Econometric Reviews

ISSN: 0747-4938 (Print) 1532-4168 (Online) Journal homepage: https://www.tandfonline.com/loi/lecr20

Multivariate Local Polynomial Kernel Estimators: Leading Bias and Asymptotic Distribution

Jingping Gu , Qi Li & Jui-Chung Yang

To cite this article: Jingping Gu , Qi Li & Jui-Chung Yang (2015) Multivariate Local Polynomial Kernel Estimators: Leading Bias and Asymptotic Distribution, Econometric Reviews, 34:6-10, 979-1010, DOI: 10.1080/07474938.2014.956615

To link to this article: https://doi.org/10.1080/07474938.2014.956615

Accepted author version posted online: 05 Sep 2014.
Published online: 05 Sep 2014.

Econometric Reviews, 34(6–10):979–1010, 2015

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2014.956615

Taylor & Francis
Taylor & Francis Group

Multivariate Local Polynomial Kernel Estimators: Leading Bias and Asymptotic Distribution

Jingping Gu[1], Qi Li[2,3], and Jui-Chung Yang[2]

[1]*Department of Economics, University of Arkansas, Fayetteville, Arkansas, USA*
[2]*Department of Economics, Texas A&M University, College Station, Texas, USA*
[3]*International School of Economics and Management, Capital University of Economics and Business, Beijing, China*

Masry (1996b) provides estimation bias and variance expression for a general local polynomial kernel estimator in a general multivariate regression framework. Under smoother conditions on the unknown regression function and by including more refined approximation terms than that in Masry (1996b), we extend the result of Masry (1996b) to obtain explicit leading bias terms for the whole vector of the local polynomial estimator. Specifically, we derive the leading bias and leading variance terms of nonparametric local polynomial kernel estimator in a general nonparametric multivariate regression model framework. The results can be used to obtain optimal smoothing parameters in local polynomial estimation of the unknown conditional mean function and its derivative functions.

Keywords Kernel estimation; Leading bias; Local polynomial method.

JEL Classification C14.

1. INTRODUCTION

Nonparametric kernel estimation method is the most popular nonparametric estimation method. Recently, nonparametric local polynomial kernel estimation method has attracted much attention among statisticians and econometricians. Local polynomial method has the advantage that it can simultaneously estimate a nonparametric regression function and its derivative functions. Moreover, local polynomial method has better estimation accuracy at the boundary region of the data support, and minimax efficiency over the local constant estimator, see Fan and Gijbels (1996). Deriving the local polynomial estimation leading bias and variance expression enables one to select optimal smoothing parameters that balance the estimation squared bias and variance, say, by using some plug-in methods. It also readily gives the asymptotic normal

Address correspondence to Qi Li, Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA; E-mail: qi@econmail.tamu.edu

distribution of the local polynomial estimator while allowing for optimally selected smoothing parameters. Fan et al. (1996) derived leading bias and variance terms of a local polynomial estimator for the case of a univariate nonparametric regression model. Ruppert and Wand (1994) analyzed conditional estimation mean squared error (conditional on the covariate X) and discussed an iterative procedure that lead to optimal selection of smoothing parameters. Masry (1996a,b) considered general multivariate local polynomial estimator and derived the rate of convergence and the asymptotic normal distribution of the estimator with time series data. More recently, local polynomial method has been used to construct specification tests and to estimate various semiparametric/nonparametric statistical models. For example, Liu et al. (2000) used local polynomial method to construct consistent model specification tests. Li et al. (2003) suggested using local polynomial method to estimate a single index model (via estimation of average derivatives). Su and Ullah (2008) proposed to use local polynomial method to estimate a system of simultaneous equation models. Xiao (2009) used local polynomial method to estimate a semiparametric cointegration model.

In this article we strength some of the regularity conditions used in Masry (1996b) and derive the leading estimation bias term for all the components of the vector local polynomial estimator. The explicit expression of the leading bias makes optimal selection of smoothing parameters quite convenient. It also enables one to examine the conditions under which the optimal smoothing parameters are (asymptotically) unique and to carry out inferences with optimally selected smoothing parameters.

The leading bias plays a central part in nonparametric estimation because optimal smoothing requires one to select the smoothing parameters to balance the estimation leading squared bias and variance. In this article, we derive a general expression for the leading bias of each component in the multivariate local polynomial estimator (LPE) with general order p. We provide detailed analysis for the multivariate local linear estimator (LLE, $p = 1$).

The leading bias and asymptotic distribution of the univariate LPE have been studied by Fan and Gijbels (1996) and Fan et al. (1996). For the multivariate case (d-dimemsional), Ruppert and Wand (1994) discussed the conditional (conditional on the covariate X) leading bias and variance for the estimate of the conditional mean function $m(\cdot)$ for the multivariate LLE; and considered the general local polynomial estimator for the univariate x case. Masry (1996a,b) gave the expression of the leading bias for the vector of the multivariate LPE, i.e., Masry derived the leading bias term for the vector $D_{h,p}m(\cdot) \stackrel{def}{=} \{m(\cdot), hDm(\cdot), h^2D^2m(\cdot), \ldots, h^pD^pm(\cdot)\}$, where $D^jm(\cdot)$ is the jth-order partial derivative function of $m(\cdot)$. Since the different components of $D_{h,p}m(\cdot)$ may have different order of bias. Masry (1996b) did not give explicit leading bias terms to all the components of $D_{h,p}m(\cdot)$. In this article, we strengthen some smoothness conditions used in Masry (1996b) and derive explicit leading bias for all components of $D_{h,p}m(\cdot)$. Specifically, we assume that the unknown function $D_{h,p}m(\cdot)$ has continuous derivatives of total order $p + 2$, and we derive the expression of the leading bias for each element of the multivariate

LPE. For the univariate case, our finding is consistent with Fan et al. (1996). The result of this article allows one to derive asymptotic distribution for the LPE with optimally selected bandwidth. It also provides the formula that can be used to examine the conditions under which the optimally selected bandwidth is unique (e.g., Li and Zhou, 2005).

2. LEADING BIAS FOR UNIVARIATE LOCAL LINEAR ESTIMATOR: AN ILLUSTRATIVE EXAMPLE

2.1. The Univariate Local Linear Estimator: When x is An Interior Point

In this subsection, we review the univariate LLE case and use it as an illustrative example to show the differences of leading bias terms derived in Fan et al. (1996) (under additional smoothness of the regression function) and that derived in Masry (1996b) with weaker regularity conditions. The leading bias of univariate LLE had been studied by Fan et al. (1996), and the multivariate case had been discussed by Masry (1996b). We compare the differences of the leading bias terms derived by Fan et al. and by Masry in a univariate setting. First, we formally define the leading bias term of a generic scalar nonparametric estimator. For a given $x \in \mathbb{R}$, let $\hat{\theta}(x)$ be an estimator of a nonparametrically specified function $\theta(x)$, if $\hat{\theta}(x)$ admits the following expansion:

$$\hat{\theta}(x) - \theta(x) = a_n B_n(x) + b_n C_n(x) + (s.o.), \tag{2.1}$$

where $B_n(x) \neq 0$ is finite and non-stochastic, $C_n(x)$ has zero mean and finite asymptotic variance, a_n and b_n are nonstochastic positive sequences such that $a_n \to 0$ and $b_n \to 0$ as $n \to \infty$, and $(s.o.)$ denote terms that have probability order smaller than a_n or b_n (i.e., $o_p(a_n \vee b_n)$). Then we call $a_n B_n(x)$ as the leading bias term of $\hat{\theta}(x)$.[1] In our applications, a_n and b_n depend on smoothing parameters. When a_n and b_n have the same order, we say that the smoothing parameter has an optimal order. In this case, we also say that $a_n^2 B_n(x)^2 + b_n^2 E[C_n(x)^2]$ is the leading estimation mean squared error (for estimating $\theta(x)$).

Consider a univariate nonparametric regression model

$$Y_i = m(X_i) + u_i = m(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \ldots, n, \tag{2.2}$$

where Y_i and X_i are two real-valued random variables. $\{Y_i, X_i\}$ are independent and identically distributed (i.i.d.) as $\{Y, X\}$, X has a density $f(\cdot)$ with support $\text{supp}(f) \in \mathbb{R}$. $\sigma^2(x) = \text{var}(Y \mid X = x)$ is finite. $\{\varepsilon_i\}$ are i.i.d. with zero mean and unit variance, and are independent of $\{X_i\}$. $m(\cdot)$ has continuous derivatives of total order 3 for all $x \in \text{supp}(f)$.

[1] Since we only consider a nonparametric functional estimate, all estimators considered in this article have nonzero (leading) biases (in finite samples), although these estimators are asymptotically unbiased.

By approximating the unknown function $m(\cdot)$ with its first order Taylor expansion, we have

$$m(X_i) \approx m(x) + m'(x)(X_i - x), \tag{2.3}$$

where $m'(x) = dm(x)/dx$. The LLE, which estimates $m(\cdot)$ and $m'(\cdot)$ simultaneously, is the minimizer of the following objective function (a weighted least squares estimator):

$$\begin{pmatrix} \hat{b}_0(x) \\ \hat{b}_1(x) \end{pmatrix} = \underset{b_0(x), b_1(x)}{\arg\min} \sum_{i=1}^{n} \{Y_i - b_0(x) - b_1(x)(X_i - x)\}^2 \frac{1}{h} K\left(\frac{X_i - x}{h}\right), \tag{2.4}$$

where $K(\cdot)$ is the kernel function, and h is the bandwidth parameter, $K(\cdot)$ is a nonnegative, symmetric, and bounded function with $\int K(v)dv = 1$ and $\int K(v)v^4 dv < \infty$. Also, we will use the notation $\int_{\mathbb{R}} K(v)v^q dv = \mu_q$ and $\int_{\mathbb{R}} K^2(v)v^r dv = \lambda_r$. Note that $\mu_0 = 1$ and $\mu_j = \lambda_j = 0$ for odd integer j. We require that $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

Masry (1996b) defined his local polynomial estimators as the estimate of $\{m(\cdot), hm'(\cdot)\}^\top \equiv \{m(\cdot), hdm(\cdot)/dx\}^\top$, while Fan et al. (1996) defined their LPE as the estimate of $\{m(\cdot), m'(\cdot)\}^\top \equiv \{m(\cdot), dm(\cdot)/dx\}^\top$. We will use the notation $\hat{\beta}_j$ to denote the estimator of Masry (1996), and $\hat{b}_j$ to denote the estimator of Fan et al. (1996). Then we have

$$\begin{aligned} b_0(x) &= \beta_0(x) = m(x), \quad \hat{b}_0(x) = \hat{\beta}_0(x) = \hat{m}(x), \\ \beta(x) &= (\beta_0(x), \beta_1(x))^\top, \quad \widehat{\beta}(x) = (\hat{\beta}_0(x), \hat{\beta}_1(x))^\top, \\ b_1(x) &= h^{-1}\beta_1(x) = m'(x), \quad \hat{b}_1(x) = h^{-1}\hat{\beta}_1(x) = \hat{m}'(x), \\ \mathbf{b}(x) &= (b_0(x), b_1(x))^\top, \quad \hat{\mathbf{b}}(x) = (\hat{b}_0(x), \hat{b}_1(x))^\top. \end{aligned} \tag{2.5}$$

Fan et al. (1996) derived the following leading bias of the local linear estimator

$$Bias_{L,F}(\hat{\mathbf{b}}(x)) \equiv Bias_{L,F}(\hat{\mathbf{b}}(x) \mid \mathscr{X}) = \begin{bmatrix} \frac{1}{2}h^2\mu_2 \frac{d^2m(x)}{dx^2} \\ \frac{1}{2}h^2 \frac{\mu_4 - \mu_2^2}{\mu_2} \frac{f'(x)}{f(x)} \frac{d^2m(x)}{dx^2} + \frac{1}{6}h^2 \frac{\mu_4}{\mu_2} \frac{d^3m(x)}{dx^3} \end{bmatrix}, \tag{2.6}$$

where the subscript L means "leading" and subscript F denotes that the bias is derived based on Fan et al. (1996).[2]

[2]Note that we have written $Bias_{L,F}\left(\hat{\mathbf{b}}(x)|\mathscr{X}\right)$ as $Bias_{L,F}\left(\hat{\mathbf{b}}(x)\right)$ in (2.6) because the right-hand side of (2.6) does not depend on $\mathscr{X}$; we often omit the conditional variable $\mathscr{X}$ in (2.6) for notational simplicity. This notation is also consistent with Eq. (2.1) when describing leading bias and variance of an estimator.

The leading bias of the local linear estimator reported in Masry (1996b) is given by

$$Bias_{L,M}(\widehat{\beta}(x)) = \begin{bmatrix} \frac{1}{2}h^2\mu_2\frac{d^2m(x)}{dx^2} \\ o(h^2) \end{bmatrix}, \tag{2.7}$$

where the subscript M denotes bias derived under (weak) conditions imposed by Masry (1996b).

Since $(b_0(x), b_1(x))^\top = (\beta_0(x), h^{-1}\beta_1(x))^\top$, (2.7) is equivalent to

$$Bias_{L,M}(\hat{\mathbf{b}}(x)) = \begin{bmatrix} \frac{1}{2}h^2\mu_2\frac{d^2m(x)}{dx^2} \\ o(h) \end{bmatrix}. \tag{2.8}$$

Comparing the leading bias expressions (2.6) derived by Fan et al. (1996) and (2.8) obtained by Masry (1996b), we observe that for $\hat{b}_0(x) = \hat{\beta}_0(x) = \hat{m}(x)$, the local linear estimator for $m(x)$, the leading bias are identical. However, for $\hat{b}_1(x) = h^{-1}\hat{\beta}_1(x) = \hat{m}'(x)$, the local linear estimator for $m'(x)$, Fan et al. (1996) obtained an explicit expression for the leading bias of order h^2, while Masry (1996b) only obtained the result of $o(h)$ without giving the explicit expression for the leading bias term.

To explain why Fan et al. and Masry obtained different bias expressions for the local linear estimator, we need first introduce some notations: For the local linear estimator ($p = 1$), by assuming that $m(\cdot)$ has continuous derivatives of total order $p + 2 = 3$ for $x \in \text{supp}(f)$, Fan et al. (1996) derived the conditional bias of $\hat{\mathbf{b}}(x)$ as (Eq. (2.1) and (4.3) of Fan et al. (1996)

$$Bias\left(\hat{\mathbf{b}}(x)\middle|\mathscr{X}\right) = \mathbb{E}\left(\hat{\mathbf{b}}(x)\middle|\mathscr{X}\right) - \mathbf{b}(x) = \mathbf{S}_n^{-1}\left(m_2(x)\mathbf{B}_{n,2} + m_3(x)\mathbf{B}_{n,3}\right) + (s.o.), \tag{2.9}$$

where $\mathscr{X} = \{X_i\}_{i=1}^n$, $m_2(x) = \frac{1}{2}\frac{d^2m(x)}{dx^2}$, $m_3(x) = \frac{1}{6}\frac{d^3m(x)}{dx^3}$,

$$\mathbf{S}_n = \begin{bmatrix} s_{n,0} & s_{n,1} \\ s_{n,1} & s_{n,2} \end{bmatrix}, \quad \mathbf{B}_{n,2} = \begin{bmatrix} s_{n,2} \\ s_{n,3} \end{bmatrix}, \quad \mathbf{B}_{n,3} = \begin{bmatrix} s_{n,3} \\ s_{n,4} \end{bmatrix}, \tag{2.10}$$

and

$$s_{n,j}(x) = \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)(X_i - x)^j, \tag{2.11}$$

for $j \in \{0, 1, \ldots, 4\}$.

We are now ready to explain the difference of bias terms in Fan et al. and in Masry. The reason for the differences in asymptotic bias expressions are twofold as follows:

1. Fan et al. (1996) assumed that the unknown function $m(\cdot)$ has continuous derivatives of total order $p+2=3$, while Masry (1996b) only assumed that it has total order $p+1=2$ (Eq. (2.11) in Masry, 1996b). Hence, Masry (1996b) could not obtain a bias term $\frac{1}{6}h^2\frac{\mu_4}{\mu_2}\frac{d^3m(x)}{dx^3}$ because he did not assume that $m(\cdot)$ is three-time differentiable. This term becomes the $o(h)$ term in (2.8).
2. Fan et al. (1996) kept both $O(1)$ terms and the $O(h)$ terms in the expression of $\mathbf{S}_n$ and $\mathbf{B}_{n,2}$, see (2.10) and (2.11), while Masry (1996b) only considered the leading $O(1)$ terms in $\mathbf{S}_n$. It can be shown that the bias term $\frac{1}{2}h^2\frac{\mu_4-\mu_2^2}{\mu_2}\frac{f'(x)}{f(x)}\frac{d^2m(x)}{dx^2}$ in (2.6) is related to the $O(h)$ terms in the expression of $\mathbf{S}_n$ and $\mathbf{B}_{n,2}$. Since Masry (1996b) did not give the explicit expression for the terms of order $O(h)$ in $\mathbf{S}_n$ and $\mathbf{B}_{n,2}$, he put the term $\frac{1}{2}h^2\frac{\mu_4-\mu_2^2}{\mu_2}\frac{f'(x)}{f(x)}\frac{d^2m(x)}{dx^2}$ as $o(h)$.

From the above discussion it is clear that in order to obtain explicit leading bias expressions for all components of $\hat{\mathbf{b}}(x)$, one needs to assume $m(\cdot)$ to be $p+2$ time-differentiable and utilizing Taylor expansion up to the order of $p+2$. Also, one needs to keep explicit expressions for terms of orders $O(1)$ *and* $O(h)$ in $\mathbf{S}_n$ and $\mathbf{B}_{n,2}$. In Section 3, we will use this strategy to study a general multivariate nonparametric regression model and obtain leading bias and variance terms for a pth -order local polynomial estimator.

2.2. The Univariate Local Linear Estimator: When x is a Boundary Point

In this subsection, we consider the case that x lies at the boundary of the support of f. In the last subsection we illustrated that, since Masry (1996b) did not keep Taylor expansion up to the term of the order h^{p+2} because he did not assume that m is $(p+2)$-time differentiable, and did not provide explicit expressions for terms of order $O(h)$ in $\mathbf{S}_n$ and $\mathbf{B}_{n,2}$, he had to put the leading bias of $\hat{b}_1(x)$ as $o(h)$. We will show that when x is at the boundary of the support of X, one only needs to keep Taylor expansion up to the term of the order h^{p+1} because the term with order h^{p+1} is not zero, and the term of order h^{p+2} becomes negligible. Also, we only need to keep the $O(1)$ term in $\mathbf{S}_n$ and $\mathbf{B}_{n,2}$, and we do not need to consider the $O(h)$ terms in the expression of $\mathbf{S}_n$ and $\mathbf{B}_{n,2}$ when x is a boundary point.

Without loss of generality, let $\mathrm{supp}(f)=\mathbb{R}_+$ so that $\{0\}$ is the boundary point. Suppose that $K(\cdot)$ is a kernel function with a bounded support $\mathrm{supp}(K)=[-1,1]$. Let $x=ch$ with $0\le c\le 1$. Then by a standard change-of-variable and Taylor-expansion argument, with $x_1=x+hv=ch+hv$, we have

$$\begin{aligned}\mathbb{E}\left(s_{n,j}(ch)\right) &= nh^j\int_0^1 K\left(\frac{x_1-ch}{h}\right)\left(\frac{x_1-ch}{h}\right)^j f(x_1)\,dx_1\\ &= nh^{j+1}\int_{-c}^{\infty} K(v)v^j f(ch+hv)\,dv\end{aligned}$$

$$= nh^{j+1} \int_{-c}^{\infty} K(v) v^j [f(ch) + f'(ch)hv + O(h^2)] dv$$
$$= nh^{j+1} \left\{ f(ch) \int_{-c}^{\infty} K(v) v^j dv + hf'(ch) \int_{-c}^{\infty} K(v) v^{j+1} dv + O\left(h^2\right) \right\}.$$

Denote

$$\int_{-c}^{\infty} K(v) v^j dv = \mu_j(c).$$

Then it is easy to show that

$$s_{n,j} = nh^{j+1} \left\{ f(ch)\mu_j(c) + hf'(ch)\mu_{j+1}(c) + O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right) \right\}, \tag{2.12}$$

where $\mu_j(c) \neq 0$ for all $j \in \mathbb{N}$ and for any $c \in [0, 1]$.

First, we derive the leading bias term using Fan et al. (1996)'s formula. When $x = ch$, in light of Eq. (2.12), Eq. (2.9) becomes

$$Bias\ (\hat{\mathbf{b}}(ch) \mid \mathscr{X}) = \begin{pmatrix} Bias\ (\hat{b}_0(ch) \mid \mathscr{X}) \\ Bias\ (\hat{b}_1(ch) \mid \mathscr{X}) \end{pmatrix}$$
$$= h^2 \begin{pmatrix} 1 & 0 \\ 0 & h^{-1} \end{pmatrix} [m_2(ch)(\mathscr{M}_1(c))^{-1} \mathscr{B}_2(c) + hB^*(ch)] + (s.o.), \tag{2.13}$$

with

$$B^*(ch) = \frac{f'(ch)m_2(ch) + f(ch)m_3(ch)}{f(ch)} (\mathscr{M}_1(c))^{-1} \mathscr{B}_3(c)$$
$$- \frac{f'(ch)}{f(ch)} m_2(ch)(\mathscr{M}_1(c))^{-1} \mathscr{M}_1^1(c)(\mathscr{M}_1(c))^{-1} \mathscr{B}_2(c), \tag{2.14}$$

where,

$$\mathscr{M}_1(c) = \begin{bmatrix} \mu_0(c) & \mu_1(c) \\ \mu_1(c) & \mu_2(c) \end{bmatrix}, \quad \mathscr{M}_1^1(c) = \begin{bmatrix} \mu_1(c) & \mu_2(c) \\ \mu_2(c) & \mu_3(c) \end{bmatrix}, \quad \mathscr{B}_2(c) = \begin{bmatrix} \mu_2(c) \\ \mu_3(c) \end{bmatrix}, \quad \mathscr{B}_3(c) = \begin{bmatrix} \mu_3(c) \\ \mu_4(c) \end{bmatrix}. \tag{2.15}$$

Therefore, when $x = ch$, the leading bias of $\hat{\mathbf{b}}(ch)$, according to Fan et al. (1996), is as follows:

$$\mathbf{Bias}_{L,F}\left(\hat{b}_0(ch) \middle| \mathscr{X}\right) = \frac{1}{2} h^2 \frac{d^2 m(ch)}{dx^2} \frac{(\mu_2(c))^2 - \mu_1(c)\mu_3(c)}{\mu_0(c)\mu_2(c) - (\mu_1(c))^2}, \tag{2.16}$$
$$\mathbf{Bias}_{L,F}\left(\hat{b}_1(ch) \middle| \mathscr{X}\right) = \frac{1}{2} h \frac{d^2 m(ch)}{dx^2} \frac{-\mu_1(c)\mu_2(c) + \mu_0(c)\mu_3(c)}{\mu_0(c)\mu_2(c) - (\mu_1(c))^2}.$$

Next we derive the leading bias term by Masry (1996b)'s formula. When $x = ch$, according to Eq. (2.12), Masry's leading bias is given by

$$h^2 \left(\mathscr{M}_1(c)\right)^{-1} \mathscr{B}_2(c) m_2(ch), \tag{2.17}$$

where,

$$\mathscr{M}_1(c) = \begin{bmatrix} \mu_0(c) & \mu_1(c) \\ \mu_1(c) & \mu_2(c) \end{bmatrix}, \quad \mathscr{B}_2(c) = \begin{bmatrix} \mu_2(c) \\ \mu_3(c) \end{bmatrix}. \tag{2.18}$$

Therefore, when $x = ch$, the leading bias of $\widehat{\beta}(ch)$, according to Masry (1996b), is as follows:

$$\begin{aligned} \mathbf{Bias}_{L,M}\left(\hat{\beta}_0(ch)\middle|\mathscr{X}\right) &= \frac{1}{2}h^2 \frac{d^2 m(ch)}{dx^2} \frac{(\mu_2(c))^2 - \mu_1(c)\mu_3(c)}{\mu_0(c)\mu_2(c) - (\mu_1(c))^2}, \\ \mathbf{Bias}_{L,M}\left(\hat{\beta}_1(ch)\middle|\mathscr{X}\right) &= \frac{1}{2}h^2 \frac{d^2 m(ch)}{dx^2} \frac{-\mu_1(c)\mu_2(c) + \mu_0(c)\mu_3(c)}{\mu_0(c)\mu_2(c) - (\mu_1(c))^2}. \end{aligned} \tag{2.19}$$

The leading bias of $\hat{\mathbf{b}}(ch)$ becomes

$$\begin{aligned} \mathbf{Bias}_{L,M}\left(\hat{b}_0(ch)\middle|\mathscr{X}\right) &= \frac{1}{2}h^2 \frac{d^2 m(ch)}{dx^2} \frac{(\mu_2(c))^2 - \mu_1(c)\mu_3(c)}{\mu_0(c)\mu_2(c) - (\mu_1(c))^2}, \\ \mathbf{Bias}_{L,M}\left(\hat{b}_1(ch)\middle|\mathscr{X}\right) &= \frac{1}{2}h \frac{d^2 m(ch)}{dx^2} \frac{-\mu_1(c)\mu_2(c) + \mu_0(c)\mu_3(c)}{\mu_0(c)\mu_2(c) - (\mu_1(c))^2}. \end{aligned} \tag{2.20}$$

Comparing the leading bias in Eq. (2.16) and (2.20), we observe that when x is a boundary point, formulas of Fan et al. (1996) and Masry (1996b) have identical leading biases; both give to the $O\left(h^2\right)$ leading bias for $\hat{b}_0(x)$ and the $O(h)$ leading bias for $\hat{b}_1(x)$. This is what one should expect, because when x is a boundary point, the leading bias term in Masry (1996b) will not be zero due to $\int_{-c}^{\infty} k(v)v^l dv \neq 0$ for any integer l. This differs from the interior x case where $\int_{-\infty}^{\infty} k(v)v^l = 0$ when l is an odd integer.

3. THE MULTIVARIATE LOCAL POLYNOMIAL ESTIMATOR

In this section, we discuss the leading bias of multivariate local polynomial estimator. We follow closely the notations of Masry (1996b) except that we assume the unknown function $m(\cdot)$ to have derivatives of total order $p+2$. We also keep explicit expressions for both $O(1)$ and $O(\mathbf{h})$ terms in $\mathbf{S}_n$ and $\mathbf{B}_{n,2}$. By making these two differences with Masry (1996b), we derive the leading bias for all components of the vector multivariate local polynomial estimator.

Consider a general multivariate nonparametric regression model

$$Y_i = m(\mathbf{X}_i) + u_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where Y_i is a real valued random variable taking value in $\mathbb{R}$, and $\mathbf{X}_i$ is a real valued d-dimensional vector of random variables taking value in $\mathbb{R}^d$. By approximating the unknown function $m(\cdot)$ with its first p terms of the Taylor expansion, and using the notations introduced below, we have

$$m(\mathbf{X}_i) \approx \sum_{0 \le |\mathbf{k}| \le p} \frac{1}{\mathbf{k}!} \left(D^{\mathbf{k}} m\right)(\mathbf{x}) (\mathbf{X}_i - \mathbf{x})^{\mathbf{k}}, \tag{3.2}$$

the local polynomial estimator, which estimates $\left\{(\mathbf{k}!)^{-1} \left(D^{\mathbf{k}} m\right)(\mathbf{x})\right\}$, is the minimizer of the following weighted least squares problem

$$\underset{0 \le |\mathbf{k}| \le p}{\{\hat{b}_{\mathbf{k}}(\mathbf{x})\}} = \underset{\{b_{\mathbf{k}}(\mathbf{x})\}_{0 \le |\mathbf{k}| \le p}}{\arg\min} \sum_{i=1}^{n} \left\{ Y_i - \sum_{0 \le |\mathbf{k}| \le p} b_{\mathbf{k}}(\mathbf{x}) (\mathbf{X}_i - \mathbf{x})^{\mathbf{k}} \right\}^2 K_{\mathbf{h},i\mathbf{x}}. \tag{3.3}$$

The following notations are used in (3.2) and (3.3) above and for the remaining parts of the article. Most of our notations are consistent with those in Masry (1996b).

Notation 1.

(i). For $k \in \{0, 1, \ldots, p\}$, let $N_k = \binom{k+d-1}{d-1}$ and $\mathcal{N}_p = \sum_{k=0}^{p} N_k$.

(ii). The bandwidth $\mathbf{h} = \{h_1, \ldots, h_d\}$. $\mathbf{H}$ is a $d \times d$ diagonal matrix with diagonal terms $\{h_1, \ldots, h_d\}$. $\|\mathbf{h}\| = \max(h_1, \ldots, h_d)$. Also $\mathbf{H}^{(k)}$ for any $k \in \{1, \ldots, d\}$ is an $N_k \times N_k$ diagonal matrix with diagonal terms $\text{vech}(\mathbf{H}^k)$, i.e., $\text{diag}(\mathbf{H}^{(k)}) = \{h_1^k, h_1^{k-1}h_2, \ldots, h_d^k\}$.

(iii). $K(\mathbf{u})$ is the production kernel function,

$$K(\mathbf{u}) = \prod_{l=1}^{d} k(u_l) = k(u_1) \times \cdots \times k(u_d), \tag{3.4}$$

and

$$\begin{aligned} K_{\mathbf{h},i\mathbf{x}} &= \left(\prod_{l=1}^{d} \frac{1}{h_l}\right) K\left(\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\right) = \prod_{l=1}^{d} \frac{1}{h_l} k\left(\frac{X_{i,l} - x_l}{h_l}\right) \\ &= \frac{1}{h_1} k\left(\frac{X_{i,1} - x_1}{h_1}\right) \times \cdots \times \frac{1}{h_d} k\left(\frac{X_{i,d} - x_d}{h_d}\right). \end{aligned} \tag{3.5}$$

(iv). Let $\mathbf{k} = \{k_1, \ldots, k_d\}$ and $|\mathbf{k}| = \sum_{l=1}^{d} k_l$,

$$\mathbf{k}! = k_1! \times \cdots \times k_d!, \quad \mathbf{x}^{\mathbf{k}} = x_1^{k_1} \times \cdots \times x_d^{k_d}, \tag{3.6}$$

$$\sum_{0\le|\mathbf{k}|\le p} = \sum_{j=0}^{p} \sum_{\substack{k_1=0 \\ |\mathbf{k}|=k_1+\cdots+k_d=j}}^{j} \cdots \sum_{k_d=0}^{j}, \quad \left(D^{\mathbf{k}} m\right)(\mathbf{x}) = \frac{\partial^{|\mathbf{k}|} m(\mathbf{x})}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}, \tag{3.7}$$

$$m_{\mathbf{k}}(\mathbf{x}) = \frac{1}{\mathbf{k}!}(D^{\mathbf{k}} m)(\mathbf{x}), \quad \beta_{\mathbf{k}}(\mathbf{x}) = \mathbf{h}^{\mathbf{k}} m_{\mathbf{k}}(\mathbf{x}) = \frac{\mathbf{h}^{\mathbf{k}}}{\mathbf{k}!}\left(D^{\mathbf{k}} m\right)(\mathbf{x}). \tag{3.8}$$

$\mathbf{m}(\mathbf{x}) = [m(\mathbf{x}), \mathbf{m}_{\mathbf{1}}^{\top}(\mathbf{x}), \ldots, \mathbf{m}_{\mathbf{p}}^{\top}(\mathbf{x})]^{\top}$, and $\beta(\mathbf{x}) = [\beta_0(\mathbf{x}), \beta_{\mathbf{1}}^{\top}(\mathbf{x}), \ldots, \beta_{\mathbf{p}}^{\top}(\mathbf{x})]^{\top}$ are $\mathcal{N}_p \times 1$ column vectors. $\mathbf{m}_{\mathbf{k}}(\mathbf{x})$ and $\beta_{\mathbf{k}}(\mathbf{x})$ are $N_{|\mathbf{k}|} \times 1$ column vectors composed of $m_{\mathbf{k}}(\mathbf{x})$ and $\beta_{\mathbf{k}}(\mathbf{x})$, respectively, by the lexicographical order.

(v). Let $g_k^{-1} : \mathbb{N}^d \to \mathbb{N}$ be a one-to-one mapping which arranges the N_k d-tuples $\{\mathbf{k}\}$ into a sequence in a lexicographical order. That is, $g_k^{-1}\left(\{k, 0, \ldots, 0\}\right) = 1$, $g_k^{-1}\left(\{k-1, 1, 0, \ldots, 0\}\right) = 2$, ..., and $g_k^{-1}\left(\{0, \ldots, 0, k\}\right) = N_k$.

(vi). Let $\mathcal{M}_p$ and $\mathcal{G}_p$ be $\mathcal{N}_p \times \mathcal{N}_p$ matrices, and $\mathcal{B}_k$ be an $\mathcal{N}_p \times N_k$ matrix,

$$\mathcal{M}_p = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,p} \\ \vdots & \vdots & & \vdots \\ M_{p,0} & M_{p,1} & \cdots & M_{p,p} \end{bmatrix}, \quad \mathcal{B}_k = \begin{bmatrix} M_{0,k} \\ M_{1,k} \\ \vdots \\ M_{p,k} \end{bmatrix}, \quad \mathcal{G}_p = \begin{bmatrix} \mathbf{\Gamma}_{0,0} & \mathbf{\Gamma}_{0,1} & \cdots & \mathbf{\Gamma}_{0,p} \\ \mathbf{\Gamma}_{1,0} & \mathbf{\Gamma}_{1,1} & \cdots & \mathbf{\Gamma}_{1,p} \\ \vdots & \vdots & & \vdots \\ \mathbf{\Gamma}_{p,0} & \mathbf{\Gamma}_{p,1} & \cdots & \mathbf{\Gamma}_{p,p} \end{bmatrix}, \tag{3.9}$$

where $M_{i,j}$ and $\mathbf{\Gamma}_{i,j}$ are $N_i \times N_j$ matrices,

$$M_{i,j} = \int_{\mathbb{R}^d} \mathrm{vech}\left(\mathbf{u}^{\otimes i}\right) \mathrm{vech}^{\top}\left(\mathbf{u}^{\otimes j}\right) K(\mathbf{u}) du, \quad \Gamma_{i,j} = \int_{\mathbb{R}^d} \mathrm{vech}\left(\mathbf{u}^{\otimes i}\right) \mathrm{vech}^{\top}\left(\mathbf{u}^{\otimes j}\right) K^2(\mathbf{u}) du. \tag{3.10}$$

For $l \in \{0, 1, \ldots, d\}$, let $\mathcal{M}_p^l$ be an $\mathcal{N}_p \times \mathcal{N}_p$ matrix, and $\mathcal{B}_k^l$ be an $\mathcal{N}_p \times N_k$ matrix,

$$\mathcal{M}_p^l = \begin{bmatrix} M_{0,0}^l & M_{0,1}^l & \cdots & M_{0,p}^l \\ M_{1,0}^l & M_{1,1}^l & \cdots & M_{1,p}^l \\ \vdots & \vdots & & \vdots \\ M_{p,0}^l & M_{p,1}^l & \cdots & M_{p,p}^l \end{bmatrix}, \quad \mathcal{B}_k^l = \begin{bmatrix} M_{0,k}^l \\ M_{1,k}^l \\ \vdots \\ M_{p,k}^l \end{bmatrix}, \tag{3.11}$$

where $M_{i,j}^l$ is an $N_i \times N_j$ matrix,

$$M_{i,j}^l = \int_{\mathbb{R}^d} u_l \, \mathrm{vech}\left(\mathbf{u}^{\otimes i}\right) \mathrm{vech}^{\top}\left(\mathbf{u}^{\otimes j}\right) K(\mathbf{u}) du. \tag{3.12}$$

where u_l is the lth component of the vector $\mathbf{u}$.

Assumption 1.

(a). $\{\mathbf{X}_i\}$ *are i.i.d. with a common density* $f(\cdot)$ *with support* $\text{supp}(f) \subset \mathbb{R}^d$. $f(\mathbf{x}) > 0$ *for* $\mathbf{x} \in \text{supp}(f)$. $\sigma^2(\mathbf{x}) = \text{var}\left(Y \,|\mathbf{X} = \mathbf{x}\right)$ *is finite.* $\{\varepsilon_i\}$ *are i.i.d. with zero mean and unit variance and are independent of* $\{\mathbf{X}_i\}$.

(b). $m(\mathbf{x})$ *has continuous derivatives of total order* $p+2$ *for all* $x \in \text{supp}(f)$.

(c). $k(\cdot)$ *is any nonnegative, symmetric, and bounded kernel function with* $\int_{\mathbb{R}} k(v)dv = 1$, $k(v) = k(-v)$, *and*

$$\int_{\mathbb{R}} k(v)v^q dv = \mu_q \in [0,\infty), \quad \int_{\mathbb{R}} k^2(v)v^r dv = \lambda_r \in [0,\infty), \tag{3.13}$$

for $q \in \{0,1,\ldots,2p+2\}$, $r \in \{0,1,\ldots,2p\}$. *Note that* $\mu_0 = 1$, *and* $\mu_q = \lambda_q = 0$ *when* q *is odd.*

(d). *As* $n \to \infty$, $h_l \to 0$ *for all* $l \in \{0,1,\ldots,d\}$, *and* $nh_1h_2\cdots h_d \to \infty$.

Let $\widehat{(D^{\mathbf{k}}m)}(\mathbf{x}) = \mathbf{k}!\hat{b}_{\mathbf{k}}(\mathbf{x})$ and $\hat{\beta}_{\mathbf{k}}(\mathbf{x}) = \mathbf{h}^{\mathbf{k}}\hat{b}_{\mathbf{k}}(\mathbf{x}) = \frac{\mathbf{h}^{\mathbf{k}}}{\mathbf{k}!}\widehat{(D^{\mathbf{k}}m)}(\mathbf{x})$. As in Masry (1996b)'s Eqs. (1.9)–(1.12), the minimization of Eq. (3.3) leads to,

$$t_{n,j}(\mathbf{x}) = \sum_{0\le|\mathbf{k}|\le p} \hat{\beta}_{\mathbf{k}}(\mathbf{x})s_{n,j+k}(\mathbf{x}), \tag{3.14}$$

where $0 \le |j| \le p$, and

$$t_{n,j}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}}[\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})]^j Y_i, \tag{3.15}$$

$$s_{n,j+k}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}}[\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})]^{j+\mathbf{k}}. \tag{3.16}$$

Or, by arranging $s_{n,j+k}(\mathbf{x})$ and $t_{n,j}(\mathbf{x})$ in a lexicographical order, Eq. (3.14) can be written as

$$\boldsymbol{\tau}_n(\mathbf{x}) = \mathbf{S}_n(\mathbf{x})\widehat{\beta}(\mathbf{x}), \tag{3.17}$$

where $\boldsymbol{\tau}_n(\mathbf{x})$ is an $\mathcal{N}_p \times 1$ vector, $\mathbf{S}_n(\mathbf{x})$ is an $\mathcal{N}_p \times \mathcal{N}_p$ matrix, and $\widehat{\beta}(\mathbf{x})$ is an $\mathcal{N}_p \times 1$ vector. As in Masry (1996b)'s Eqs. (2.2) and (2.5),

$$\boldsymbol{\tau}_n(\mathbf{x}) = \begin{bmatrix} \boldsymbol{\tau}_{n,0}(\mathbf{x}) \\ \boldsymbol{\tau}_{n,1}(\mathbf{x}) \\ \vdots \\ \boldsymbol{\tau}_{n,p}(\mathbf{x}) \end{bmatrix}, \quad \mathbf{S}_n(\mathbf{x}) = \begin{bmatrix} \mathbf{S}_{n,0,0}(\mathbf{x}) & \mathbf{S}_{n,0,1}(\mathbf{x}) & \cdots & \mathbf{S}_{n,0,p}(\mathbf{x}) \\ \mathbf{S}_{n,1,0}(\mathbf{x}) & \mathbf{S}_{n,1,1}(\mathbf{x}) & \cdots & \mathbf{S}_{n,1,p}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \mathbf{S}_{n,p,0}(\mathbf{x}) & \mathbf{S}_{n,p,1}(\mathbf{x}) & \cdots & \mathbf{S}_{n,p,p}(\mathbf{x}) \end{bmatrix}, \tag{3.18}$$

in which $\boldsymbol{\tau}_{n,j}(\mathbf{x})$ is an $N_j \times 1$ vector composed by $\{t_{n,j}(\mathbf{x})\}$, and $\mathbf{S}_{n,j,k}(\mathbf{x})$ an $N_j \times N_k$ matrix composed by $\{s_{n,j+k}(\mathbf{x})\}$.

Following Masry (1996b)'s Eqs. (2.9) and (2.10), let

$$t^*_{n,j}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}}[\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})]^j [Y_i - m(\mathbf{X}_i)] = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \left[\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\right]^j u_i. \quad (3.19)$$

Therefore,

$$t_{n,j}(\mathbf{x}) - t^*_{n,j}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \left[\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\right]^j m(\mathbf{X}_i). \quad (3.20)$$

By Assumption (b), $m(\mathbf{X}_i)$ could be approximated by its first $p+2$ terms of the Taylor expansion,

$$\begin{aligned} m(\mathbf{X}_i) &= \sum_{0\le|\mathbf{k}|\le(p+2)} \frac{1}{\mathbf{k}!}\left(D^{\mathbf{k}} m\right)(\mathbf{x})(\mathbf{X}_i - \mathbf{x})^{\mathbf{k}} + o_p\left(\|\mathbf{h}\|^{p+2}\right) \quad (3.21)\\ &= \sum_{0\le|\mathbf{k}|\le p} \beta_{\mathbf{k}}(\mathbf{x})\left[\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\right]^{\mathbf{k}} + \sum_{(p+1)\le|\mathbf{k}|\le(p+2)} \frac{1}{\mathbf{k}!}\left(D^{\mathbf{k}} m\right)(\mathbf{x})(\mathbf{X}_i - \mathbf{x})^{\mathbf{k}} + o_p\left(\|\mathbf{h}\|^{p+2}\right). \end{aligned}$$

By substituting Eq. (3.21) and Eq. (3.14) into Eq. (3.20), we obtain

$$\begin{aligned} t^*_{n,j}(\mathbf{x}) = &\sum_{0\le|\mathbf{k}|\le p} \left(\hat{\beta}_{\mathbf{k}}(\mathbf{x}) - \beta_{\mathbf{k}}(\mathbf{x})\right) s_{n,j+k}(\mathbf{x}) \quad (3.22)\\ &- \sum_{(p+1)\le|\mathbf{k}|\le(p+2)} \frac{\mathbf{h}^{\mathbf{k}}}{\mathbf{k}!}\left(D^{\mathbf{k}} m\right)(\mathbf{x}) s_{n,j+k}(\mathbf{x}) + o_p\left(\|\mathbf{h}\|^{p+2}\right) s_{n,j}(\mathbf{x}). \end{aligned}$$

Again, following Masry (1996b), we rewrite Eq. (3.22) in matrix form,

$$\begin{aligned} \boldsymbol{\tau}^*_n(\mathbf{x}) = &\mathbf{S}_n(\mathbf{x})\left(\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x})\right) - \mathbf{B}_{n,p+1}(\mathbf{x})\mathbf{H}^{(p+1)}\mathbf{m}_{p+\mathbf{1}}(\mathbf{x}) \quad (3.23)\\ &-\mathbf{B}_{n,p+2}(\mathbf{x})\mathbf{H}^{(p+2)}\mathbf{m}_{p+\mathbf{2}}(\mathbf{x}) + o_p\left(\|\mathbf{h}\|^{p+2}\right)\mathbf{B}_{n,0}(\mathbf{x}), \end{aligned}$$

where $\boldsymbol{\tau}^*_n(\mathbf{x})$ is an $\mathcal{N}_p \times 1$ vector, $\mathbf{B}_{n,k}(\mathbf{x})$ for any $k \in \{0,\ldots,p+2\}$ is an $\mathcal{N}_p \times N_k$ matrix, $\mathbf{H}^{(k)}$ is an $N_k \times N_k$ matrix, and $\mathbf{m}_{\mathbf{k}}(\mathbf{x})$ is an $N_k \times 1$ vector of $|\mathbf{k}|$-order partial derivatives. As in Eq. (2.13) of Masry (1996b),

$$\boldsymbol{\tau}^*_n(\mathbf{x}) = \begin{bmatrix} \boldsymbol{\tau}^*_{n,0}(\mathbf{x}) \\ \boldsymbol{\tau}^*_{n,1}(\mathbf{x}) \\ \vdots \\ \boldsymbol{\tau}^*_{n,p}(\mathbf{x}) \end{bmatrix}, \quad \mathbf{B}_{n,k}(\mathbf{x}) = \begin{bmatrix} \mathbf{S}_{n,0,k}(\mathbf{x}) \\ \mathbf{S}_{n,1,k}(\mathbf{x}) \\ \vdots \\ \mathbf{S}_{n,p,k}(\mathbf{x}) \end{bmatrix}, \quad (3.24)$$

in which $\boldsymbol{\tau}^*_{n,j}(\mathbf{x})$ is a $N_j \times 1$ vector composed by $\{t^*_{n,j}(\mathbf{x})\}$, and, again, $\mathbf{S}_{n,j,k}(\mathbf{x})$ an $N_j \times N_k$ matrix composed by $\{s_{n,j+k}(\mathbf{x})\}$.

Thus,

$$\begin{aligned}\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) &= (\mathbf{S}_n(\mathbf{x}))^{-1}\boldsymbol{\tau}^*_n(\mathbf{x}) + (\mathbf{S}_n(\mathbf{x}))^{-1}\{\mathbf{B}_{n,p+1}(\mathbf{x})\mathbf{H}^{(p+1)}\mathbf{m}_{p+\mathbf{1}}(\mathbf{x}) + \mathbf{B}_{n,p+2}(\mathbf{x})\mathbf{H}^{(p+2)}\mathbf{m}_{p+\mathbf{2}}(\mathbf{x})\} \\ &\quad + o_p(\|\mathbf{h}\|^{p+2})(\mathbf{S}_n(\mathbf{x}))^{-1}\mathbf{B}_{n,0}(\mathbf{x}). \end{aligned} \tag{3.25}$$

Lemma 1. *Under Assumptions (a)–(d), the following equations hold:*

1.

$$\mathbf{S}_n(\mathbf{x}) = \mathscr{M}_p f(\mathbf{x}) + \sum_{l=1}^{d} h_l \mathscr{M}^l_p f_l(\mathbf{x}) + O_p\left(\|\mathbf{h}\|^2 + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right), \tag{3.26}$$

where $f_l(x) = \frac{\partial f(x)}{\partial x_l}$.

2.

$$\mathbf{B}_{n,p+1}(\mathbf{x}) = \mathscr{B}_{p+1} f(\mathbf{x}) + \sum_{l=1}^{d} h_l \mathscr{B}^l_{p+1} f_l(\mathbf{x}) + O_p\left(\|\mathbf{h}\|^2 + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right). \tag{3.27}$$

3.

$$\mathbf{B}_{n,p+2}(\mathbf{x}) = \mathscr{B}_{p+2} f(\mathbf{x}) + O_p\left(\|\mathbf{h}\| + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right). \tag{3.28}$$

4.

$$\sqrt{nh_1h_2\cdots h_d}\,\boldsymbol{\tau}^*_n(\mathbf{x}) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2(\mathbf{x}) f(\mathbf{x}) \mathscr{G}_p\right). \tag{3.29}$$

Proof. The proofs are trivial and are therefore omitted here.

By Eq. (3.25), Lemma 1, and using the identity $(A + hB + o(h))^{-1} = A^{-1} - hA^{-1}BA^{-1} + o(h)$, we have

$$(\mathbf{S}_n(\mathbf{x}))^{-1} = \frac{1}{f(\mathbf{x})}\mathscr{M}^{-1}_p - \sum_{l=1}^{d} h_l \frac{f_l(\mathbf{x})}{f^2(\mathbf{x})}\mathscr{M}^{-1}_p \mathscr{M}^l_p \mathscr{M}^{-1}_p + O_p\left(\|\mathbf{h}\|^2 + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right), \tag{3.30}$$

and

$$\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) = (\mathbf{S}_n(\mathbf{x}))^{-1}\boldsymbol{\tau}^*_n(\mathbf{x}) + \underbrace{\mathscr{M}^{-1}_p \mathscr{B}_{p+1}\mathbf{H}^{(p+1)}\mathbf{m}_{p+1}(\mathbf{x})}_{O(\|\mathbf{h}\|^{p+1})}$$

$$+\underbrace{\sum_{l=1}^{d} h_l \frac{f_l(\mathbf{x})}{f(\mathbf{x})}\left(\mathscr{M}_p^{-1}\mathscr{B}_{p+1}^{l}-\mathscr{M}_p^{-1}\mathscr{M}_p^{l}\mathscr{M}_p^{-1}\mathscr{B}_{p+1}\right)\mathbf{H}^{(p+1)}\mathbf{m}_{p+1}(\mathbf{x})+\mathscr{M}_p^{-1}\mathscr{B}_{p+2}\mathbf{H}^{(p+2)}\mathbf{m}_{p+2}(\mathbf{x})}_{O(\|\mathbf{h}\|^{p+2})}$$

$$+o_p\left(\|\mathbf{h}\|^{p+2}\right)+O_p\left(\frac{\|\mathbf{h}\|^{p+1}}{\sqrt{nh_1h_2\cdots h_d}}\right). \tag{3.31}$$

Theorem 1. *Under Assumptions (a)–(d), assume that* $nh_1h_2\cdots h_d \min\left\{h_1^{2p},h_2^{2p},\ldots,h_d^{2p}\right\} \to \infty$ *and* $nh_1h_2\cdots h_d\|\mathbf{h}\|^{2p+6}\to 0$ *as* $n\to\infty$. *Then we have the following situations:*

1.

$$\sqrt{nh_1h_2\cdots h_d}\left(\widehat{\beta}(\mathbf{x})-\beta(\mathbf{x})-\mathbf{Bias}(\mathbf{x})\right)\xrightarrow{d} N\left(\mathbf{0},\boldsymbol{\Sigma}(\mathbf{x})\right), \tag{3.32}$$

where

$$\mathbf{Bias}(\mathbf{x})=\underbrace{\mathscr{M}_p^{-1}\mathscr{B}_{p+1}\mathbf{H}^{(p+1)}\mathbf{m}_{p+1}(\mathbf{x})}_{O(\|\mathbf{h}\|^{p+1})} \tag{3.33}$$

$$+\underbrace{\sum_{l=1}^{d} h_l \frac{f_l(\mathbf{x})}{f(\mathbf{x})}\left(\mathscr{M}_p^{-1}\mathscr{B}_{p+1}^{l}-\mathscr{M}_p^{-1}\mathscr{M}_p^{l}\mathscr{M}_p^{-1}\mathscr{B}_{p+1}\right)\mathbf{H}^{(p+1)}\mathbf{m}_{p+1}(\mathbf{x})+\mathscr{M}_p^{-1}\mathscr{B}_{p+2}\mathbf{H}^{(p+2)}\mathbf{m}_{p+2}(\mathbf{x})}_{O(\|\mathbf{h}\|^{p+2})}.$$

and

$$\boldsymbol{\Sigma}(\mathbf{x})=\frac{\sigma^2(\mathbf{x})}{f(\mathbf{x})}\mathscr{M}_p^{-1}\mathscr{G}_p\mathscr{M}_p^{-1}. \tag{3.34}$$

2. *Let* $\hat{\beta}_{\mathbf{k}}(\mathbf{x})=\frac{\mathbf{h}^{\mathbf{k}}}{\mathbf{k}!}\left(\widehat{D^{\mathbf{k}}m}\right)(\mathbf{x})=\left(\widehat{\beta}(\mathbf{x})\right)_j$ *be the* jth *element in* $\widehat{\beta}(\mathbf{x})$. *That is,* $j=g_k^{-1}(\mathbf{k})+\sum_{l=1}^{k-1}N_l$. *Recall that* $\mathbf{h}^{\mathbf{k}}=h_1^{k_1}h_2^{k_2}\cdots h_d^{k_d}$, $\left(D^{\mathbf{k}}m\right)(\mathbf{x})=\frac{\partial^{|\mathbf{k}|}m(\mathbf{x})}{\partial x_1^{k_1}\partial x_2^{k_2}\cdots\partial x_d^{k_d}}$, $\mathbf{k}!=k_1!\times\cdots\times k_d!$, *and* $|\mathbf{k}|=k_1+\cdots+k_d$. *Then,*

$$\sqrt{nh_1h_2\cdots h_d}\mathbf{h}^{\mathbf{k}}\left(\left(\widehat{D^{\mathbf{k}}m}\right)(\mathbf{x})-\left(D^{\mathbf{k}}m\right)(\mathbf{x})-\frac{\mathbf{k}!}{\mathbf{h}^{\mathbf{k}}}\left(\mathbf{Bias}(\mathbf{x})\right)_j\right)\xrightarrow{d} N\left(\mathbf{0},\left(\boldsymbol{\Sigma}(\mathbf{x})\right)_{j,j}\right), \tag{3.35}$$

where, when $p-|\mathbf{k}|$ *is odd,*

$$\left(\mathbf{Bias}(\mathbf{x})\right)_j=\left(\mathscr{M}_p^{-1}\mathscr{B}_{p+1}\mathbf{H}^{(p+1)}\mathbf{m}_{p+1}(\mathbf{x})\right)_j, \tag{3.36}$$

and, when $p - |\mathbf{k}|$ *is even,*

$$(\mathbf{Bias}(\mathbf{x}))_j = \left(\sum_{l=1}^{d} h_l \frac{f_l(\mathbf{x})}{f(\mathbf{x})} (\mathscr{M}_p^{-1} \mathscr{B}_{p+1}^l - \mathscr{M}_p^{-1} \mathscr{M}_p^l \mathscr{M}_p^{-1} \mathscr{B}_{p+1}) \mathbf{H}^{(p+1)} \mathbf{m}_{p+1}(\mathbf{x}) + \mathscr{M}_p^{-1} \mathscr{B}_{p+2} \mathbf{H}^{(p+2)} \mathbf{m}_{p+2}(\mathbf{x}) \right)_j . \tag{3.37}$$

Proof. 1. It follows from Eq. (3.31) and the proof of Theorem 4 of Masry (1996b).

2. By Assumption (c), $\mu_q = \int k(v) v^q dv = 0$ when q is odd, and $\mu_q \neq 0$ when q is even. Therefore, when $p - |\mathbf{k}|$ is odd, the jth element of the $O\left(\|\mathbf{h}\|^{p+1}\right)$ part in $\mathbf{Bias}(\mathbf{x})$ is nonzero, hence, the $O\left(\|\mathbf{h}\|^{p+2}\right)$ part is negligible. When $p - |\mathbf{k}|$ is even, the jth element of the $O\left(\|\mathbf{h}\|^{p+1}\right)$ part is zero, and the $O\left(\|\mathbf{h}\|^{p+2}\right)$ part becomes the leading bias.

In order to obtain a nonzero leading bias term for either odd $p - |\mathbf{k}|$ or even $p - |\mathbf{k}|$, we need to keep two terms in the asymptotic bias expression, the $O\left(\|\mathbf{h}\|^{p+1}\right)$ part and the $O\left(\|\mathbf{h}\|^{p+2}\right)$ part. When $p - |\mathbf{k}|$ is odd, the $O\left(\|\mathbf{h}\|^{p+1}\right)$ part in $\mathbf{Bias}(\mathbf{x})$ is nonzero, and the $O\left(\|\mathbf{h}\|^{p+2}\right)$ part is negligible. However, when $p - |\mathbf{k}|$ is even, the $O\left(\|\mathbf{h}\|^{p+1}\right)$ part is zero because of Assumption (c), and the $O\left(\|\mathbf{h}\|^{p+2}\right)$ part becomes the leading bias.

The optimal bandwidth minimizing mean squared error (MSE) of $\left(\widehat{D^{\mathbf{k}} m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}} m\right)(\mathbf{x})$ can be derived easily.

Lemma 2.

1. *When* $p - |\mathbf{k}|$ *is odd, the optimal bandwidth minimizing the MSE of* $\left(\widehat{D^{\mathbf{k}} m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}} m\right)(\mathbf{x})$ *is* $h_l = O\left(n^{-1/(d+2p+2)}\right)$ *for* $l \in \{1, \ldots, d\}$.
2. *When* $p - |\mathbf{k}|$ *is even, the optimal bandwidth minimizing the MSE of* $\left(\widehat{D^{\mathbf{k}} m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}} m\right)(\mathbf{x})$ *is* $h_l = O\left(n^{-1/(d+2p+4)}\right)$ *for* $l \in \{1, \ldots, d\}$.

Proof. Denote $A_n = O_{e,p}(a_n)$ if a sequence $\{A_n\}$ is of exact order a_n in probability. That is, $A_n = O_{e,p}(a_n)$ if and only if $A_n = O_p(a_n)$ but $A_n \neq o_p(a_n)$. *When* $p - |\mathbf{k}|$ *is odd,*

$$\left(\widehat{D^{\mathbf{k}} m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}} m\right)(\mathbf{x}) = O_{e,p}\left(\|\mathbf{h}\|^{p+1} \mathbf{h}^{\mathbf{k}} + \frac{1}{\sqrt{n h_1 h_2 \cdots h_d \mathbf{h}^{2\mathbf{k}}}} \right). \tag{3.38}$$

Thus, the optimal bandwidth minimizing the MSE *of* $\left(\widehat{D^{\mathbf{k}} m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}} m\right)(\mathbf{x})$ *is* $h_l = O\left(n^{-1/(d+2p+2)}\right)$ *for* $l \in \{1, \ldots, d\}$. *When* $p - |\mathbf{k}|$ *is even,*

$$\left(\widehat{D^{\mathbf{k}} m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}} m\right)(\mathbf{x}) = O_{e,p}\left(\|\mathbf{h}\|^{p+2} \mathbf{h}^{\mathbf{k}} + \frac{1}{\sqrt{n h_1 h_2 \cdots h_d \mathbf{h}^{2\mathbf{k}}}} \right). \tag{3.39}$$

Thus, the optimal bandwidth minimizing the MSE of $\left(\widehat{D^{\mathbf{k}} m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}} m\right)(\mathbf{x})$ is $h_l = O\left(n^{-1/(d+2p+4)}\right)$ for $l \in \{1, \ldots, d\}$.

In practice, the bandwidth could be chosen by the plug-in method. For each $j = g_k^{-1}(\mathbf{k}) + \sum_{l=1}^{k-1} N_l$, with any well-chosen nonnegative weight function $\upsilon(\mathbf{x})$,[3] the optimal bandwidth is the minimizer of the truncated weighted integrated MSE (TWIMSE):

$$\{h_1^{opt}, h_2^{opt}, \ldots, h_d^{opt}\} = \arg\min_{\mathbf{h}} TWIMSE_j$$
$$= \arg\min_{\mathbf{h}} \int \left\{ (\mathbf{Bias}(\mathbf{x}))_j^2 + \frac{1}{nh_1h_2\cdots h_d} (\mathbf{\Sigma}(\mathbf{x}))_{j,j} \right\} \upsilon(\mathbf{x}) dx. \quad (3.40)$$

Or, if we would like to simultaneously choose the optimal bandwidth for the j_1th, j_2th, ..., j_Sth estimates in $\hat{\mathbf{b}}(\mathbf{x})$, let e_S be a $\mathcal{N}_p \times 1$ vector with the $j_1^{th}, j_2^{th}, \ldots, j_S^{th}$ elements equal to one and all other components equal to zero, $\mathbf{E}_S = diag(e_S)$, and the optimal bandwidth is the minimizer of the TWIMSE,

$$\left\{h_1^{opt}, h_2^{opt}, \ldots, h_d^{opt}\right\} = \arg\min_{\mathbf{h}} \int \left\{ (\mathbf{Bias}(\mathbf{x}))^\top \mathbf{E}_S \mathbf{Bias}(\mathbf{x}) + \frac{1}{nh_1h_2\cdots h_d} e_S^\top \mathbf{\Sigma}(\mathbf{x}) e_S \right\} \upsilon(\mathbf{x}) d\mathbf{x}$$
$$= \arg\min_{\mathbf{h}} \int \left\{ \sum_{s=1}^{S} (\mathbf{Bias}(\mathbf{x}))_{j_s}^2 + \frac{1}{nh_1h_2\cdots h_d} e_S^\top \mathbf{\Sigma}(\mathbf{x}) e_S \right\} \upsilon(\mathbf{x}) dx. \quad (3.41)$$

Note that $e_S^\top \mathbf{\Sigma}(\mathbf{x}) e_S$ does not equal to $\sum_{s=1}^{S} \text{avar}(\hat{\beta}_{j_s}(\mathbf{x}))$ here since $\mathbf{\Sigma}(\mathbf{x})$ is not diagonal.

Although the leading bias and variance expressions provide the basis for using some plug-in method to estimate leading bias and variance, as a referee correctly pointed out, plug-in computed smoothing parameters are stochastic and our Theorem 1 does not cover the case of stochastic smoothing parameters. However, under conditions similar to those given in Li and Zhou (2005), one can show that there exist unique (nonrandom) smoothing parameter values, say $h_{s,o}$, $s = 1, \ldots, q$, that minimize a truncated weighted integrated leading estimation mean squared error. Then one can further show that some feasible plug-in methods selected smoothing parameters will be asymptotically equivalent to the nonrandom optimal smoothing paramters in the sense that $\hat{h}_s / h_{s,o} \xrightarrow{p} 1$, where $\hat{h}_s$ is the plug-in method selected smoothing parameter, and $h_{s,o}$ is the nonrandom optimal smoothing parameter mentioned above. The asymptotic distribution of the local polynomial estimator (Theorem 1), when $h_{s,o}$ is replaced by $\hat{h}_s$, $s = 1, \ldots, q$, remains unchanged by following the stochastic equi-continuity arguments as in Li and Li (2010).

Note that in order to have the explicit leading bias term, one crucial assumption is to assume $(p+2)$ order of continuous differentiability as the prior knowledge: if this condition is violated, even though we include more refined approximated terms, we are not able to have the explicit leading bias terms, and then the plug-in bandwidth selection is not feasible for all the components from the local polynomial kernel estimators (we owe the discussion of this paragraph to a referee).

[3]The weight function $v(x)$ is nonnegative and has a bounded support. It also trims out the boundary regions when $\mathbf{X}$ has bounded support.

Next we compare the bias given in Eq. (3.33) and the bias implied by Masry (1996b) under weaker regularity conditions. According to Theorem 4 of Masry (1996b), and assuming $h_1 = \cdots = h_d = h = O\left(n^{-1/(d+2p+2)}\right)$ for notational simplicity,

$$\sqrt{nh^d}\left(\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) - \mathbf{Bias}_{L,M}(\mathbf{x})\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{\Sigma}(\mathbf{x})\right), \tag{3.42}$$

where

$$\mathbf{Bias}_{L,M}(\mathbf{x}) = h^{p+1}\mathscr{M}_p^{-1}\mathscr{B}_{p+1}\mathbf{m}_{p+1}(\mathbf{x}). \tag{3.43}$$

With $h_1 = \cdots = h_d = h = O(n^{-1/(d+2p+2)})$, the asymptotic bias given by Masry (1996b) is the same as the $O\left(\|\mathbf{h}\|^{p+1}\right)$ part in Eq. (3.33). That is, for each $j = g_k^{-1}(\mathbf{k}) + \sum_{l=1}^{k-1} N_l$, when $p - |\mathbf{k}|$ *is odd*, $\left(\mathbf{Bias}_{L,M}(\mathbf{x})\right)_j$ is non-zero, and is leading bias of $(\mathbf{k}!)^{-1}\left(\left(\widehat{D^{\mathbf{k}}m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}}m\right)(\mathbf{x})\right)$. But when $p - |\mathbf{k}|$ is even, $\left(\mathbf{Bias}_{L,M}(\mathbf{x})\right)_j$ is zero, and the leading bias of $(\mathbf{k}!)^{-1}\left(\left(\widehat{D^{\mathbf{k}}m}\right)(\mathbf{x}) - \left(D^{\mathbf{k}}m\right)(\mathbf{x})\right)$ should be the $O\left(\|\mathbf{h}\|^{p+2}\right)$ part in Eq. (3.33). However, Masry (1996b) did not give an explicit expression for the $O\left(\|\mathbf{h}\|^{p+2}\right)$ order term.

Again, the reason for the differences in asymptotic bias expressions are twofold. First, we assume that the unknown function $m(\cdot)$ has continuous derivatives of total order $p+2$, while Masry (1996b) only assumed that it has total order $p+1$ (Eq. (2.11) in Masry). Hence, Masry cannot obtain a bias term associated with $\mathbf{m}_{p+2}(\mathbf{x})$, because Masry did not assume that $m(\cdot)$ is $(p+2)$-time differentiable. Hence, Masry could not obtain the bias term $\mathscr{M}_p^{-1}\mathscr{B}_{p+2}\mathbf{H}^{(p+2)}\mathbf{m}_{p+2}(\mathbf{x})$, and this term becomes $o(h^{p+1})$ in his bias calculation. Second, we keep both $O(1)$ terms and the $O(\mathbf{h})$ terms in the expression of $\mathbf{S}_n(\mathbf{x})$ and $\mathbf{B}_{n,p+1}(\mathbf{x})$ (in Lemma 1), while Masry (1996b) only considered the leading $O(1)$ terms. As we have shown earlier that the bias term $\sum_{l=1}^{d} h_l \frac{f_l(\mathbf{x})}{f(\mathbf{x})}\left(\mathscr{M}_p^{-1}\mathscr{B}_{p+1}^l - \mathscr{M}_p^{-1}\mathscr{M}_p^l\mathscr{M}_p^{-1}\mathscr{B}_{p+1}\right)\mathbf{H}^{(p+1)}\mathbf{m}_{p+1}(\mathbf{x})$ in Eq. (3.33) is related to the $O(\mathbf{h})$ terms in the expression of $\mathbf{S}_n(\mathbf{x})$ and $\mathbf{B}_{n,p+1}(\mathbf{x})$. Since Masry (1996b) did not give the explicit expression for the terms of order $O(\mathbf{h})$ in the expression of $\mathbf{S}_n(\mathbf{x})$ and $\mathbf{B}_{n,p+1}(\mathbf{x})$, he has to put the term $\sum_{l=1}^{d} h_l \frac{f_l(\mathbf{x})}{f(\mathbf{x})}\left(\mathscr{M}_p^{-1}\mathscr{B}_{p+1}^l - \mathscr{M}_p^{-1}\mathscr{M}_p^l\mathscr{M}_p^{-1}\mathscr{B}_{p+1}\right)\mathbf{H}^{(p+1)}\mathbf{m}_{p+1}(\mathbf{x})$ as $o(h^{p+1})$.

Now we turn to the choice of bandwidth. Since for the whole vector of $\widehat{\beta}(\mathbf{x})$,

$$\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) = O\left(\|\mathbf{h}\|^{p+1} + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right), \tag{3.44}$$

Masry (1996b) chose the bandwidth $h_1 = \cdots = h_d = h = O(n^{-1/(d+2p+2)})$ to minimize the MSE of $\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x})$. However, for each $\mathbf{k}$, when $p - |\mathbf{k}|$ is even, the bandwidth chosen by Masry (1996b) would be smaller than the optimum $O\left(n^{-1/(d+2p+4)}\right)$, and the estimator with a smaller bandwidth under-smoothes the model and leads to a larger MSE compared with an estimator that uses optimal smoothing parameters.

Finally, it is easy to check that for the univariate case, our result is the same as given in Fan et al. (1996) as it should.

Corollary 1. *If $\{X_i\}$ is one-dimensional ($d = 1$), under Assumptions (a)–(d), assume that as $nh^{2p+1} \to \infty$ and $nh^{2p+7} \to 0$ as $n \to \infty$, we have*

$$\sqrt{nh}\left(\widehat{\beta}(x) - \beta(x) - \mathbf{Bias}(x)\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{\Sigma}(x)\right), \tag{3.45}$$

where

$$\begin{aligned}\mathbf{Bias}(x) = {} & h^{p+1}\mathcal{M}_p^{-1}\mathcal{B}_{p+1}m_{p+1}(x) + h^{p+2}\frac{f'(x)}{f(x)}\left(\mathcal{M}_p^{-1}\mathcal{B}_{p+2} - \mathcal{M}_p^{-1}\mathcal{M}_p^{1}\mathcal{M}_p^{-1}\mathcal{B}_{p+1}\right)m_{p+1}(x) \\ & + h^{p+2}\mathcal{M}_p^{-1}\mathcal{B}_{p+2}m_{p+2}(x),\end{aligned} \tag{3.46}$$

and

$$\mathbf{\Sigma}(x) = \frac{\sigma^2(x)}{f(x)}\mathcal{M}_p^{-1}\mathcal{G}_p\mathcal{M}_p^{-1}. \tag{3.47}$$

We can also rewrite the asymptotic bias as

$$\mathbf{Bias}(x) = h^{p+1}\left\{\mathcal{M}_p^{-1}\mathcal{B}_{p+1}m_{p+1}(x) + h\mathbf{Bias}^*(x)\right\}, \tag{3.48}$$

$$\mathbf{Bias}^*(x) = \frac{f'(x)m_{p+1}(x) + f(x)m_{p+2}(x)}{f(x)}\mathcal{M}_p^{-1}\mathcal{B}_{p+2} - \frac{f'(x)m_{p+1}(x)}{f(x)}\mathcal{M}_p^{-1}\mathcal{M}_p^{1}\mathcal{M}_p^{-1}\mathcal{B}_{p+1},$$

with some difference in notation, the asymptotic bias in Corollary 1 is the same with Eq. (2.2) in Theorem 1 of Fan et al. (1996).

The leading bias and variance expressions derived in this section apply to both interior and boundary x cases. In the next section we apply the result of Section 3 to derive the explicit leading bias expressions for local linear estimator.

4. LEADING BIASES OF THE LINEAR ESTIMATOR

In this section, we derive detailed bias expressions for multivariate local linear estimator.

4.1. The Multivariate Local Linear Estimator: When $\mathbf{x}$ is An Interior Point

We discuss the multivariate LLE when $\mathbf{x} \in \text{int}\,(\text{supp}(f))$. Consider the multivariate nonparametric regression model (3.1). As in (3.3), the local linear estimator ($p = 1$) is the minimizer of the following weighted least squares problem:

$$\left\{\hat{b}_0(\mathbf{x}), \hat{b}_{e_1}(\mathbf{x}), \ldots, \hat{b}_{e_d}(\mathbf{x})\right\} = \underset{b_0(\mathbf{x}), b_{e_1}(\mathbf{x}), \ldots, b_{e_d}(\mathbf{x})}{\arg\min} \sum_{i=1}^{n}\left\{Y_i - b_0(\mathbf{x}) - \sum_{k=1}^{d} b_{e_k}(\mathbf{x})\left(X_{i,k} - x_k\right)\right\}^2 K_{\mathbf{h},i\mathbf{x}}. \tag{4.1}$$

For notational simplicity, we denote $\hat{b}_{e_j}(\mathbf{x})$ by $\hat{b}_j(\mathbf{x})$, where e_j is a $d \times 1$ vector with jth element equal to one and others equal to zero, i.e., $\hat{b}_{e_j}(\mathbf{x}) \equiv \hat{b}_j(\mathbf{x}) \equiv \frac{\partial \widehat{m(\mathbf{x})}}{\partial x_j}$. $b_j(\mathbf{x})$, $\hat{\beta}_j(\mathbf{x})$, and $\beta_j(\mathbf{x})$ are similarly defined.

As in Eqs. (1.9)–(1.12) of Masry (1996b), the minimization of Eq. (4.1) leads to

$$t_{n,0}(\mathbf{x}) = \hat{\beta}_0(\mathbf{x}) s_{n,0}(\mathbf{x}) + \sum_{k=1}^{d} \hat{\beta}_k(\mathbf{x}) s_{n,e_k}(\mathbf{x}), \tag{4.2}$$

$$t_{n,e_j}(\mathbf{x}) = \hat{\beta}_0(\mathbf{x}) s_{n,e_j}(\mathbf{x}) + \sum_{k=1}^{d} \hat{\beta}_k(\mathbf{x}) s_{n,e_j+e_k}(\mathbf{x}), \tag{4.3}$$

or, in a matrix form,

$$\widehat{\beta}(\mathbf{x}) = (\mathbf{S}_n(\mathbf{x}))^{-1} \boldsymbol{\tau}_n(\mathbf{x}), \tag{4.4}$$

where $j \in \{1, \ldots, d\}$, e_j is a $d \times 1$ vector with jth element equal to one and others equal to zero, and the definition $s_{n,e_j+e_k}(\mathbf{x})$, $t_{n,e_j}(\mathbf{x})$, $\mathbf{S}_n(\mathbf{x})$, and $\tau_n(\mathbf{x})$ are in the Appendix. Recall that $\hat{\beta}_0(\mathbf{x}) = \hat{b}_0(\mathbf{x})$ and $\hat{\beta}_k(\mathbf{x}) = h_k \hat{b}_k(\mathbf{x})$, where $\hat{\beta}_k(\mathbf{x})$ ($\hat{b}_k(\mathbf{x})$), for $k = 1, \ldots, d$, is the kth component of the first derivative estimator $\widehat{\beta}_{\mathbf{1}}(\mathbf{x})$ ($\hat{\mathbf{b}}_{\mathbf{1}}(\mathbf{x})$).

Theorem 2. (The multivariate LLE, i.e., $d \geq 1$, $p = 1$). *For the multivariate local linear estimator, under Assumptions (a)–(d), as* $nh_1h_2 \cdots h_d \min\left\{h_1^2, h_2^2, \ldots, h_d^2\right\} \to \infty$ *and* $nh_1h_2 \cdots h_d \|\mathbf{h}\|^8 \to 0$ *as* $n \to \infty$, *we have the following situations:*

1.

$$\sqrt{nh_1h_2 \cdots h_d}\left(\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) - \mathbf{Bias}(\mathbf{x})\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}(\mathbf{x})\right), \tag{4.5}$$

where $Bias(\mathbf{x}) = [\mathbf{Bias}_0(\mathbf{x}), \mathbf{Bias}_1(\mathbf{x}), \ldots, \mathbf{Bias}_d(\mathbf{x})]^\top$,

$$\mathbf{Bias}_0(\mathbf{x}) = \sum_{l=1}^{d} h_l^2 \mu_2 m_{ll}(\mathbf{x}), \tag{4.6}$$

$$\mathbf{Bias}_k(\mathbf{x}) = h_k^3 \frac{\mu_4 - \mu_2^2}{\mu_2} \frac{f_k(\mathbf{x})}{f(\mathbf{x})} m_{kk}(\mathbf{x}) + h_k \sum_{l=1, l\neq k}^{d} h_l^2 \mu_2 \frac{f_l(\mathbf{x})}{f(\mathbf{x})} m_{kl}(\mathbf{x}) + h_k^3 \frac{\mu_4}{\mu_2} m_{kkk}(\mathbf{x}) + h_k \sum_{l=1, l\neq k}^{d} h_l^2 \mu_2 m_{kll}(\mathbf{x}), \tag{4.7}$$

for $k = 1, \ldots, d$, *where* $m_{kll}(\mathbf{x}) = \frac{1}{1!2!}\frac{\partial^3}{\partial x_k \partial x_l^2} m(\mathbf{x})$, $m_{kk}(\mathbf{x})$, $m_{kl}(\mathbf{x})$, *and* $m_{kkk}(\mathbf{x})$ *are similarly defined, and*

$$\mathbf{\Sigma}(\mathbf{x}) = \frac{\sigma^2(\mathbf{x})}{f(\mathbf{x})} \begin{bmatrix} \lambda_0^d & \mathbf{0}_{1\times d} \\ \mathbf{0}_{d\times 1} & \frac{\lambda_0^{d-1}\lambda_2}{\mu_2^2} I_d \end{bmatrix}. \tag{4.8}$$

2.

$$\sqrt{nh_1h_2\cdots h_d}\left(\hat{b}_0(\mathbf{x}) - b_0(\mathbf{x}) - \sum_{l=1}^{d} h_l^2 \mu_2 m_{ll}(\mathbf{x})\right) \xrightarrow{d} N\left(0, \frac{\lambda_0^d \sigma^2(\mathbf{x})}{f(\mathbf{x})}\right). \tag{4.9}$$

3.

$$\sqrt{nh_1h_2\cdots h_d}h_k\left(\hat{b}_k(\mathbf{x}) - b_k(\mathbf{x}) - \frac{1}{h_k}\mathbf{Bias}_k(\mathbf{x})\right) \xrightarrow{d} N\left(0, \frac{\lambda_0^{d-1}\lambda_2\sigma^2(\mathbf{x})}{\mu_2^2 f(\mathbf{x})}\right), \tag{4.10}$$

for $k = 1, \ldots, d$, *where* $\mathbf{Bias}_k$ *is defined in (4.7).*

The proof of Theorem 2 is in the Appendix. Again, there are two parts, the $O\left(\|\mathbf{h}\|^2\right)$ and the $O\left(\|\mathbf{h}\|^3\right)$ parts in the leading bias. For $\mathbf{Bias}_0(\mathbf{x})$, the bias of estimating the unknown function $m(\mathbf{x})$, the $O(\|\mathbf{h}\|^2)$ part is nonzero, and the $O\left(\|\mathbf{h}\|^3\right)$ part is negligible. While for $\mathbf{Bias}_k(\mathbf{x})$, the bias of estimating $\frac{1}{h_k}\frac{\partial m(\mathbf{x})}{\partial x_k}$ $(k = 1, .., d)$, the the coefficient associated with the $\|\mathbf{h}\|^2$ part is zero, and the $O\left(\|\mathbf{h}\|^3\right)$ part becomes the leading bias.

It can be shown that the asymptotic bias of $\hat{b}_0(\mathbf{x}) - b_0(\mathbf{x})$, the first component of the vector $\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x})$, derived above is consistent with equation (2.3) in Theorem 2.1 of Ruppert and Wand (1994).

The next corollary gives the optimal smoothing formulas.

Corollary 2.

1. *The optimal bandwidth minimizing the MSE of* $\hat{b}_0(\mathbf{x}) - b_0(\mathbf{x})$ *is* $h_l = O\left(n^{-1/(d+4)}\right)$ *for* $l \in \{1, \ldots, d\}$.
2. *The optimal bandwidth minimizing the MSE of* $\hat{b}_k(\mathbf{x}) - b_k(\mathbf{x})$ *for any* $k = 1, \ldots, d$ *is* $h_l = O\left(n^{-1/(d+6)}\right)$ *for* $l \in \{1, \ldots, d\}$.

To choose the bandwidth by the plug-in method, for any well-chosen nonnegative weight function $v(\mathbf{x})$, the optimal bandwidth for $m(\mathbf{x})$ is the minimizer of TWIMSE of $\hat{b}_0(\mathbf{x}) - b_0(\mathbf{x})$,

$$\left\{h_1^{opt}, h_2^{opt}, \ldots, h_d^{opt}\right\} = \arg\min_{\mathbf{h}} \int \left\{\left(\sum_{l=1}^{d} h_l^2 \mu_2 m_{ll}(\mathbf{x})\right)^2 + \frac{\lambda_0^d \sigma^2(\mathbf{x})}{nh_1h_2\cdots h_d f(\mathbf{x})}\right\} v(\mathbf{x})dx. \tag{4.11}$$

If the interest is to estimate the first derivative functions. One way to select the smoothing parameters is to minimize sum of TWIMSE over all the partial derivatives:

$$\begin{aligned}
&\{h_1^{opt}, h_2^{opt}, \ldots, h_d^{opt}\} \\
&= \arg\min_{\mathbf{h}} \int \sum_{k=1}^{d} \left\{ \left(h_k^3 \frac{\mu_4 - \mu_2^2}{\mu_2} \frac{f_k(\mathbf{x})}{f(\mathbf{x})} m_{kk}(\mathbf{x}) + h_k \sum_{l=1, l\neq k}^{d} h_l^2 \mu_2 \frac{f_l(\mathbf{x})}{f(\mathbf{x})} m_{kl}(\mathbf{x}) \right.\right. \\
&\quad \left.\left. + h_k^3 \frac{\mu_4}{\mu_2} m_{kkk}(\mathbf{x}) + h_k \sum_{l=1, l\neq k}^{d} h_l^2 \mu_2 m_{kll}(\mathbf{x}) \right)^2 + \frac{\lambda_0^{d-1} \lambda_2 \sigma^2(\mathbf{x})}{n h_1 h_2 \cdots h_d h_k^2 \mu_2^2 f(\mathbf{x})} \right\} \upsilon_k(\mathbf{x}) d\mathbf{x},
\end{aligned} \tag{4.12}$$

where $\upsilon_k(\mathbf{x})$ is a weight function, $k = 1, \ldots, d$. Note the right-hand side of (4.12) is the sum of truncated, weighted MSE over each component k.

From Eq. (4.12) it is easy to see that the optimal smoothing parameter has an order of $n^{-1/(d+6)}$. Let $h_s = \sqrt{a_s}\, n^{-1/(d+6)}$. then the minimization problem of Eq. (4.12) is equivalent to

$$\begin{aligned}
\{a_1^0, a_2^0, \ldots, a_d^0\} = \arg\min_{a_1,\ldots,a_d} \int \sum_{k=1}^{d} \left\{ \left(a_k^3 \frac{\mu_4 - \mu_2^2}{\mu_2} \frac{f_k(\mathbf{x})}{f(\mathbf{x})} m_{kk}(\mathbf{x}) + a_k \sum_{l=1, l\neq k}^{d} a_l^2 \mu_2 \frac{f_l(\mathbf{x})}{f(\mathbf{x})} m_{kl}(\mathbf{x}) \right.\right. \\
\left.\left. + a_k^3 \frac{\mu_4}{\mu_2} m_{kkk}(\mathbf{x}) + a_k \sum_{l=1, l\neq k}^{d} a_l^2 \mu_2 m_{kll}(\mathbf{x}) \right)^2 + \frac{\lambda_0^{d-1} \lambda_2 \sigma^2(\mathbf{x})}{n a_1 a_2 \cdots a_d a_k^2 \mu_2^2 f(\mathbf{x})} \right\} \upsilon_k(\mathbf{x}) d\mathbf{x}.
\end{aligned} \tag{4.13}$$

Using the similar arguments as in Li and Zhou (2005), one can derive conditions under which (4.13) leads to unique solutions for $\{a_1^0, a_2^0, \ldots, a_d^0\}$.

Again, the asymptotic bias obtained here is different from Masry (1996b). According to Theorem 4 of Masry (1996b), with $h_1 = \cdots = h_d = h = O(n^{-1/(d+2p+2)}) = O(n^{-1/(d+4)})$,

$$\begin{aligned}
\mathbf{Bias}_{L,M}(\mathbf{x}) &= h^{p+1} \mathscr{M}_p^{-1} \mathscr{B}_{p+1} \mathbf{m}_{p+\mathbf{1}}(\mathbf{x}) \\
&= h^2 \begin{bmatrix} 1 & \mathbf{0}_{1\times d} \\ \mathbf{0}_{d\times 1} & \mu_2 I_d \end{bmatrix}^{-1} \begin{bmatrix} \mu_2 & 0 & \cdots & \mu_2 \\ \mathbf{0}_{d\times 1} & \mathbf{0}_{d\times 1} & \cdots & \mathbf{0}_{d\times 1} \end{bmatrix} \begin{bmatrix} m_{11}(\mathbf{x}) \\ m_{12}(\mathbf{x}) \\ \vdots \\ m_{dd}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} h^2 \mu_2 \sum_{l=1}^{d} m_{ll}(\mathbf{x}) \\ \mathbf{0}_{d\times 1}. \end{bmatrix}.
\end{aligned} \tag{4.14}$$

That is, according to the formula of Masry (1996b),

$$\hat{b}_0(\mathbf{x}) - b_0(\mathbf{x}) = h^2\mu_2 \sum_{l=1}^{d} m_{ll}(\mathbf{x}) + o_p\left(h^2\right) + O_p\left(\frac{1}{\sqrt{nh^d}}\right), \tag{4.15}$$

$$\hat{b}_k(\mathbf{x}) - b_k(\mathbf{x}) = o_p(h) + O_p\left(\frac{1}{\sqrt{nh^d}}\right) \quad \text{for} \quad k = 1, \ldots, d.$$

The leading bias of $\hat{b}_k(\mathbf{x}) - b_k(\mathbf{x})$, $k \in \{1, \ldots, d\}$, is not derived in Masry (1996b). For $\hat{b}_k(\mathbf{x}) - b_k(\mathbf{x})$, $k \in \{1, \ldots, d\}$, the $O(\|\mathbf{h}\|^2)$ part in the leading bias is zero, and the leading bias should be the $O(\|\mathbf{h}\|^3)$ part. However, the $O(\|\mathbf{h}\|^3)$ part in the leading bias was not given by Masry (1996b).

For $\hat{b}_k(\mathbf{x})$, $k \in \{1, \ldots, d\}$, the bandwidth for the multivariate local linear estimator chosen by Masry (1996b), $h = O(n^{-1/(d+2p+2)}) = O(n^{-1/(d+4)})$, which is smaller than the optimum rate $O(n^{-1/(d+6)})$.

For the univariate case, our result is the same with the result from the formula of Fan et al. (1996), and the leading bias of $\hat{m}(\cdot)$, the first component of $\hat{\beta}(\mathbf{x})$, is the same with Ruppert and Wand (1994).

Corollary 3 (The univariate LLE, i.e., $d = 1$, $p = 1$)**.** *For the univariate local linear estimator, under Assumptions (a)–(d), assume that $nh^3 \to \infty$ and $nh^9 \to 0$ as $n \to \infty$. We have*

$$\sqrt{nh}\begin{bmatrix} 1 & 0 \\ 0 & h \end{bmatrix}(\hat{\mathbf{b}}(x) - \mathbf{b}(x) - \mathbf{Bias}(x)) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}(x)), \tag{4.16}$$

where

$$\mathbf{Bias}(x) = \begin{bmatrix} \dfrac{1}{2}h^2\mu_2\dfrac{d^2m(x)}{dx^2} \\ \dfrac{1}{2}h^3\dfrac{\mu_4 - \mu_2^2}{\mu_2}\dfrac{f'(x)}{f(x)}\dfrac{d^2m(x)}{dx^2} + \dfrac{1}{6}h^3\dfrac{\mu_4}{\mu_2}\dfrac{d^3m(x)}{dx^3} \end{bmatrix}, \tag{4.17}$$

and

$$\mathbf{\Sigma}(x) = \frac{\sigma^2(x)}{f(x)}\begin{bmatrix} \lambda_0 & 0 \\ 0 & \frac{\lambda_2}{\mu_2^2} \end{bmatrix}. \tag{4.18}$$

4.2. The Multivariate Local Linear Estimator: When $\mathbf{x}$ is a Boundary Point

Here we discuss the multivariate LLE when $\mathbf{x}$ is a boundary point. Without loss of generality, suppose that $\text{supp}(f) = \mathbb{R}_+^{d_1} \times \mathbb{R}^{d_2}$, where $d_1 + d_2 = d$, and $K(\cdot)$ is a kernel function with a bounded support $\text{supp}(K) = [-1, 1]$. Let $x = (\mathbf{x}_1, \mathbf{x}_2)$, $\mathbf{x}_1 = \{c_1h_1, \ldots c_{d_1}h_{d_1}\} \in \mathbb{R}_+^{d_1}$ with $0 \le c_1 \le 1, \ldots 0 \le c_{d_1} \le 1$, and $\mathbf{x}_2 \in \mathbb{R}^{d_2}$.

It is easy to show that when $\mathbf{x}_1$ is in the boundary,

$$\mathbf{S}_n(\mathbf{x}) = \mathscr{M}_1(\mathbf{c}) f(\mathbf{x}) + O_p\left(\|\mathbf{h}\| + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right), \tag{4.19}$$

$$\mathbf{B}_{n,2}(\mathbf{x}) = \mathscr{B}_2(\mathbf{c}) f(\mathbf{x}) + O_p\left(\|\mathbf{h}\| + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right), \tag{4.20}$$

where $\mathscr{M}_1(\mathbf{c})$ is a$(d+1)\times(d+1)$ matrix:

$$\mathscr{M}_1(\mathbf{c}) = \mathscr{M}_1\left(c_1, c_2, \ldots, c_{d_1}\right) = \begin{bmatrix} \mathscr{M}_1^{(0,0)}(\mathbf{c}) & \mathscr{M}_1^{(0,1)}(\mathbf{c}) \\ \mathscr{M}_1^{(1,0)}(\mathbf{c}) & \mathscr{M}_1^{(1,1)}(\mathbf{c}) \end{bmatrix},$$

in which $\mathscr{M}_1^{(0,0)}(\mathbf{c})$ is a scalar equal to

$$\begin{aligned}
\mathscr{M}_1^{(0,0)}(\mathbf{c}) &= \int_{c_1}^{\infty}\int_{c_2}^{\infty}\cdots\int_{c_{d_1}}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} k(u_1)\cdot k(u_2)\cdot\cdots\cdot k(u_d)\, du_1 du_2\cdots du_d \\
&= \int_{c_1}^{\infty} k(u_1)\, du_1 \cdot \int_{c_2}^{\infty} k(u_2)\, du_2 \cdot\cdots\cdot \int_{c_{d_1}}^{\infty} k\left(u_{d_1}\right) du_{d_1} \\
&\quad \cdot \int_{-\infty}^{\infty} k\left(u_{d_1+1}\right) du_{d_1+1} \cdot \int_{-\infty}^{\infty} k\left(u_{d_1+2}\right) du_{d_1+2} \cdot\cdots\cdot \int_{c_{d_2}}^{\infty} k\left(u_{d_2}\right) du_{d_2} \\
&= \mu_0(c_1)\cdot\cdots\cdot\mu_0\left(c_{d_1}\right),
\end{aligned}$$

$\mathscr{M}_1^{(1,0)}(\mathbf{c}) = \left(\mathscr{M}_1^{(0,1)}(\mathbf{c})\right)^{\top}$ is a $d\times 1$ vector with the ith element of $\mathscr{M}_1^{(1,0)}(\mathbf{c})$ equal to

$$\begin{aligned}
\mathscr{M}_1^{(1,0),i}(\mathbf{c}) &= \int_{c_1}^{\infty}\int_{c_2}^{\infty}\cdots\int_{c_{d_1}}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} u_i k(u_1)\cdot k(u_2)\cdot\cdots\cdot k(u_d)\, du_1 du_2\cdots du_d \\
&= \begin{cases} \mu_0(c_1)\cdot\cdots\cdot\mu_0(c_{i-1})\cdot\mu_1(c_i)\cdot\mu_0(c_{i+1})\cdot\cdots\cdot\mu_0\left(c_{d_1}\right) & \\ = \mathscr{M}_1^{(0,0)}(\mathbf{c})\cdot\mu_1(c_i)\Big/\mu_0(c_i), & \text{if } 1\le i\le d_1, \\ 0, & \text{if } d_1+1\le i\le d_2, \end{cases}
\end{aligned}$$

and $\mathscr{M}_1^{(1,1)}(\mathbf{c})$ is a $d\times d$ matrix with the (i,j) th element of $\mathscr{M}_1^{(1,1)}(\mathbf{c})$ equal to

$$\begin{aligned}
\mathscr{M}_1^{(1,0),(i,j)}(\mathbf{c}) &= \int_{c_1}^{\infty}\int_{c_2}^{\infty}\cdots\int_{c_{d_1}}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} u_i u_j k(u_1)\cdot k(u_2)\cdot\cdots\cdot k(u_d)\, du_1 du_2\cdots du_d \\
&= \begin{cases} \mathscr{M}_1^{(0,0)}(\mathbf{c})\cdot\mu_2(c_i)/\mu_0(c_i), & \text{if } 1\le i=j\le d_1, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c})\cdot\mu_1(c_i)\cdot\mu_1(c_j)/(\mu_0(c_i)\cdot\mu_0(c_j)), & \text{if } 1\le i\le d_1, 1\le j\le d_1, i\ne j, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c})\cdot\mu_2/\mu_0, & \text{if } d_1+1\le i=j\le d_2, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

Also $\mathscr{B}_2(\mathbf{c})$ is a $(d+1) \times N_2 = (d+1) \times \left(\binom{d}{1} + \binom{d}{2}\right)$ matrix:

$$\mathscr{B}_2(\mathbf{c}) = \mathscr{B}_2\left(c_1, c_2, \ldots, c_{d_1}\right) = \begin{bmatrix} \mathscr{B}_2^{(0)}(\mathbf{c}) \\ \mathscr{B}_2^{(1)}(\mathbf{c}) \end{bmatrix},$$

in which $\mathscr{B}_2^{(0)}(\mathbf{c})$ is a $1 \times N_2$ row vector with the ith element of $\mathscr{B}_2^{(0)}(\mathbf{c})$ equal to

$$\mathscr{B}_2^{(0),i}(\mathbf{c}) = \begin{cases} \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_2(c_k)/\mu_0(c_k), \\ \quad \text{if } g_2(i) = \{0, \ldots, 0, \underset{k\text{th}}{2}, 0, \ldots, 0\}, 1 \le k \le d_1, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_1(c_{k_1}) \cdot \mu_1(c_{k_2})/(\mu_0(c_{k_1}) \cdot \mu_0(c_{k_2})), \\ \quad \text{if } g_2(i) = \{0, \ldots, 0, \underset{k_1\text{th}}{1}, 0, \ldots, 0, \underset{k_2\text{th}}{1}, 0, \ldots, 0\}, 1 \le k_1 < k_2 \le d_1, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_2/\mu_0 = \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_2, \\ \quad \text{if } g_2(i) = \{0, \ldots, 0, \underset{k\text{th}}{2}, 0, \ldots, 0\}, d_1 + 1 \le k \le d_2, \\ 0, \quad \text{otherwise}, \end{cases}$$

and $\mathscr{B}_2^{(1)}(\mathbf{c})$ is a $d \times N_2$ matrix with the (i,j) th element of $\mathscr{B}_2^{(1)}(\mathbf{c})$ equal to

$$\mathscr{B}_2^{(1),(i,j)}(\mathbf{c}) = \begin{cases} \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_3(c_k)/\mu_0(c_k), \\ \quad \text{if } g_2(j) = \{0, \ldots, 0, \underset{k\text{th}}{2}, 0, \ldots, 0\}, 1 \le i = k \le d_1, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_1(c_i) \cdot \mu_2(c_k)/(\mu_0(c_i) \cdot \mu_0(c_k)), \\ \quad \text{if } g_2(j) = \{0, \ldots, 0, \underset{k\text{th}}{2}, 0, \ldots, 0\}, 1 \le i \le d_1, 1 \le k \le d_1, i \ne k, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_2(c_{k_1}) \cdot \mu_1(c_{k_2})/(\mu_0(c_{k_1}) \cdot \mu_0(c_{k_2})), \\ \quad \text{if } g_2(j) = \{0, \ldots, 0, \underset{k_1\text{th}}{1}, 0, \ldots, 0, \underset{k_2\text{th}}{1}, 0, \ldots, 0\}, 1 \le i = k_1 < k_2 \le d_1, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_1(c_{k_1}) \cdot \mu_2(c_{k_2})/(\mu_0(c_{k_1}) \cdot \mu_0(c_{k_2})), \\ \quad \text{if } g_2(j) = \{0, \ldots, 0, \underset{k_1\text{th}}{1}, 0, \ldots, 0, \underset{k_2\text{th}}{1}, 0, \ldots, 0\}, 1 \le k_1 < i = k_2 \le d_1, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_1(c_i) \cdot \mu_2/\mu_0(c_i), \\ \quad \text{if } g_2(j) = \{0, \ldots, 0, \underset{k\text{th}}{2}, 0, \ldots, 0\}, 1 \le i \le d_1, d_1 + 1 \le k \le d_2, \\ \mathscr{M}_1^{(0,0)}(\mathbf{c}) \cdot \mu_1(c_{k_1}) \cdot \mu_2/\mu_0(c_{k_1}), \\ \quad \text{if } g_2(j) = \{0, \ldots, 0, \underset{k_1\text{th}}{1}, 0, \ldots, 0, \underset{k_2\text{th}}{1}, 0, \ldots, 0\}, 1 \le k_1 \le d_1, d_1 + 1 \le i = k \le d_2, \\ 0, \quad \text{otherwise}, \end{cases}$$

where $g_2(\cdot)$ is defined in section 3. Since every element of $(\mathscr{M}_1(\mathbf{c}))^{-1} \mathscr{B}_2(\mathbf{c}) \mathbf{H}^{(2)} \mathbf{m}_2(\mathbf{x})$ does not equal to zero, the leading bias of $\widehat{\beta}(\mathbf{x})$ is $(\mathscr{M}_1(\mathbf{c}))^{-1} \mathscr{B}_2(\mathbf{c}) \mathbf{H}^{(2)} \mathbf{m}_2(\mathbf{x})$.

It does not seem possible to obtain a further simple expression for the leading bias for the general multivariate regression case. Below we will consider the simple case of $d = 2$.

We consider two illustrative bivariate ($d = 2$) examples. First, $d_1 = 2$ and $d_2 = 0$, i.e., $\text{supp}(f) = \mathbb{R}_+ \times \mathbb{R}_+$, and $K(\cdot)$ is a kernel function with a bounded support $\text{supp}(K) = [-1, 1]$. Let $x = (x_1, x_2)$, $x_1 = c_1 h_1$, $x_2 = c_2 h_2$ with $0 \le c_1 \le 1$, and $0 \le c_2 \le 1$. Then

$$\mathscr{M}_1(c_1, c_2) = \begin{bmatrix} \mu_0(c_1)\mu_0(c_2) & \mu_1(c_1)\mu_0(c_2) & \mu_0(c_1)\mu_1(c_2) \\ \mu_1(c_1)\mu_0(c_2) & \mu_2(c_1)\mu_0(c_2) & \mu_1(c_1)\mu_1(c_2) \\ \mu_0(c_1)\mu_1(c_2) & \mu_1(c_1)\mu_1(c_2) & \mu_0(c_1)\mu_2(c_2) \end{bmatrix},$$

$$\mathscr{B}_2(c_1, c_2) = \begin{bmatrix} \mu_2(c_1)\mu_0(c_2) & \mu_1(c_1)\mu_1(c_2) & \mu_0(c_1)\mu_2(c_2) \\ \mu_3(c_1)\mu_0(c_2) & \mu_2(c_1)\mu_1(c_2) & \mu_1(c_1)\mu_2(c_2) \\ \mu_2(c_1)\mu_1(c_2) & \mu_1(c_1)\mu_2(c_2) & \mu_0(c_1)\mu_3(c_2) \end{bmatrix},$$

in which $\mu_j(c) = \int_{-c}^{\infty} K(v) v^j dv$. In light of Theorem 1, the $O\left(\|\mathbf{h}\|^{p+1}\right) = O\left(\|\mathbf{h}\|^2\right)$ part of the leading bias is

$$(\mathscr{M}_1(c_1, c_2))^{-1} \mathscr{B}_2(c_1, c_2) \mathbf{H}^{(2)} \mathbf{m_2}(\mathbf{x})$$

which is not zero. Therefore, the leading bias of $\widehat{\beta}(\mathbf{x})$ is

$$\mathbf{Bias}_0(\mathbf{x}) = \frac{1}{M(c_1, c_2)} \left\{ h_1^2 B_{1,1}(c_1, c_2) m_{11}(\mathbf{x}) + h_1 h_2 B_{1,2}(c_1, c_2) m_{12}(\mathbf{x}) + h_2^2 B_{1,3}(c_1, c_2) m_{22}(\mathbf{x}) \right\},$$

$$\mathbf{Bias}_1(\mathbf{x}) = \frac{1}{M(c_1, c_2)} \left\{ h_1^2 B_{2,1}(c_1, c_2) m_{11}(\mathbf{x}) + h_1 h_2 B_{2,2}(c_1, c_2) m_{12}(\mathbf{x}) \right\},$$

$$\mathbf{Bias}_2(\mathbf{x}) = \frac{1}{M(c_1, c_2)} \left\{ h_1 h_2 B_{3,2}(c_1, c_2) m_{12}(\mathbf{x}) + h_2^2 B_{3,3}(c_1, c_2) m_{22}(\mathbf{x}) \right\},$$

where $m_{ll}(\mathbf{x}) = (1/2)\partial^2 m(\mathbf{x})/\partial x_l^2$, $m_{lk}(\mathbf{x}) = \partial^2 m(\mathbf{x})/\partial x_l \partial x_k$ for $l \ne k$,

$$\begin{aligned} M(c_1, c_2) &= \mu_0(c_1)\mu_0(c_2) \left[\mu_0(c_1)\mu_2(c_1)\mu_0(c_2)\mu_2(c_2) - (\mu_1(c_1))^2 (\mu_1(c_2))^2 \right] \\ &\quad + \mu_0(c_1) (\mu_1(c_1))^2 \mu_0(c_2) \left[(\mu_1(c_2))^2 - \mu_0(c_2)\mu_2(c_2) \right] \\ &\quad + \mu_0(c_1)\mu_0(c_2) (\mu_1(c_2))^2 \left[(\mu_1(c_1))^2 - \mu_0(c_1)\mu_2(c_1) \right], \\ B_{1,1}(c_1, c_2) &= \mu_2(c_1)\mu_0(c_2) \left[\mu_0(c_1)\mu_2(c_1)\mu_0(c_2)\mu_2(c_2) - (\mu_1(c_1))^2 (\mu_1(c_2))^2 \right] \\ &\quad + \mu_0(c_1)\mu_1(c_1)\mu_3(c_1)\mu_0(c_2) \left[(\mu_1(c_2))^2 - \mu_0(c_2)\mu_2(c_2) \right] \\ &\quad + \mu_2(c_1)\mu_0(c_2) (\mu_1(c_2))^2 \left[(\mu_1(c_1))^2 - \mu_0(c_1)\mu_2(c_1) \right], \\ B_{1,2}(c_1, c_2) &= \mu_1(c_1)\mu_1(c_2) \left[\mu_0(c_1)\mu_2(c_1)\mu_0(c_2)\mu_2(c_2) - (\mu_1(c_1))^2 (\mu_1(c_2))^2 \right] \\ &\quad + \mu_0(c_1)\mu_1(c_1)\mu_2(c_1)\mu_1(c_2) \left[(\mu_1(c_2))^2 - \mu_0(c_2)\mu_2(c_2) \right] \\ &\quad + \mu_0(c_1)\mu_0(c_2)\mu_1(c_2)\mu_2(c_2) \left[(\mu_1(c_1))^2 - \mu_0(c_1)\mu_2(c_1) \right], \end{aligned}$$

$$B_{1,3}(c_1,c_2) = \mu_0(c_1)\mu_2(c_2)\left[\mu_0(c_1)\mu_2(c_1)\mu_0(c_2)\mu_2(c_2) - (\mu_1(c_1))^2(\mu_1(c_2))^2\right] + \mu_0(c_1)(\mu_1(c_1))^2\mu_2(c_2)\left[(\mu_1(c_2))^2 - \mu_0(c_2)\mu_2(c_2)\right] + \mu_0(c_1)\mu_0(c_2)\mu_1(c_2)\mu_3(c_2)\left[(\mu_1(c_1))^2 - \mu_0(c_1)\mu_2(c_1)\right],$$

$$B_{2,1}(c_1,c_2) = \mu_0(c_1)\mu_0(c_2)\left[\mu_1(c_1)\mu_2(c_1) - \mu_0(c_1)\mu_3(c_1)\right]\left[(\mu_1(c_2))^2 - \mu_0(c_2)\mu_2(c_2)\right],$$

$$B_{2,2}(c_1,c_2) = \mu_0(c_1)\mu_1(c_2)\left[(\mu_1(c_1))^2 - \mu_0(c_1)\mu_2(c_1)\right]\left[(\mu_1(c_2))^2 - \mu_0(c_2)\mu_2(c_2)\right],$$

$$B_{3,2}(c_1,c_2) = \mu_1(c_1)\mu_0(c_2)\left[(\mu_1(c_1))^2 - \mu_0(c_1)\mu_2(c_1)\right]\left[(\mu_1(c_2))^2 - \mu_0(c_2)\mu_2(c_2)\right],$$

$$B_{3,3}(c_1,c_2) = \mu_0(c_1)\mu_0(c_2)\left[(\mu_1(c_1))^2 - \mu_0(c_1)\mu_2(c_1)\right]\left[\mu_1(c_2)\mu_2(c_2) - \mu_0(c_2)\mu_3(c_2)\right].$$

Second, $d_1 = 1$ and $d_2 = 1$, i.e., $\mathrm{supp}(f) = \mathbb{R}_+ \times \mathbb{R}$, and $K(\cdot)$ is a kernel function with a bounded support $\mathrm{supp}(K) = [-1,1]$. Let $x = (x_1, x_2)$, $x_1 = c_1 h_1$ with $0 \le c_1 \le 1$, and $x_2 \in \mathrm{int}(f)$. Then

$$\mathscr{M}_1(c_1) = \begin{bmatrix} \mu_0(c_1) & \mu_1(c_1) & 0 \\ \mu_1(c_1) & \mu_2(c_1) & 0 \\ 0 & 0 & \mu_0(c_1)\mu_2 \end{bmatrix}, \quad \mathscr{B}_2(c_1) = \begin{bmatrix} \mu_2(c_1) & 0 & \mu_0(c_1)\mu_2 \\ \mu_3(c_1) & 0 & \mu_1(c_1)\mu_2 \\ 0 & \mu_1(c_1)\mu_2 & 0 \end{bmatrix},$$

in which $\mu_j(c) = \int_{-c}^{\infty} K(v)v^j dv$. In light of Theorem 1, the $O\left(\|\mathbf{h}\|^{p+1}\right) = O\left(\|\mathbf{h}\|^2\right)$ part of the leading bias is

$$(\mathscr{M}_1(c_1))^{-1}\mathscr{B}_2(c_1)\mathbf{H}^{(2)}\mathbf{m_2}(\mathbf{x}),$$

which is not zero. Therefore, the leading bias of $\widehat{\beta}(\mathbf{x})$ is

$$\mathbf{Bias}_0(\mathbf{x}) = h_1^2 \frac{(\mu_2(c_1))^2 - \mu_1(c_1)\mu_3(c_1)}{\mu_0(c_1)\mu_2(c_1) - (\mu_1(c_1))^2} m_{11}(\mathbf{x}) + h_2^2\mu_2 m_{22}(\mathbf{x}),$$

$$\mathbf{Bias}_1(\mathbf{x}) = h_1^2 \frac{\mu_0(c_1)\mu_3(c_1) - \mu_1(c_1)\mu_2(c_1)}{\mu_0(c_1)\mu_2(c_1) - (\mu_1(c_1))^2} m_{11}(\mathbf{x}),$$

$$\mathbf{Bias}_2(\mathbf{x}) = h_1 h_2 \frac{\mu_1(c_1)}{\mu_0(c_1)} m_{12}(\mathbf{x}).$$

5. CONCLUDING REMARKS

Masry (1996a,b) considered a general setup of a multivariate local polynomial regression model and derived the asymptotic distribution of the LPE. Based on Masry (1996b) result and by strengthening the smoothness condition imposed on the unknown regression function, in this article we provide formulas for deriving the leading bias in a multivariate local polynomial regression model. We give detailed results for the leading bias for the

LLE. For higher order local polynomial estimators ($p \geq 2$), the leading bias could be directly derived from Theorem 1.

Based on the explicitly leading bias and variance expressions derived in this article and following the approach of Li and Zhou (2005), one can further analyze the uniqueness of the optimal smoothing parameter selection. Asymptotic normality follows from the result of Masry (1996a). By combining the result of Masry (1996a) with our derived leading bias term, the asymptotic normality result allows for optimally selected smoothing parameters for all components of the vector of the LPE. By using the stochastic equicontinuity arguments as in Li and Li (2010), one can show that the asymptotic normality result remains valid when the nonrandom smoothing parameters are replaced by some asymptotically equivalent stochastic smoothing parameters.

In this article, we only consider the i.i.d. case for notational simplicity, given that the bias calculation only relies on the identical distribution assumption; hence, the leading bias remains exactly the same whether the data is i.i.d. or weakly (strictly) stationary process. Given the results of Masry (1996a,b) who consider the time series weakly dependent data for LPE. The results of this article hold true for weakly dependent data case. Also, we only consider the case that $\mathbf{H}$ is a diagonal bandwidth. As discussed in Wand and Jones (1993) and Ruppert and Wand (1994), there are situations where using a nondiagonal bandwidth matrix is advantageous, e.g., as addressed in Wand and Jones (1993) in the density estimation context. However, an explicit treatment of nondiagonal $\mathbf{H}$ matrix in our general multivariate local polynomial case will be quite complex. This case deserves a separate treatment and is beyond the scope of the present article.

6. APPENDIX: PROOFS OF THEOREM 2 AND THEOREM 3

6.1. Proof of Theorem 2

Proof. $\mathbf{S}_n(\mathbf{x})$ and $\boldsymbol{\tau}_n(\mathbf{x})$ in Eq. (4.4) are defined as in Eq. (2.2) and (2.5) of Masry (1996b). Let

$$t_{n,e_j}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}}\left(\frac{\mathbf{X}_{i,j}-\mathbf{x}_j}{h_j}\right) Y_i, \tag{6.1}$$

$$s_{n,e_j+e_k}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}}\left(\frac{\mathbf{X}_{i,j}-\mathbf{x}_j}{h_j}\right)\left(\frac{\mathbf{X}_{i,k}-\mathbf{x}_k}{h_k}\right), \tag{6.2}$$

and

$$\boldsymbol{\tau}_n(\mathbf{x}) = \begin{bmatrix} t_{n,0}(\mathbf{x}) \\ t_{n,e_1}(\mathbf{x}) \\ \vdots \\ t_{n,e_d}(\mathbf{x}) \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \begin{bmatrix} 1 \\ \frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1} \\ \vdots \\ \frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d} \end{bmatrix} Y_i, \tag{6.3}$$

$$\mathbf{S}_n(\mathbf{x}) = \begin{bmatrix} s_{n,0}(\mathbf{x}) & s_{n,e_1}(\mathbf{x}) & \cdots & s_{n,e_d}(\mathbf{x}) \\ s_{n,e_1}(\mathbf{x}) & s_{n,e_1+e_1}(\mathbf{x}) & \cdots & s_{n,e_1+e_d}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ s_{n,e_d}(\mathbf{x}) & s_{n,e_1+e_d}(\mathbf{x}) & \cdots & s_{n,e_d+e_d}(\mathbf{x}) \end{bmatrix} \tag{6.4}$$

$$= \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \begin{bmatrix} 1 & \frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1} & \cdots & \frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d} \\ \frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1} & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^2 & \cdots & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)\left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right) \\ \vdots & \vdots & & \vdots \\ \frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d} & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)\left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right) & \cdots & \left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right)^2 \end{bmatrix}.$$

Following Eqs. (2.9) and (2.10) in Masry (1996b), let

$$t^*_{n,e_j}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \left(\frac{\mathbf{X}_{i,j}-\mathbf{x}_j}{h_j}\right) u_i. \tag{6.5}$$

Then we have

$$t_{n,e_j}(\mathbf{x}) - t^*_{n,e_j}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \left(\frac{\mathbf{X}_{i,j}-\mathbf{x}_j}{h_j}\right) m(\mathbf{X}_i). \tag{6.6}$$

Approximating $m(\mathbf{X}_i)$ by the Taylor expansion up to its third order derivatives,

$$m(\mathbf{X}_i) = \sum_{0\le|\mathbf{k}|\le 3} \frac{1}{\mathbf{k}!}\left(D^{\mathbf{k}}m\right)(\mathbf{x})\,(\mathbf{X}_i-\mathbf{x})^{\mathbf{k}} + o_p\left(\|\mathbf{h}\|^3\right) \tag{6.7}$$

$$= \sum_{0\le|\mathbf{k}|\le 1} \beta_{\mathbf{k}}(\mathbf{x})[\mathbf{H}^{-1}(\mathbf{X}_i-\mathbf{x})]^{\mathbf{k}} + \sum_{2\le|\mathbf{k}|\le 3} \frac{1}{\mathbf{k}!}\left(D^{\mathbf{k}}m\right)(\mathbf{x})\,(\mathbf{X}_i-\mathbf{x})^{\mathbf{k}} + o_p\left(\|\mathbf{h}\|^3\right).$$

and by substituting Eq. (4.2), Eq. (4.3), and Eq. (6.7) into Eq. (6.6), we find,

$$t^*_{n,0}(\mathbf{x}) = \left(\hat{\beta}_0(\mathbf{x}) - \beta_0(\mathbf{x})\right) s_{n,0}(\mathbf{x}) + \sum_{k=1}^{d}\left(\hat{\beta}_k(\mathbf{x}) - \beta_k(\mathbf{x})\right) s_{n,e_k}(\mathbf{x}) \tag{6.8}$$

$$- \sum_{2\le|\mathbf{k}|\le 3} \frac{\mathbf{h}^{\mathbf{k}}}{\mathbf{k}!}\left(D^{\mathbf{k}}m\right)(\mathbf{x}) s_{n,e_k}(\mathbf{x}) + o_p\left(\|\mathbf{h}\|^3\right) s_{n,0}(\mathbf{x}),$$

$$t^*_{n,e_j}(\mathbf{x}) = \left(\hat{\beta}_0(\mathbf{x}) - \beta_0(\mathbf{x})\right) s_{n,e_j}(\mathbf{x}) + \sum_{k=1}^{d}\left(\hat{\beta}_k(\mathbf{x}) - \beta_k(\mathbf{x})\right) s_{n,e_j+e_k}(\mathbf{x}) \tag{6.9}$$

$$- \sum_{2\le|\mathbf{k}|\le 3} \frac{\mathbf{h}^{\mathbf{k}}}{\mathbf{k}!}\left(D^{\mathbf{k}}m\right)(\mathbf{x}) s_{n,e_j+e_k}(\mathbf{x}) + o_p\left(\|\mathbf{h}\|^3\right) s_{n,e_j}(\mathbf{x}).$$

Again, following Masry (1996b), we can write, in matrix form,

$$\boldsymbol{\tau}_n^*(\mathbf{x}) = \mathbf{S}_n(\mathbf{x})\left(\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x})\right) - \mathbf{B}_{n,2}(\mathbf{x})\mathbf{H}^{(2)}\mathbf{m_2}(\mathbf{x}) - \mathbf{B}_{n,3}(\mathbf{x})\mathbf{H}^{(3)}\mathbf{m_3}(\mathbf{x}) + o_p\left(\|\mathbf{h}\|^3\right)\mathbf{B}_{n,0}(\mathbf{x}), \tag{6.10}$$

or

$$\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) = (\mathbf{S}_n(\mathbf{x}))^{-1}\boldsymbol{\tau}_n^*(\mathbf{x}) + (\mathbf{S}_n(\mathbf{x}))^{-1}\left\{\mathbf{B}_{n,2}(\mathbf{x})\mathbf{H}^{(2)}\mathbf{m_2}(\mathbf{x}) + \mathbf{B}_{n,3}(\mathbf{x})\mathbf{H}^{(3)}\mathbf{m_3}(\mathbf{x})\right\} + o_p\left(\|\mathbf{h}\|^3\right)(\mathbf{S}_n(\mathbf{x}))^{-1}\mathbf{B}_{n,0}(\mathbf{x}). \tag{6.11}$$

where, as in Eq. (2.13) of Masry (1996b),

$$\boldsymbol{\tau}_n^*(\mathbf{x}) = \begin{bmatrix} t_{n,0}^*(\mathbf{x}) \\ t_{n,e_1}^*(\mathbf{x}) \\ \vdots \\ t_{n,e_d}^*(\mathbf{x}) \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \begin{bmatrix} 1 \\ \dfrac{\mathbf{X}_{i,1} - \mathbf{x}_1}{h_1} \\ \vdots \\ \dfrac{\mathbf{X}_{i,d} - \mathbf{x}_d}{h_d} \end{bmatrix} u_i, \tag{6.12}$$

$$\mathbf{B}_{n,2}(\mathbf{x}) = \begin{bmatrix} s_{n,2e_1}(\mathbf{x}) & s_{n,e_1+e_2}(\mathbf{x}) & \cdots & s_{n,2e_d}(\mathbf{x}) \\ s_{n,3e_1}(\mathbf{x}) & s_{n,2e_1+e_1}(\mathbf{x}) & \cdots & s_{n,e_1+2e_d}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ s_{n,2e_1+e_d}(\mathbf{x}) & s_{n,e_1+e_2+e_d}(\mathbf{x}) & \cdots & s_{n,3e_d}(\mathbf{x}) \end{bmatrix} \tag{6.13}$$

$$= \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \begin{bmatrix} \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^2 & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)\left(\frac{\mathbf{X}_{i,2}-\mathbf{x}_2}{h_2}\right) & \cdots & \left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right)^2 \\ \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^3 & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^2\left(\frac{\mathbf{X}_{i,2}-\mathbf{x}_2}{h_2}\right) & \cdots & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)\left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right)^2 \\ \vdots & \vdots & & \vdots \\ \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^2\left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right) & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)\left(\frac{\mathbf{X}_{i,2}-\mathbf{x}_2}{h_2}\right)\left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right) & \cdots & \left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right)^3 \end{bmatrix},$$

$$\mathbf{B}_{n,3}(\mathbf{x}) = \begin{bmatrix} s_{n,3e_1}(\mathbf{x}) & s_{n,2e_1+e_2}(\mathbf{x}) & \cdots & s_{n,3e_d}(\mathbf{x}) \\ s_{n,4e_1}(\mathbf{x}) & s_{n,3e_1+e_1}(\mathbf{x}) & \cdots & s_{n,e_1+3e_d}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ s_{n,3e_1+e_d}(\mathbf{x}) & s_{n,2e_1+e_2+e_d}(\mathbf{x}) & \cdots & s_{n,4e_d}(\mathbf{x}) \end{bmatrix}$$

$$= \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h},i\mathbf{x}} \begin{bmatrix} \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^3 & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^2\left(\frac{\mathbf{X}_{i,2}-\mathbf{x}_2}{h_2}\right) & \cdots & \left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right)^3 \\ \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^4 & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^3\left(\frac{\mathbf{X}_{i,2}-\mathbf{x}_2}{h_2}\right) & \cdots & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)\left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right)^3 \\ \vdots & \vdots & & \vdots \\ \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^3\left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right) & \left(\frac{\mathbf{X}_{i,1}-\mathbf{x}_1}{h_1}\right)^2\left(\frac{\mathbf{X}_{i,2}-\mathbf{x}_2}{h_2}\right)\left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right) & \cdots & \left(\frac{\mathbf{X}_{i,d}-\mathbf{x}_d}{h_d}\right)^4 \end{bmatrix}. \tag{6.14}$$

In light of Lemma 1,

$$\begin{aligned} \mathbf{S}_n(\mathbf{x}) &= \mathscr{M}_1 f(\mathbf{x}) + \sum_{l=1}^{d} h_l \mathscr{M}_1^l f_l(\mathbf{x}) + O_p\left(\|\mathbf{h}\|^2 + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right) \\ &= \begin{bmatrix} f(\mathbf{x}) & h_1\mu_2 f_1(\mathbf{x}) & h_2\mu_2 f_2(\mathbf{x}) & \cdots & h_d\mu_2 f_d(\mathbf{x}) \\ h_1\mu_2 f_1(\mathbf{x}) & \mu_2 f(\mathbf{x}) & 0 & \cdots & 0 \\ h_2\mu_2 f_2(\mathbf{x}) & 0 & \mu_2 f(\mathbf{x}) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ h_d\mu_2 f_d(\mathbf{x}) & 0 & \cdots & 0 & \mu_2 f(\mathbf{x}) \end{bmatrix} + O_p\left(\|\mathbf{h}\|^2 + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right), \end{aligned} \tag{6.15}$$

$$\begin{aligned} \mathbf{B}_{n,2}(\mathbf{x}) &= \mathscr{B}_2 f(\mathbf{x}) + \sum_{l=1}^{d} h_l \mathscr{B}_2^l f_l(\mathbf{x}) + O_p\left(\|\mathbf{h}\|^2 + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right) \\ &= \begin{bmatrix} \mu_2 f(\mathbf{x}) & 0 & \cdots & \mu_2 f(\mathbf{x}) \\ h_1\mu_4 f_1(\mathbf{x}) & h_2\mu_2^2 f_2(\mathbf{x}) & \cdots & h_1\mu_2^2 f_1(\mathbf{x}) \\ h_2\mu_2^2 f_2(\mathbf{x}) & h_1\mu_2^2 f_1(\mathbf{x}) & \cdots & h_2\mu_2^2 f_2(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ h_d\mu_2^2 f_d(\mathbf{x}) & 0 & \cdots & h_d\mu_4 f_d(\mathbf{x}) \end{bmatrix} + O_p\left(\|\mathbf{h}\|^2 + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right), \end{aligned} \tag{6.16}$$

$$\begin{aligned} \mathbf{B}_{n,3}(\mathbf{x}) &= \mathscr{B}_3 f(\mathbf{x}) + O_p\left(\|\mathbf{h}\| + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right) \\ &= \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \mu_4 f(\mathbf{x}) & 0 & \cdots & 0 \\ 0 & \mu_2^2 f(\mathbf{x}) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \mu_2^2 f(\mathbf{x}) & 0 & \cdots & \mu_4 f(\mathbf{x}) \end{bmatrix} + O_p\left(\|\mathbf{h}\| + \frac{1}{\sqrt{nh_1h_2\cdots h_d}}\right). \end{aligned} \tag{6.17}$$

Therefore,

$$
\begin{aligned}
&(\mathbf{S}_n(\mathbf{x}))^{-1}\left\{\mathbf{B}_{n,2}(\mathbf{x})\mathbf{H}^{(2)}\mathbf{m_2}(\mathbf{x})+\mathbf{B}_{n,3}(\mathbf{x})\mathbf{H}^{(3)}\mathbf{m_3}(\mathbf{x})\right\}\\
&=\underbrace{\mathscr{M}_1^{-1}\mathscr{B}_2\mathbf{H}^{(2)}\mathbf{m_2}(\mathbf{x})}_{O(\|\mathbf{h}\|^2)}\\
&\quad+\underbrace{\sum_{l=1}^{d}h_l\frac{f_l(\mathbf{x})}{f(\mathbf{x})}\left(\mathscr{M}_1^{-1}\mathscr{B}_2^l-\mathscr{M}_1^{-1}\mathscr{M}_1^l\mathscr{M}_1^{-1}\mathscr{B}_2\right)\mathbf{H}^{(2)}\mathbf{m_2}(\mathbf{x})+\mathscr{M}_1^{-1}\mathscr{B}_3\mathbf{H}^{(3)}\mathbf{m_3}(\mathbf{x})}_{O(\|\mathbf{h}\|^3)}\\
&=\left[\mathbf{Bias}_0(\mathbf{x})\ \mathbf{Bias}_1(\mathbf{x})\cdots\mathbf{Bias}_d(\mathbf{x})\right]^\top,
\end{aligned}
\tag{6.18}
$$

where for $k\in\{1,\ldots,d\}$,

$$
\begin{aligned}
\mathbf{Bias}_0(\mathbf{x})&=\sum_{l=1}^{d}h_l^2\mu_2 m_{ll}(\mathbf{x}),\\
\mathbf{Bias}_k(\mathbf{x})&=h_k^3\frac{\mu_4-\mu_2^2}{\mu_2}\frac{f_k(\mathbf{x})}{f(\mathbf{x})}m_{kk}(\mathbf{x})+h_k\sum_{l=1,l\neq k}^{d}h_l^2\mu_2\frac{f_l(\mathbf{x})}{f(\mathbf{x})}m_{kl}(\mathbf{x})\\
&\quad+h_k^3\frac{\mu_4}{\mu_2}m_{kkk}(\mathbf{x})+h_k\sum_{l=1,l\neq k}^{d}h_l^2\mu_2 m_{kll}(\mathbf{x}),
\end{aligned}
\tag{6.19}
$$

And the rest follows from Theorem 1.

ACKNOWLEDGMENTS

We thank two referees for their insightful comments that greatly improved the paper.

FUNDING

Li's research is partially supported by a China National Nature Science Foundation, Project # 71133001.

REFERENCES

Fan, J., Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

Fan, J., Gijbels, I., Hu, T. C., Huang, L. S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica* 6:113–127.

Li, D., Li, Q. (2010). Nonparametric/semiparametric estimation and testing of econometric models with data dependent smoothing parameters. *Journal of Econometrics* 157:179–190.

Li, Q., Lu, X., Ullah, A. (2003). Multivariate local polynomial regression for estimating average derivatives. *Journal of Nonparametric Statistics* 15:607–624.

Li, Q., Zhou, J. (2005). The uniqueness of cross-validation selected smoothing parameters in kernel estimation of nonparametric models. *Econometric Theory* 21:1017–1025.

Liu, Z., Stengos, T., Li, Q. Model check by kernel local polynomial methods. *Statistics and Probability Letters* 48:327–334.

Masry, E. (1996a). Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications* 65:81–101.

Masry, E. (1996b). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis* 17:571–599.

Ruppert, D., Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics* 22:1346–1370.

Su, L., Ullah, A. (2008). Local polynomial estimation of nonparametric simultaneous equation models. *Journal of Econometrics* 144:193–218.

Wand, M. P., Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of American Statistical Association* 88:520–528.

Xiao, Z. (2009). Functional coefficient co-integration models. *Journal of Econometrics* 152:81–92.