_0}{h})x_{i1},\ldots,(\frac{u_i-u_0}{h})x_{id})^T$,

$$R(u_i,\mathbf{x}_i)\mathbf{x}_i^* \approx \left[\left(\frac{(u_i-u_0)^2}{2}\left\{\sum_{j=1}^{d}g_j''(u_0)x_{ij}\right\}x_{ik}\right)_{1\le k\le d}, \left(\frac{(u_i-u_0)^3}{2h}\left\{\sum_{j=1}^{d}g_j''(u_0)x_{ij}\right\}x_{ik}\right)_{1\le k\le d}\right]^T_{2d\times 1}.$$

Since

$$\begin{aligned}
\mathrm{E}\left\{\mathrm{E}\left\{\left[\sum_{j=1}^{d}g_j''(u_0)x_{ij}\right]x_{ik}|u_i\right\}\frac{(u_i-u_0)^2}{2}K_i\right\} &= \mathrm{E}\left\{\sum_{j=1}^{d}g_j''(u_0)\mathrm{E}(x_{ij}x_{ik}|u_i)\frac{(u_i-u_0)^2}{2}K_i\right\}\\
&= \mathrm{E}\left\{\sum_{j=1}^{d}g_j''(u_0)\Gamma_{jk}(u_i)\frac{(u_i-u_0)^2}{2}K_i\right\}\\
&= \sum_{j=1}^{d}g_j''(u_0)\mathrm{E}\left\{\Gamma_{jk}(u_i)\frac{(u_i-u_0)^2}{2}K_i\right\}\\
&= \sum_{j=1}^{d}g_j''(u_0)\frac{1}{h}\int\Gamma_{jk}(u_i)\frac{(u_i-u_0)^2}{2}K\left(\frac{u_i-u_0}{h}\right)q(u_i)\mathrm{d}u_i\\
&= \frac{h^2}{2}q(u_0)\sum_{j=1}^{d}g_j''(u_0)\Gamma_{jk}(u_0)\int t^2K(t)\mathrm{d}t(1+o(1))
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left\{\mathrm{E}\left\{\left[\sum_{j=1}^{d}g_j''(u_0)x_{ij}\right]x_{ik}|u_i\right\}\frac{(u_i-u_0)^3}{2h}K_i\right\} &= \mathrm{E}\left\{\sum_{j=1}^{d}g_j''(u_0)\Gamma_{jk}(u_i)\frac{(u_i-u_0)^3}{2h}K_i\right\}\\
&= \sum_{j=1}^{d}g_j''(u_0)\frac{1}{2h}\mathrm{E}\left\{\Gamma_{jk}(u_i)(u_i-u_0)^3K_i\right\}\\
&= \frac{h^2}{2}q(u_0)\sum_{j=1}^{d}g_j''(u_0)\Gamma_{jk}(u_0)\int t^3K(t)\mathrm{d}t(1+o(1)),
\end{aligned}$$

then

$$\mathrm{E}(\boldsymbol{\nu}_n) = \delta_1 q(u_0)\frac{h^2}{2}\sum_{j=1}^{d}g_j''(u_0)\boldsymbol{\psi}_j(1+o(1)),$$

where $\boldsymbol{\psi}_j = \binom{\mu_2}{\mu_3}\otimes\left(\Gamma_{jk}(u_0)\right)_{1\le k\le d}^T$ is a $2d\times 1$ vector for $j=1,\ldots,d$. Since $\mathrm{var}(\boldsymbol{\nu}_n) = \mathrm{var}\left\{K_i\rho''(\epsilon_i)R(u_i,\mathbf{x}_i)\mathbf{x}_i^*\right\}/n = O(h^3/n)$, then based on the result $\boldsymbol{\nu}_n = \mathrm{E}(\boldsymbol{\nu}_n) + O_p(\sqrt{\mathrm{var}(\boldsymbol{\nu}_n)})$ and the assumption $nh\to\infty$, it follows that

$$\boldsymbol{\nu}_n = \delta_1 q(u_0)\frac{h^2}{2}\sum_{j=1}^{d}g_j''(u_0)\boldsymbol{\psi}_j(1+o_p(1)).$$

Then

$$\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^* = -[L''(\boldsymbol{\theta}^*)]^{-1}L'(\boldsymbol{\theta}^*)(1+o_p(1))$$

$$= -\left[\delta_1 q(u_0)\mathbf{S}\right]^{-1}(-\mathbf{w}_n - \boldsymbol{\nu}_n)(1+o_p(1))$$
$$= \frac{\mathbf{S}^{-1}\mathbf{w}_n}{\delta_1 q(u_0)}(1+o_p(1)) + \mathbf{S}^{-1}\frac{h^2}{2}\sum_{j=1}^{d} g_j''(u_0)\boldsymbol{\psi}_j(1+o_p(1)). \tag{4.2}$$

Based on the assumption $\mathrm{E}[\rho'(\epsilon_i)] = 0$, we can easily get $\mathrm{E}(\mathbf{w}_n) = 0$.

$$\mathrm{var}(\mathbf{w}_n) = \frac{1}{n}\mathrm{var}\left\{K_i\rho'(\epsilon_i)\mathbf{x}_i^*\right\} = \frac{1}{n}\mathrm{E}\left\{K_i^2\rho'(\epsilon_i)^2\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\} = \frac{1}{n}\delta_2\mathrm{E}\left\{\mathrm{E}\left\{\mathbf{x}_i^*\mathbf{x}_i^{*T}|u_i\right\}K_i^2\right\}.$$

Since $\mathbf{x}_i^*\mathbf{x}_i^{*T} = \left(x_{ij}x_{ik}\left(\frac{u_i-u_0}{h}\right)^l\right)_{1\le j,k\le d, l=0,1,2}$ and

$$\mathrm{E}\left\{\mathrm{E}\left\{x_{ij}x_{ik}\left(\frac{u_i-u_0}{h}\right)^l\middle|u_i\right\}K_i^2\right\} = \mathrm{E}\left\{\mathrm{E}\left\{x_{ij}x_{ik}|u_i\right\}\left(\frac{u_i-u_0}{h}\right)^l K_i^2\right\}$$
$$= \mathrm{E}\left\{\Gamma_{jk}(u_i)\left(\frac{u_i-u_0}{h}\right)^l K_i^2\right\}$$
$$= \frac{1}{h}q(u_0)\Gamma_{jk}(u_0)\int t^l K^2(t)\mathrm{d}t(1+o(1)),$$

then

$$\mathrm{E}\left\{\mathrm{E}\left\{\mathbf{x}_i^*\mathbf{x}_i^{*T}|u_i\right\}K_i^2\right\} = \frac{1}{h}q(u_0)\boldsymbol{\Lambda}(1+o(1)),$$

where $\boldsymbol{\Lambda} = \begin{pmatrix}\nu_0 & \nu_1\\ \nu_1 & \nu_2\end{pmatrix}\otimes\boldsymbol{\Gamma}(u_0)$ is a $2d\times 2d$ matrix. So $\mathrm{var}(\mathbf{w}_n) = \frac{1}{nh}\delta_2 q(u_0)\boldsymbol{\Lambda}(1+o(1))$. We next use the Lyapunov central limit theorem to obtain the asymptotic distribution of $\mathbf{w}_n$. The Lyapunov conditions are checked as follows. For any unit vector $\mathbf{d}\in\mathbb{R}^{2d}$, let $\mathbf{d}^T\mathbf{w}_n = \sum_{i=1}^n \xi_i$, where $\xi_i = \frac{1}{n}K_i\rho'(\epsilon_i)\mathbf{d}^T\mathbf{x}_i^*$. Since

$$\mathrm{E}(\xi_i^2) = \mathrm{E}\left\{\frac{1}{n^2}K_i^2\rho'(\epsilon_i)^2\mathbf{d}^T\mathbf{x}_i^*\mathbf{x}_i^{*T}\mathbf{d}\right\} = \frac{1}{n^2}\delta_2\mathbf{d}^T\mathrm{E}\left\{K_i^2\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\}\mathbf{d} = \frac{1}{n^2 h}\delta_2 q(u_0)\mathbf{d}^T\boldsymbol{\Lambda}\mathbf{d}(1+o(1)),$$

then $o\left(\left(\sum_{i=1}^n \mathrm{E}\,|\xi_i|^2\right)^3\right) = o\left(\left(\frac{1}{nh}\right)^3\right)$. Let $\delta_4 = \mathrm{E}\left\{\rho'(\epsilon_i)^3\right\}$, then

$$\mathrm{E}(\xi_i^3) = \mathrm{E}\left\{\frac{1}{n^3}K_i^3\rho'(\epsilon_i)^3(\mathbf{d}^T\mathbf{x}_i^*)^3\right\} = \frac{1}{n^3}\delta_3\mathrm{E}\left\{K_i^3(\mathbf{d}^T\mathbf{x}_i^*)^3\right\} = O\left(\frac{1}{n^3h^2}\right).$$

So $\left(\sum_{i=1}^n \mathrm{E}\,|\xi_i|^3\right)^2 = O\left(\left(\frac{1}{n^2h^2}\right)^2\right)$. Since $\left(\frac{1}{n^2h^2}\right)^2(nh)^3 = \frac{1}{nh}\to 0$, then $\left(\frac{1}{n^2h^2}\right)^2 = o\left(\left(\frac{1}{nh}\right)^3\right)$, which is equivalent to $\left(\sum_{i=1}^n \mathrm{E}\,|\xi_i|^3\right)^2 = o\left(\left(\sum_{i=1}^n \mathrm{E}\,|\xi_i|^2\right)^3\right)$. Based on Lyapunov Central Limit Theorem,

$$\frac{\mathbf{w}_n}{\sqrt{\mathrm{var}(\mathbf{w}_n)}} \xrightarrow{D} N(\mathbf{0}_{2d}, \mathbf{I}_{2d}),$$

where $\mathbf{0}_{2d}$ is a $2d\times 1$ vector with each entry being 0; $\mathbf{I}_{2d}$ is a $2d\times 2d$ identity matrix. Previously, we already computed that $\mathrm{var}(\mathbf{w}_n) = \frac{1}{nh}\delta_2 q(u_0)\boldsymbol{\Lambda}(1+o(1))$, by Slutsky's Theorem,

$$\sqrt{nh}\mathbf{w}_n \xrightarrow{D} N(\mathbf{0}_{2d}, \delta_2 q(u_0)\boldsymbol{\Lambda}).$$

Based on (4.2), we have the following result

$$\sqrt{nh}\left\{\mathbf{H}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) - \mathbf{S}^{-1}\frac{h^2}{2}\sum_{j=1}^{d} g_j''(u_0)\boldsymbol{\psi}_j(1+o_p(1))\right\} \xrightarrow{D} N(\mathbf{0}_{2d}, \delta_1^{-2}\delta_2 q(u_0)^{-1}\mathbf{S}^{-1}\boldsymbol{\Lambda}\mathbf{S}^{-1}).$$

References

[1] R. Beran, Asymptotic efficient adaptive rank estimates in location models, Ann. Statist. 2 (1974) 63–74.
[2] P.J. Bickel, On adaptive estimation, Ann. Statist. 10 (1982) 647–671.
[3] Z. Cai, J. Fan, R. Li, Efficient estimation and inferences for varying-coefficient models, J. Amer. Statist. Assoc. 95 (2000) 888–902.
[4] C.-T. Chiang, J.A. Rice, C.O. Wu, Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variable, J. Amer. Statist. Assoc. 96 (2001) 605–619.
[5] W.S. Cleveland, E. Grosse, W.M. Shyu, Local regression models, in: J.M. Chambers, T.J. Hastie (Eds.), Statistical Models in S, Wadsworth/Brooks-Cole, Pacific Grove, CA, 1991, pp. 309–376.

[6] F.C. Drost, C.A.J. Klaassen, Efficient estimation in semiparametric GRACH models, J. Econometrics 81 (1997) 193–221.
[7] R.L. Eubank, C.F. Huang, Y.M. Maldonado, N. Wang, S. Wang, R.J. Buchanan, Smoothing spline estimation in varying-coefficient models, J. R. Stat. Soc. Ser. B 66 (2004) 653–667.
[8] J. Fan, Local linear regression smoothers and their minimax efficiencies, Ann. Statist. 21 (1993) 196–216.
[9] J. Fan, M. Farmen, I. Gijbels, Local maximum likelihood estimation and inference, J. R. Stat. Soc. Ser. B 60 (1998) 591–608.
[10] J. Fan, I. Gijbels, Variable bandwidth and local linear regression smoothers, Ann. Statist. 20 (1992) 2008–2036.
[11] J. Fan, T. Huang, Profile likelihood inferences on semiparametric varying-coefficient partially linear models, Bernoulli 11 (2005) 1031–1057.
[12] J. Fan, W. Zhang, Statistical estimation in varying coefficient models, Ann. Statist. 27 (1999) 1491–1518.
[13] J. Fan, J.T. Zhang, Two-step estimation of functional linear models with applications to longitudinal data, J. R. Stat. Soc.Ser. B 62 (2000) 303–322.
[14] J. Fan, W. Zhang, Statistical methods with varying coefficient models, Stat. Interface 1 (1) (2008) 179–195.
[15] J. Fan, C. Zhang, J. Zhang, Generalized likelihood ratio statistics and Wilks phenomenon, Ann. Statist. 29 (2001) 153–193.
[16] O.W. Gilley, R.K. Pace, On the Harrison and Rubinfeld data, J. Environ. Econ. Manag. 31 (1996) 403–405.
[17] T.J. Hastie, R.J. Tibshirani, Varying-coefficient models, J. R. Stat. Soc.Ser. B 55 (1993) 757–796.
[18] D.J. Hodgson, Adaptive estimation of cointegrating regressions with ARMA errors, J. Econometrics 85 (1998) 231–267.
[19] D.R. Hoover, J.A. Rice, C.O. Wu, L.P. Yang, Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, Biometrika 85 (1998) 809–822.
[20] J.Z. Huang, H. Shen, Functional coefficient regression models for non-linear time series: a polynomial spline approach, Scand. J. Statist. 31 (2004) 515–534.
[21] J.Z. Huang, C.O. Wu, L. Zhou, Varying-coefficient models and basis function approximations for the analysis of repeated measurements, Biometrika 89 (2002) 111–128.
[22] J.Z. Huang, C.O. Wu, L. Zhou, Polynomial spline estimation and inference for varying coefficient models with longitudinal data, Statist. Sinica 14 (2004) 763–788.
[23] G. Kauermann, G. Tutz, On model diagnostics using varying coefficient models, Biometrika 86 (1999) 119–128.
[24] O.B. Linton, Z. Xiao, A nonparametric regression estimator that adapts to error distribution of unknown form, Econometric Theory 23 (2007) 371–413.
[25] C.F. Manski, Adaptive estimation of non-linear regression models, Econometric Rev. 3 (1984) 145–194.
[26] A. Schick, On efficient estimation in regression models, Ann. Statist. 21 (1993) 1486–1521.
[27] J.G. Staniswalis, The kernel estimate of a regression function in likelihood based models, J. Amer. Statist. Assoc. 84 (1989) 276–283.
[28] D.G. Steigerwald, Adaptive estimation in time series regression models, J. Econometrics 54 (1992) 251–275.
[29] C.J. Stone, Adaptive maximum likelihood estimators of a location parameter, Ann. Statist. 3 (1975) 267–284.
[30] Y. Sun, H. Yan, W. Zhang, Z. Lu, A semiparametric spatial dynamic model, Ann. Statist. 42 (2014) 700–727.
[31] L. Wang, B. Kai, R. Li, Local rank inference for varying coefficient models, J. Amer. Statist. Assoc. 104 (2009) 1631–1645.
[32] H. Wang, Y. Xia, Shrinkage estimation of the varying coefficient model, J. Amer. Statist. Assoc. 104 (2009) 747–757.
[33] Q. Wang, W. Yao, An adaptive estimation of MAVE, J. Multivariate Anal. 104 (2012) 88–100.
[34] C.O. Wu, C.-T. Chiang, D.R. Hoover, Asymptotoc confidence regions for kernel smoothing of a varying coefficient model with longitudinal data, J. Amer. Statist. Assoc. 93 (1998) 1388–1402.
[35] Y. Xia, H. Tong, W. Li, L. Zhu, An adaptive estimation of dimension reduction space, J. R. Stat. Soc. Ser. B 64 (2002) 363–388.
[36] W. Yao, A note on EM algorithm for mixture models, Statist. Probab. Lett. 83 (2013) 519–526.
[37] W. Yao, Z. Zhao, Kernel density based linear regression estimates, Comm. Statist. Theory Methods 42 (2013) 4499–4512.
[38] A. Yuan, Semiparametric inference with kernel likelihood, J. Nonparametr. Stat. 21 (2009) 207–228.
[39] A. Yuan, J.G. De Gooijer, Semiparametric regression with kernel error model, Scand. J. Statist. 34 (2007) 841–869.
[40] W. Zhang, S.Y. Lee, Variable bandwidth selection in varying-coefficient models, J. Multivariate Anal. 74 (2000) 116–134.