Data Sharpening on Unknown Manifold

Masaki Kudo[1] and Kanta Naito[1*]

[1] *Graduate School of Science and Engineering*

Shimane University, Matsue 690-8504, Japan

Abstract

This paper is concerned with data sharpening technique in nonparametric regression under the setting where the multivariate predictor is embedded in an unknown low-dimensional manifold. Theoretical asymptotic bias is derived, which reveals that the proposed data sharpening estimator has a reduced bias compared to the usual local linear estimator. The asymptotic normality of the data sharpening estimator is also developed. It can be confirmed from simulation and applications to real data that the bias reduction for the data sharpening estimator supported on unknown manifold is evident.

Keywords Data Sharpening, Bias Reduction, Nonparametric Regression, Manifold

1 Introduction

Bias reduction for kernel estimators is an important topic in nonparametric regression. Among the many nonparametric approaches, the local linear estimator is known to be an efficient standard tool. The local linear estimator with multivariate predictor and its asymptotic behavior were studied in [11]. The bias of the local linear estimator is $O_p(h^2)$ where h denotes scalar bandwidth as shown in Theorem 2.1 of [11]. Estimators based on high-order polynomials or high-order kernels have been used as conventional approaches for reducing bias. [5] proposed a comprehensive method called data sharpening for reducing bias. Their proposed estimator is derived by adding the usual local linear estimator applied to the data and the local linear estimator of its residuals. Compared to the bias of the local linear estimator, the data sharpening estimator has smaller bias in the order of $O_p(h^4)$. The data sharpening estimator was studied further in [9], where an effective method for the bandwidth selection for a data sharpening estimator was fully discussed. The data sharpening technique is closely related to the boosting. In fact the data sharpening can be seen as one-step L_2-boosting [3]: smooth the residual of initial estimator, and then the obtained residual smoother is added to the initial estimator. Also it is worth to note that this idea can be found in [12] as twicing.

*E-mail: `naito@riko.shimane-u.ac.jp`

On the other hand, there has been researched on identifying intrinsic low-dimensional structures underlying original high-dimensional data. It is assumed in this direction of research that the observed high-dimensional data are essentially lying on a low-dimensional smooth manifold. Recently, this assumption has been divided into known manifolds and unknown manifolds, which is reflected in the recent research relating to low-dimensional manifolds. There are several examples of research based on known manifolds. Directional statistics [8] is a representative area based on a known low-dimensional manifold, where the circle and sphere frequently appear as the low-dimensional manifold. Regression for circular and spherical data was discussed in [6] and [7]. While there has been an abundance of works concerning known manifolds, it seems that research based on unknown low-dimensional manifolds has still not sufficiently emerged. The local linear estimator on an unknown manifold was discussed in [2], where the asymptotic bias and variance were obtained using the local chart connecting the low-dimensional manifold and Euclidean space. Further works related to smoothing techniques on low-dimensional manifold can be found in [1] and [4].

In this paper, our purpose is to extend the data sharpening estimator on unknown manifolds by combining [2], [11] and [13]. In the research regarding regression on a manifold, we can use the definition in [2] without modification for the asymptotic term of the local linear estimator. We define the same data sharpening estimator as [2] and [9] on a low-dimensional manifold and investigate its asymptotic behavior. Thus, the present work can be seen as a generalization of the data sharpening technique in Euclidean space to an unknown low-dimensional manifold.

This paper is organized as follows. Section 2 introduces the proposed data sharpening estimator with a short review of the local linear estimator. The asymptotic results and assumptions are collected in Section 3. In particular, the asymptotic bias and variance of the data sharpening estimator are derived, and its asymptotic normality is also developed. To confirm the bias reduction, the data sharpening estimator is compared with the local linear estimator theoretically in Section 4. Asymptotic terms of the bias, variance and mean squared error (MSE) for the data sharpening estimator are shown graphically in some specific cases. In Section 5, the practical performance of the data sharpening estimator is investigated through simulation and applications to real data sets. Proofs of theoretical results and calculations are contained in Section 6.

2 Model and Data Sharpening

Throughout this paper, we use the notations LL and DS to denote Local Linear and Data Sharpening, respectively.

2.1 Model

Let $\{(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R} \mid 1 \leq i \leq n\}$ be a random sample drawn from the model

$$Y_i = m(X_i) + v(X_i)^{1/2}\varepsilon_i, \quad i \in \{1, \cdots, n\},$$

where $\varepsilon_1, \cdots, \varepsilon_n$ are independent and identically distributed (i.i.d.) random variables satisfying $E[\varepsilon_i] = 0$, $Var[\varepsilon_i] = 1$ that are independent of $X_1, \cdots, X_n$; $m : \mathbb{R}^p \to \mathbb{R}$ is the target regression function defined as

$$m(x) = E[Y|X = x]$$

for $x = (x_1, \cdots, x_p) \in \mathbb{R}^p$; and v is a positive variance function. In this paper, we assume that $X_1, \cdots, X_n$ are embedded into a low-dimensional manifold $\mathcal{X}$. Let positive integer d ($d \leq p$) be the dimension of manifold $\mathcal{X}$. Manifold $\mathcal{X}$ is represented as the image of a local chart φ. The local chart φ is a bijective and C^3 mapping from $\mathcal{B}^d_{0,r}$ into $\mathcal{X} \cap \mathcal{B}^p_{x,\mu}$ for some $r > 0$ and $\mu > 0$, where $\mathcal{B}^q_{y,\nu} = \{z \in \mathbb{R}^q | \ ||z - y|| < \nu\}$ ($q = d$ or $q = p$) is the ball with its center y and radius ν, and $||z|| = \sqrt{z^T z}$ for $z \in \mathbb{R}^q$. We suppose that the local chart φ satisfies $\varphi(0) = x$. For a given $x \in \mathcal{X}$, our problem is to estimate $m(x)$ nonparametrically.

2.2 Data Sharpening

Commonly used nonparametric regression estimators for $m(x)$ are multivariate versions of the Nadaraya-Watson kernel estimator, the local polynomial estimator, and the smoothing spline. [5] proposed some kernel regression estimators derived through a method called Data Sharpening (DS), which aims to reduce the bias of usual kernel estimators. The DS estimator is obtained by adding the residual smoother to the original regression estimator. The DS methods for the Nadaraya-Watson and LL estimators were discussed in [5] with a scalar bandwidth when $d = p$ and φ is the identical mapping on $\mathbb{R}^p$. The general theory for multivariate LL regression has been developed in [11]. Asymptotic conditional bias and variance of the LL estimator were derived in [11], where they treated not only a scalar bandwidth but also a general bandwidth matrix. For a scalar bandwidth $h > 0$, it is well known that the bias of the LL estimator is $O_p(h^2)$, whereas the DS estimator has bias of the order $O_p(h^4)$.

The LL estimator $\widehat{m}_{LL}(x, h)$ is defined from the solution $\hat{\alpha}(x)$ of the following weighted least squares problem:

$$(\hat{\alpha}(x), \hat{\beta}(x)) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{arg-min}} \sum_{i=1}^{n} \{Y_i - \alpha - \beta^T(X_i - x)\}^2 K_h\left(X_i - x\right),$$

where $K_h(U) = h^{-p}K(h^{-1}U)$, K is a p-variate kernel function. By direct calculations, $\widehat{m}_{LL}(x,h)$ for $m(x)$ is derived as

$$\widehat{m}_{LL}(x,h) = \hat{\alpha}(x) = e_1^T(X_x^T W_x X_x)^{-1} X_x^T W_x \mathbf{Y},$$

where

$$X_x = \begin{bmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{bmatrix}, \ \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

$W_x = \mathrm{diag}(K_h(X_1 - x), \cdots, K_h(X_n - x))$, $e_1 = [1\ 0 \cdots 0]^T \in \mathbb{R}^{p+1}$.

The DS method consists of two steps: the first step is smoothing by $\widehat{m}_{LL}(x,h)$, and then calculating the residuals by $r_i = Y_i - \widehat{m}_{LL}(X_i,h)$ for $i = 1, \cdots, n$. In the second step, let $\widehat{r}_{LL}(x,h)$ be the LL estimator applied to the residual data $\{(X_i, r_i)|\ 1 \le i \le n\}$, which is obtained as

$$\widehat{r}_{LL}(x,h) = e_1^T(X_x^T W_x X_x)^{-1} X_x^T W_x(\mathbf{Y} - \widetilde{M}),$$

where $\widetilde{M} = [\widehat{m}_{LL}(X_1,h) \cdots \widehat{m}_{LL}(X_n,h)]^T$. The DS estimator $\widehat{m}_{DS}(x,h)$ is then defined as

$$\widehat{m}_{DS}(x,h) = \widehat{m}_{LL}(x,h) + \widehat{r}_{LL}(x,h) = e_1^T(X_x^T W_x X_x)^{-1} X_x^T W_x(2\mathbf{Y} - \widetilde{M}).$$

3 Theory

This section summarizes the asymptotic behavior for the DS estimator on the manifold as well as the required assumptions. Assumptions for the LL estimator on the manifold were already developed in [2]. Some assumptions in [2] apply here, but additional assumptions are needed to develop asymptotics for $\widehat{m}_{DS}(x,h)$.

3.1 Assumptions

We denote closure of $\mathcal{B}^p_{x,\mu}$ as $\overline{\mathcal{B}^p_{x,\mu}} = \{y \in \mathbb{R}^p|\ ||y - x|| \le \mu\}$.

1. The true regression function m is bounded and C^4 on $\mathbb{R}^p$.

2. The variance function v is bounded above and from zero on $\mathbb{R}^p$.

3. The map φ given in Section 2.1 further satisfies the following. There exists a random element Z in $\mathbb{R}^d$ such that $\mathbb{P}(X \in \mathcal{S}) = \mathbb{Q}(Z \in \varphi^{-1}(\mathcal{S}))$ for any open set $\mathcal{S}$ of $\mathcal{X} \cap \mathcal{B}^p_{x,\mu}$, where $\mathbb{Q}$ is the induced measure

on $\mathcal{B}^d_{0,r}$. And the measure $\mathbb{Q}$ has the non-degenerate density, which is denoted by f. The density f is C^2 on $\mathcal{B}^d_{0,r}$ and $f(0) > 0$.

4. The kernel function K is a radially symmetric density and C^3.

5. For bandwidth h and sample size n, $nh^d \to \infty$ holds as $n \to \infty$ and $h \to 0$.

6. If the function w satisfies $w(y) \leq M(1 + ||y||^2)^3$ $(y \in \mathcal{X})$ for some $M > 0$ and $\gamma \in \{1, 2, 4\}$, then as $h \to 0$,

$$E\left[w(X)K\left(\frac{X-x}{h}\right)^\gamma 1(X \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})^c)\right] = o(h^{d+4})$$

and $\varphi^{-1}(\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X}) = \mathcal{B}^d_{0,h^{1-\eta}}$ hold for some $\eta(0 < \eta < 1)$.

7. If function w' is continuous on $\overline{\mathcal{B}^p_{x,\mu}} \times \overline{\mathcal{B}^p_{x,\mu}}$, then as $h \to 0$,

$$\begin{aligned} &E\bigg[w'(X_1, X_2)K\left(\frac{X_1 - X_2}{h}\right)K\left(\frac{X_2 - x}{h}\right)K\left(\frac{X_1 - x}{h}\right) \\ &\quad\times 1(X_1 \in (\mathcal{B}^p_{x,h^{1-\eta'}} \cap \mathcal{X})^c)1(X_2 \in (\mathcal{B}^p_{x,h^{1-\eta'}} \cap \mathcal{X})^c)\bigg] = o(h^{2d}), \\ &E\bigg[w'(X_1, X_2)K\left(\frac{X_1 - X_2}{h}\right)^2 K\left(\frac{X_1 - x}{h}\right)^2 \\ &\quad\times 1(X_1 \in (\mathcal{B}^p_{x,h^{1-\eta'}} \cap \mathcal{X})^c)1(X_2 \in (\mathcal{B}^p_{x,h^{1-\eta'}} \cap \mathcal{X})^c)\bigg] = o(h^{2d}) \end{aligned}$$

and $\varphi^{-1}(\mathcal{B}^p_{x,h^{1-\eta'}} \cap \mathcal{X}) = \mathcal{B}^d_{0,h^{1-\eta'}}$ hold for some $\eta'(0 < \eta' < 1)$.

8. If function w^* is continuous on $\overline{\mathcal{B}^p_{x,\mu}} \times \overline{\mathcal{B}^p_{x,\mu}} \times \overline{\mathcal{B}^p_{x,\mu}}$, then as $h \to 0$,

$$\begin{aligned} &E\bigg[w^*(X_1, X_2, X_3)K\left(\frac{X_3 - X_2}{h}\right)K\left(\frac{X_3 - X_1}{h}\right)K\left(\frac{X_2 - x}{h}\right)K\left(\frac{X_1 - x}{h}\right) \\ &\times 1(X_1 \in (\mathcal{B}^p_{x,h^{1-\eta^*}} \cap \mathcal{X})^c)1(X_2 \in (\mathcal{B}^p_{x,h^{1-\eta^*}} \cap \mathcal{X})^c)1(X_3 \in (\mathcal{B}^p_{x,h^{1-\eta^*}} \cap \mathcal{X})^c)\bigg] = o(h^{3d}) \end{aligned}$$

and $\varphi^{-1}(\mathcal{B}^p_{x,h^{1-\eta^*}} \cap \mathcal{X}) = \mathcal{B}^d_{0,h^{1-\eta^*}}$ hold for some $\eta^*(0 < \eta^* < 1)$.

3.2 Notation

We define the following notations, some of which are also used in [2] and [9].

For a function $g : \mathbb{R}^q \to \mathbb{R}$ $(q = d$ or $q = p)$, $\nabla g(x)$ and $\nabla^2 g(x)$ are the gradient (column) vector and the Hessian matrix of function g evaluated at x, respectively. The Jacobi $(d \times p)$ matrix of φ evaluated at

$a \in \mathbb{R}^d$ is

$$\mathcal{J}(a|\varphi) = [\nabla\varphi_1(z) \cdots \nabla\varphi_p(z)]^T\Big|_{z=a} = \left[\frac{\partial}{\partial z_j}\varphi_i(z)\right]_{1\le i\le p, 1\le j\le d}\Bigg|_{z=a}.$$

For g and $G : \mathbb{R}^p \to \mathbb{R}$, we define

$$\begin{aligned} B(x|g,\varphi) &= C(x|\varphi)^{-1}\int_{\mathbb{R}^d} u^T\mathcal{J}(\varphi^{-1}(x)|\varphi)^T\nabla^2 g(x)\mathcal{J}(\varphi^{-1}(x)|\varphi)uK(\mathcal{J}(\varphi^{-1}(x)|\varphi)u)du, \\ R(x|G,\varphi) &= \int_{\mathbb{R}^d} G(\mathcal{J}(\varphi^{-1}(x)|\varphi)u)^2 du, \end{aligned}$$

where

$$C(x|\varphi) = \int_{\mathbb{R}^d} K(\mathcal{J}(\varphi^{-1}(x)|\varphi)u)du.$$

Additionally, we put set

$$\begin{aligned} G^*(y) &= 2G(y) - C(x|\varphi)^{-1}G * G(y), \\ G * G(y) &= \int_{\mathbb{R}^d} G(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)G(y - \mathcal{J}(\varphi^{-1}(x)|\varphi)v))dv. \end{aligned}$$

When $d = p$ and φ is the identity mapping $id_{\mathbb{R}^p}$ on $\mathbb{R}^p$. Each notation and symbol are the same as those used in [11], for example $C(x|id_{\mathbb{R}^p}) = 1$.

Using our notations, functions $J_1(x)$ and $J_2(x)$ used in [2] are expressed as $J_1(x) = B(x|m,\varphi)$ and $J_2(x) = v(x)R(x|K,\varphi)/\{f(\varphi^{-1}(x))C(x|\varphi)^2\}$.

3.3 Asymptotic Conditional Bias and Variance

The asymptotic behavior of the DS estimator is summarized in the following theorem.

Theorem 1. Let $x \in \mathcal{X}$. Then under the assumptions in Section 3.1, asymptotic conditional bias and variance of $\widehat{m}_{DS}(x,h)$ are

$$\begin{aligned} Bias[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n] &= -\frac{h^4}{4}B(x|J_1,\varphi) + o_p(h^4), \\ Var[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n] &= \frac{1}{nh^d}\frac{R(x|K^*,\varphi)}{R(x|K,\varphi)}J_2(x)(1+o_p(1)). \end{aligned}$$

Remark 1. We compare Theorem 1 with Theorem 2.1 in [2]. It is realized that the DS method can be applied on an unknown manifold as the estimator with bias reduction. Our assumptions are essentially equal to those given in [2] for the LL estimator, in which the bias of $\widehat{m}_{LL}(x,h)$ is $O_p(h^2)$, but the bias of $\widehat{m}_{DS}(x,h)$ is $O_p(h^4)$ as in Theorem 1, indicating a bias reduction. On the other hand, $1/(nh^d)$ appears in the variance of $\widehat{m}_{DS}(x,h)$, which is the same order as the variance of $\widehat{m}_{LL}(x,h)$. The difference of variance for $\widehat{m}_{DS}(x,h)$ and $\widehat{m}_{LL}(x,h)$ is captured in the ratio $r(x|K^*,K,\varphi) = R(x|K^*,\varphi)/R(x|K,\varphi)$.

3.4 Asymptotic Normality

Asymptotic distributional result on the DS estimator is obtained as follows:

THEOREM 2. Let $E[\varepsilon^4] < \infty$ and $x \in \mathcal{X}$. Then under the same assumptions in Theorem 1 and $h = \kappa n^{-1/(d+8)}$ for a positive constant κ, $\widehat{m}_{DS}(x,h)$ has asymptotic normality:

$$\left(\widehat{m}_{DS}(x,h) - m(x) - \frac{h^4}{4}B(x|J_1,\varphi)\right) \Big/ \sqrt{Var[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n]} \xrightarrow{D} N(0,1),$$

where "$\xrightarrow{D}$" designates convergence in distribution, and $N(0,1)$ is the standard normal distribution.

REMARK 2. The asymptotic normality of $\widehat{m}_{DS}(x,h)$ on $\mathbb{R}^p$ was already developed in [5] and [13]. Although asymptotic normality of the LL estimator in [2] has not been discussed, it can be proved that

$$\left(\widehat{m}_{LL}(x,h) - m(x) - \frac{h^2}{2}J_1(x)\right) \Big/ \sqrt{Var[\widehat{m}_{LL}(x,h)|X_1,\cdots,X_n]} \xrightarrow{D} N(0,1)$$

holds for $h = \kappa n^{-1/(d+4)}$ using with the similar calculations as for the proof of Theorem 2.

4 Theoretical comparison

In this section, we report the theoretical performance of the DS estimator for $m : \mathbb{R}^p \to \mathbb{R}$ on the d-dimensional closed and smooth manifold $\mathcal{X} \subset \mathbb{R}^p$. We investigate a bias reduction on a sharp-peaked point by comparing the behavior of the leading terms in Theorem 1 of this paper and Theorem 2.1 in [2]. In particular, our targets for comparison are the following terms: $(h^4/4)J_1(x)$ versus $(h^8/16)B(x|J_1,\varphi)$ in squared bias, and $(nh^d)^{-1}J_2(x)$ versus $(nh^d)^{-1}r(x|K^*,K,\varphi)J_2(x)$ in variance.

4.1 Settings: True function, kernel, local chart and bandwidth

For theoretical comparison, we consider the case $p = 2$ and $d = 1$. We utilize the true regression function m as

$$m(x) = m(x_1,x_2) = x_1 + 2\exp(-400(x_2 - 0.5)^2)$$

for $x \in \mathcal{X} \subset \mathbb{R}^2$ and the Gaussian density

$$K(x) = \frac{1}{2\pi} \exp\left(-\frac{x^T x}{2}\right) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right)$$

is used as the kernel. The local chart φ is supported on $[0,1]$, and hence, we have considered $\mathcal{X} = \{\varphi(z) \in \mathbb{R}^2 | z \in [0,1]\}$. We assume that the density function f has uniform distribution on $[0,1]$. By the use of Gaussian kernel K and setting $g(x) = ||\mathcal{J}(\varphi^{-1}(x)|\varphi)||$, we obtain

$$\begin{aligned}
\frac{h^2}{2} J_1(x) &= \frac{h^2}{2} g(x)^{-2} \mathcal{J}(\varphi^{-1}(x)|\varphi)^T \nabla^2 m(x) \mathcal{J}(\varphi^{-1}(x)|\varphi), && (1)\\
\frac{1}{nh} J_2(x) &= \frac{1}{nh} \frac{1}{\sqrt{2}} \frac{v(x)}{\sqrt{2\pi} f(\varphi^{-1}(x))} g(x), && (2)\\
-\frac{h^4}{4} B(x|J_1, \varphi) &= -\frac{h^4}{4} g(x)^{-2} \mathcal{J}(\varphi^{-1}(x)|\varphi)^T \nabla^2 J_1(x) \mathcal{J}(\varphi^{-1}(x)|\varphi), && (3)\\
\frac{1}{nh} \frac{R(x|K^*, \varphi)}{R(x|K, \varphi)} J_2(x) &= \frac{1}{nh} \sqrt{2} \left(2\sqrt{2} - \frac{4}{\sqrt{3}} + \frac{1}{2}\right) J_2(x) && (4)
\end{aligned}$$

where $\nabla^2 m(x)$ is

$$\nabla^2 m(x) = \begin{bmatrix} 0 & 0 \\ 0 & \frac{\partial^2}{\partial x_2^2} m(x_1, x_2) \end{bmatrix},$$

with

$$\frac{\partial^2}{\partial x_2^2} m(x_1, x_2) = 1600\{800(x_2 - 0.5)^2 - 1\} \exp(-400(x_2 - 0.5)^2).$$

Here, we set $G(x) = \{G_0(x)\}^2$ and $G_0(x) = \varphi_2'(\varphi^{-1}(x))$. For smooth function $\eta(x)$, we use subscript to denote the partial derivatives by variables corresponding to those indices throughout in this section as

$$\begin{aligned}
\frac{\partial^3}{\partial x_i \partial x_j \partial x_k} \eta(x) &= \eta_{ijk}(x),\\
\frac{\partial^4}{\partial x_i \partial x_j \partial x_k \partial x_\ell} \eta(x) &= \eta_{ijk\ell}(x)
\end{aligned}$$

and so on. Then, the each component of $\nabla^2 J_1(x)$ is

$$\begin{aligned}
J_{1,11}(x) &= g(x)^{-3} m_{22}(x) \left\{6g(x)^{-1} g_1(x)^2 G(x) - 2g_{11}(x) G(x) - 4g_1(x) G_1(x) + g(x) G_{11}(x)\right\},\\
J_{1,12}(x) &= g(x)^{-3} m_{22}(x) \left\{6g(x)^{-1} g_2(x) g_1(x) G(x) - 2g_{12}(x) G(x) - 2g_1(x) G_2(x)\right.\\
&\quad \left. -2G_1(x) g_2(x) - 2g(x) G_{12}(x)\right\} + g(x)^{-3} m_{222}(x) \left\{-2g_1(x) G(x) + g(x) G_1(x)\right\},
\end{aligned}$$

$$
\begin{aligned}
J_{1,22}(x) &= g(x)^{-3}m_{22}(x)\left\{6g(x)^{-1}g_2(x)^2G(x) + g(x)G_{22}(x) - 2g_{22}(x)G(x) - 4g_2(x)G_2(x)\right\} \\
&\quad + g(x)^{-2}m_{2222}(x)G(x) - 2g(x)^{-3}m_{222}(x)\left\{g_2(x)G(x) - g(x)G_2(x)\right\},
\end{aligned}
$$

where

$$
m_{222}(x) = -1600 \times 800\{800(x_2-0.5)^3 - 3(x_2-0.5)\}\exp(-400(x_2-0.5)^2)
$$

and

$$
m_{2222}(x) = -1280000\{-640000(x_2-0.5)^4 + 4800(x_2-0.5)^2 - 3\}\exp(-400(x_2-0.5)^2).
$$

We focus on functions

$$
\frac{h^4}{4}J_1(x)^2 \tag{5}
$$

and

$$
\frac{h^8}{16}B(x|J_1,\varphi)^2 \tag{6}
$$

calculated from (1) and (3). The functions in (2) and (4) are also used in comparison of variances.

Here, it is necessary to determine the bandwidth parameter h. We consider the bandwidth parameter as the minimizer of AMISE:

$$
h_{LL} = \underset{h>0}{\text{arg-min}}\left\{\frac{h^4}{4}\int_0^1 J_1(\varphi(z))^2 f(z)dz + \frac{1}{nh}\int_0^1 J_2(\varphi(z))f(z)dz\right\}
$$

and

$$
h_{DS} = \underset{h>0}{\text{arg-min}}\left\{\frac{h^8}{16}\int_0^1 B(\varphi(z)|J_1,\varphi)^2 f(z)dz + \frac{1}{nh}\int_0^1 \frac{R(\varphi(z)|K^*,\varphi)}{R(\varphi(z)|K,\varphi)}J_2(\varphi(z))f(z)dz\right\}.
$$

By differentiate with h, we obtain

$$
h_{LL} = \left(\frac{\int_0^1 J_2(\varphi(z))dz}{\int_0^1 J_1(\varphi(z))^2 dz}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} \tag{7}
$$

and

$$h_{DS} \quad = \quad \left(\frac{2\displaystyle\int_0^1 \frac{R(\varphi(z)|K^*,\varphi)}{R(\varphi(z)|K,\varphi)} J_2(\varphi(z))dz}{\displaystyle\int_0^1 B(\varphi(z)|J_1,\varphi)^2 dz} \right)^{\frac{1}{9}} n^{-\frac{1}{9}}. \tag{8}$$

In fact, by estimating integrals using Monte Carlo method we utilize

$$\widehat{h}_{LL} \quad = \quad \left(\frac{\displaystyle\frac{1}{M}\sum_{j=1}^{M} J_2(\varphi(z_{(j)}))}{\displaystyle\frac{1}{M}\sum_{j=1}^{M} J_1(\varphi(z_{(j)}))^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}} \tag{9}$$

and

$$\widehat{h}_{DS} \quad = \quad \left(\frac{\displaystyle\frac{2}{M}\sum_{j=1}^{M} \frac{R(\varphi(z_{(j)})|K^*,\varphi)}{R(\varphi(z_{(j)})|K,\varphi)} J_2(\varphi(z_{(j)})))}{\displaystyle\frac{1}{M}\sum_{j=1}^{M} B(\varphi(z_{(j)})|J_1,\varphi)^2} \right)^{\frac{1}{9}} n^{-\frac{1}{9}}, \tag{10}$$

where each $z_{(j)}$ is generated uniformly on $[0,1]$, $1 \le j \le M$ and $M = 10000$.

4.2 Case of a Power Function

We consider the situation where the local chart is

$$\varphi(z) = \rho_\alpha(z) = \begin{bmatrix} z^\alpha \\ z \end{bmatrix}$$

and $0 \le z \le 1$. For the case $\alpha = 3$, the true curve $m(\rho_3(z))$ achieves its maximum at $z = 1/2$, as displayed in Figure 1. By using $\rho_\alpha(z)$, we have $\rho_\alpha^{-1}(x) = x_2$, $g(x) = \sqrt{\alpha^2 x_2^{2(\alpha-1)} + 1}$, $G_0(x) = 1$ and $\mathcal{J}(\rho_\alpha^{-1}(x)|\rho_\alpha) = [\alpha x_2^{\alpha-1} \ 1]^T$. The $g(x)$ and $m_{22}(x)$ are both functions containing only x_2. Thus, we obtain

$$\begin{aligned}
\mathcal{J}(\rho_\alpha^{-1}(x)|\rho_\alpha)^T \nabla^2 m(x) \mathcal{J}(\rho_\alpha^{-1}(x)|\rho_\alpha) &= m_{22}(x), \\
\mathcal{J}(\rho_\alpha^{-1}(x)|\rho_\alpha)^T \nabla^2 J_1(x) \mathcal{J}(\rho_\alpha^{-1}(x)|\rho_\alpha) &= J_{1,22}(x), \\
g_2(x) &= \frac{\alpha^2(\alpha-1)x_2^{2\alpha-3}}{g(x)}
\end{aligned}$$

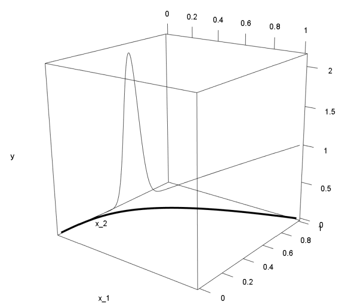


Figure 1. True curve of m and the local chart ρ_3. The manifold $\mathcal{X}$ is the image of ρ_3 and is embedded in the plane $y = 0$ in $\mathbb{R}^3$ as a thick solid curve. The true curve of $m(\rho_3(z))$ is drawn as a thin solid curve on embedded $\mathcal{X}$.

and

$$g_{22}(x) = \frac{\alpha^2(\alpha-1)\left\{(2\alpha-3)x_2^{2\alpha-4}g(x)^2 - x_2^{2\alpha-3}g_2(x)\right\}}{g(x)^2}.$$

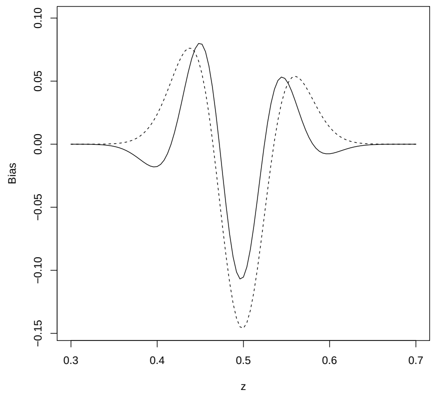
For our consideration, we utilized the case $\alpha = 3$ and $n = 1000$, and used $\widehat{h}_{LL} = 0.01688$ and $\widehat{h}_{DS} = 0.02275$ calculated by (9) and (10). Functions (1), (2), (3) and (4) have been regarded as univariate function with z by plugging $x = \rho_3(z)$. According to Figure 2, a bias reduction occurs around $z = 1/2$. In particular, the squared bias of the DS estimator is smaller than that of the LL estimator at point 0.5 in panel (b) of Figure 2. Note that solid curves in panels (a) and (b) are close to the dashed curves on the intervals $[0.3, 0.4]$ and $[0.6, 0.7]$. However, the placement of the solid curve and the dashed curve is the opposite on the interval $[0.4, 0.5]$ and $[0.5, 0.6]$. Panel (c) displays the difference in asymptotic variances, in which the DS estimator has a slightly bigger variance than that of the LL estimator. The reason for this could be considered to be the fact that the variance of the DS estimator includes a convolution $K * K$. However, from the view of the asymptotic MSE in panel (d), the DS estimator is superior in the sense of smaller MSE around of $z = 1/2$.

5 Practical Performance

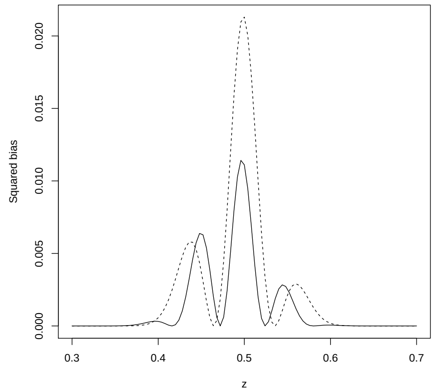
This section investigates bias reduction of the DS estimator using simulated and real data. In both cases, our aim is to confirm that the DS estimator can trace sudden variations hidden in data.

5.1 Simulation

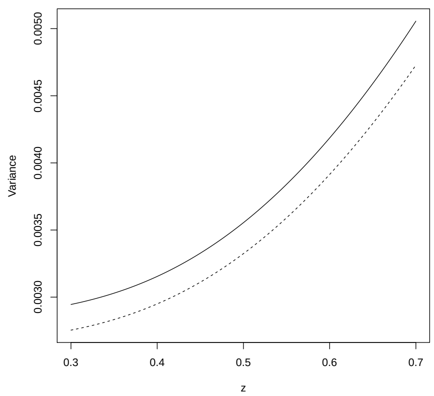
The purpose of this simulation is to verify the bias reduction of the DS estimator for sudden variation by estimating bias with the simulated data. We utilize m, K, ρ_3 used in Section 4 in this simulation. Data sets are generated by the following fixed design. First, we generated $n = 1000$ points uniformly from the



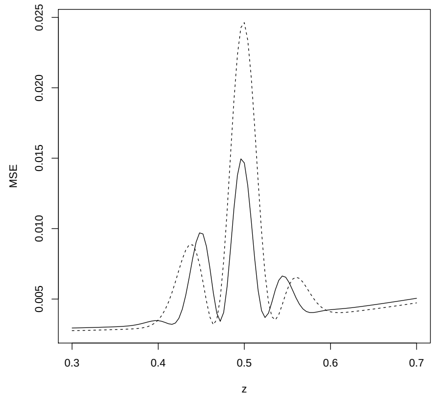
(a) Bias based on ρ_3



(b) Squared bias based on ρ_3



(c) Variance based on ρ_3



(d) MSE based on ρ_3

Figure 2. Comparison of leading terms by Theorem 1 between the DS estimator (solid line) and LL estimator (dashed line) in the case of $n = 1000$ and local chart ρ_3. Here, the leading terms are used as the function with z by $x = \rho_3(z)$ and each bandwidth parameter is $\widehat{h}_{LL}$ and $\widehat{h}_{DS}$. Panel (a) displays (1) and (3), panel (b) displays (5) and (6), panel (c) displays (2) and (4) and panel (d) displays (5) + (2) and (6) + (4). We focus on the interval $[0.3, 0.7]$ to emphasize the difference.

interval $[0,1]$ as Z, denoted $\{z_i \in [0,1] | z_i \sim U(0,1), 1 \le i \le n\}$. The predictor X is defined as $\{(x_{1i}, x_{2i}) \in \mathbb{R}^2 | (x_{1i}, x_{2i}) = \varphi(z_i), 1 \le i \le n\}$. In each step k, we generate n random errors $\{\varepsilon_i^{(k)} | \varepsilon_i^{(k)} \sim N(0,1), 1 \le i \le n\}$, and then the response variable Y is obtained as $\{y_i^{(k)} \in \mathbb{R} | y_i^{(k)} = m(x_{1i}, x_{2i}) + \varepsilon_i^{(k)}, 1 \le i \le n\}$. Let $\mathcal{D}^{(k)} = \{(x_{1i}, x_{2i}, y_i^{(k)}) \in \mathbb{R}^3 | 1 \le i \le n\}$ be the simulated data in step k. By the shape of m, the simulated data sets seem to have the structure of sudden variation around $z = 0.5$ for the case of $\varphi = \rho_3$. For example, $\mathcal{D}^{(1)}$ is shown in Figure 3. We compose $\widehat{m}_{DS}(x, h; \mathcal{D}^{(k)})$ and $\widehat{m}_{LL}(x, h; \mathcal{D}^{(k)})$ by using $\mathcal{D}^{(k)}$. We iterate above

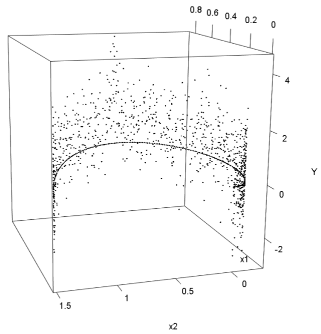


Figure 3. Example of simulated data showing $\mathcal{D}^{(1)}$ for the case of $\varphi = \rho_3$, in which the predictors are embedded.

steps from $k = 1$ to $N = 10000$ and calculate the following two values:

$$\widehat{Bias}[\widehat{m}_{LL}(x, \widehat{h}_{LL})] = \frac{1}{N} \sum_{k=1}^{N} \widehat{m}_{LL}(x, \widehat{h}_{LL}; \mathcal{D}^{(k)}) - m(x) \tag{11}$$

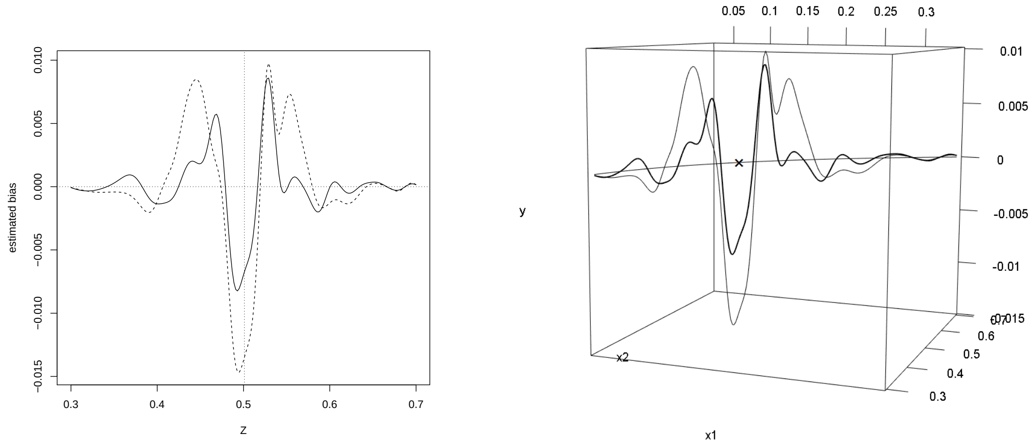
and

$$\widehat{Bias}[\widehat{m}_{DS}(x, \widehat{h}_{DS})] = \frac{1}{N} \sum_{k=1}^{N} \widehat{m}_{DS}(x, \widehat{h}_{DS}; \mathcal{D}^{(k)}) - m(x), \tag{12}$$

where $\widehat{h}_{LL}$ and $\widehat{h}_{DS}$ are the optimal values used in Section 4. We implemented this simulation 5 times, and then calculated the mean of the obtained (11) and (12), which are displayed in Figure 4.

In the left panel (a) of Figure 4, it can be seen that the solid curve (DS) nears the horizontal dotted line ($y = 0$) around the vertical dotted line ($z = 0.5$). From this we can claim that the reduction of estimated bias of the DS estimator occurs around the vertical dotted line. The right panel (b) of Figure 4 shows that the DS estimator supported on the 1-dimensional manifold $\mathcal{X}$ has a smaller bias around the point of sudden variation $x = \rho_3(0.5)$, which can also be observed in panel (a). From the this result, we can consider that bias reduction for the DS estimator on an unknown manifold also occurs around the point of sudden variation.

Next, we compare this simulation and the theoretical results from Figures 2. We compare panel (a) of Figure 2 with panel (a) of Figure 4. The following facts can be observed: the curve of estimated bias has a



(a) Based on ρ_3

(b) Based on ρ_3

Figure 4. Comparison of (11) and (12). Panel (a): (12) (solid) and (11) (dashed) with $x = \rho_3(z)$, drawn as the curve of z; (b): (12) (thick) and (11) (thin) with $x = \rho_3(z)$ in $\mathbb{R}^3$; In (b), the solid curve on $y = 0$ denotes $\mathcal{X}$, and the mark $\times$ means the point of sudden variation.

similar shape to the curve of asymptotic bias in Figure 4 for each estimator, although the scales of curves (11) and (12) in Figure 4 differ to those of (1) and (3) in Figures 2 because each theoretical bandwidth used in this simulation might be much smaller than the optimal bandwidth for each simulated data set $\mathcal{D}^{(k)}$. Reduced bias for the DS estimator has certainly occurred, as it is observed that the curve of (12) is closer to zero than the curve of (11) around $z = 0.5$ in Figure 4. These two points correspond to places where the magnitude of asymptotic bias for the DS estimator is smaller than that of the LL estimator in Figure 2. It can be confirmed from the above facts that data sharpening is an efficient smoothing technique endowed with the property of bias reduction.

5.2 Real Data: directional data

We demonstrate here the applicability of the DS estimator to real data. To study the bias reduction, we investigate the fitting of the DS estimator. If the fitting of the DS estimator is better than that of the LL estimator, this indicates that the DS estimator can trace the sudden variation of data because of the bias reduction. We used the sets of directional data tabulated in Tables 1.1 and 1.2 in [8], and the data of angle and velocity for 199 winds included in the R package “NPCirc” [10].

Table 1.1 is the frequencies of the vanishing angles of 714 nonmigratory British mallards with 0° defined as north. Table 1.2 shows the orientation of the least projection elongations of sand grains in thin sections, cut parallel to laminations, of Recent Gulf Coast beach sand. The aforementioned tables in [8] include frequencies for each 18 subintervals of degree divided from the interval of degree $[0°, 360°]$. We regard the middle value of each bin of subinterval as the observation of the predictor and the height of bin as the observation of the response variable.

We denote each data set as $\mathcal{D}_i = \{(x_k, y_k) | 1 \leq k \leq n_i\}$ and note that each x_k is embedded in the unit circle $\mathcal{X} = \{z \in \mathbb{R}^2 | \ ||z|| < 1\}$. $\mathcal{D}_1$ and $\mathcal{D}_2$ correspond to Tables 1.1 and 1.2 in [8], respectively, with sample size $n_i = 18 (i = 1, 2)$. $\mathcal{D}_3$ designates the data from " NPCirc" with sample size $n_3 = 199$, and $\mathcal{D}_4$ is based on $\mathcal{D}_3$ without data $(2.60, 8.4)$. Although the local chart φ is unknown for practical situation of real data, we exploit the local chart as follows for the purpose of expressing estimators graphically.

$$\varphi(\theta) = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$

with $\theta \in [0, 2\pi)$ because $\mathcal{D}_k$'s are sets of circular data. Note that the LL and DS estimators can be composed without the local chart.

To determine the bandwidth parameter h, we implemented the following leave-one-out cross validation for the LL and DS estimators:

$$RSS_{LL}(h) = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \widehat{m}_{LL}^{(-i)}(x_i, h) \right\}^2 \tag{13}$$

and

$$RSS_{DS}(h) = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \widehat{m}_{DS}^{(-i)}(x_i, h) \right\}^2 \tag{14}$$

for each value of h and each data set, where $\widehat{m}_{DS}^{(-i)}(x_i, h)$ is the DS estimator based on the data without (x_i, y_i), evaluated at $x = x_i$ and $\widehat{m}_{LL}^{(-i)}(x_i, h)$ is defined similarly. We utilized the bandwidth parameter h as the minimizer of RSS:

$$h_{LL} = \underset{h>0}{\text{arg-min}} RSS_{LL}(h) \tag{15}$$

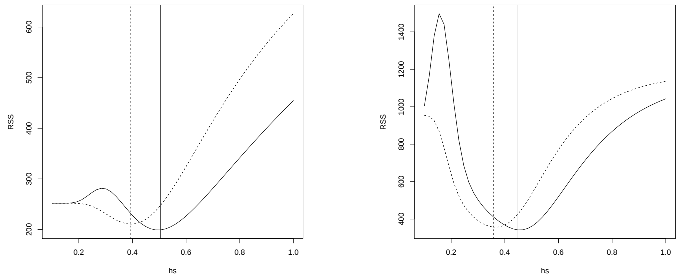
and

$$h_{DS} = \underset{h>0}{\text{arg-min}} RSS_{DS}(h). \tag{16}$$

The curve of RSS and the chosen value of h for each $\mathcal{D}_k$ are exhibited in Figure 5. Actual values of (15) and (16) as well as corresponding minimum of (13) and (14) are tabulated in Table 1.

Table 1: The optimal bandwidths (15) and (16), and the minimum values of (13) and (14).

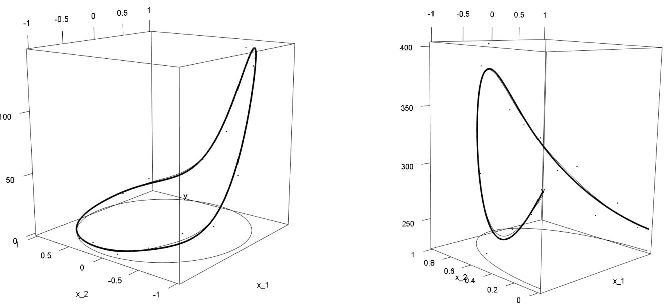
DATA	h_{LL}	h_{DS}	$RSS_{LL}(h_{LL})$	$RSS_{DS}(h_{DS})$
$\mathcal{D}_1$	0.394	0.504	210.714	199.277
$\mathcal{D}_2$	0.357	0.449	357.091	341.957



(a) $\mathcal{D}_1$ (b) $\mathcal{D}_2$

Figure 5. RSS as the function of bandwidth h. The solid curve is (14), the dashed curve is (13) and the vertical lines are the minimizers h_{LL} (15) and h_{DS} (16) of (13) and (14), respectively, in both panels.

We first look at the obtained regression curves in Figure 6. Panel (a) shows the curves of the DS estimator and LL estimator based on the unit circle in $\mathbb{R}^3$. Since the predictors in $\mathcal{D}_2$ were observed on the interval of degree $[0°, 180°]$, panel (b) displays the same curves based on the half unit circle. We can see that the two curves of the LL and DS estimators are almost the same. A possible reason is that the sample size $n = 18$ is small for each $\mathcal{D}_i$. However Figure 5 and Table 1 reveal that the DS estimator is better than the LL estimator from the perspective of the best RSS for each $\mathcal{D}_k$.



(a) $\mathcal{D}_1$ (b) $\mathcal{D}_2$

Figure 6. The DS estimator and LL estimator in $\mathbb{R}^3$. The thin curve is the LL estimator, the thick curve is the DS estimator, the dotted points are data, and the solid line on the bottom plane is $\mathcal{X}$.

To compare the DS estimator and LL estimator with changing bandwidth, we compose two estimators using the same h in an interval including (13) and (14). The results are shown in Figure 7. Panels (a) and

(b) show the curves of the LL and DS estimators by $\mathcal{D}_1$, while panels (c) and (d) display those by $\mathcal{D}_2$. The thin curves show the estimator using the optimal bandwidths, and the two horizontal dotted lines indicate the range of the maximums of the LL estimators with different bandwidths.

From panels (a) and (b) of Figure 7, we can observe that the range of maximums of the LL estimator for different bandwidths is wider than that of the DS estimator. This difference in the range of maximums indicates that the performance of the LL estimator for sudden variation largely changes depending on bandwidths, while a similar performance of the DS estimator is observed regardless of the values of bandwidths which are not largely different from the optimal one.

In panels (c) and (d) of Figure 7, we observe a sudden variation appearing around the interval $[1.5, 2.0]$. From panels (c) and (d), the range of maximums in panel (d) seem to be almost the same as the range of maximums in panel (c) in contrast to panels (a) and (b). One difference between panel (a) and panel (c) is the variance of data around the peak. That is, data around the peak in panel (a) is dense whereas data around the peak in the panel (c) like as sparse. However, examining panels (c) and (d), the maximums of the DS estimators are bigger than those of the LL estimators with the same h. In other words, even if the data seem to be sparse in a sudden variation, the DS estimator can trace the sudden variation better than the LL estimator.

The block is our focus in panels (c) and (d), and actually the place where the data and curves increase slowly. It can be observed from panel (c) that one of the LL estimators is far from optimal one, but panel (d) shows the DS estimator behaving stably in the same area. Thus, it can be claimed that the DS estimator performs more stably than the LL estimator in the area where the data varies slightly.

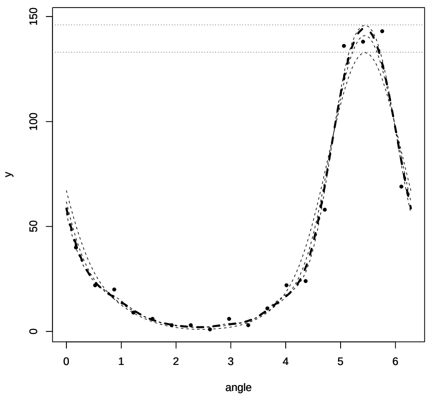
By summarizing the above considerations, the curve of the LL estimator is significantly affected by changing the bandwidth h whereas the curves of the DS estimator are changed at the place of a sudden variation by changing h. Further, the DS estimator can indicate the shape of a sudden variation better than the LL estimator with the same h. This suggests the effect bias reduction is present when using the DS estimator.

Next we address the results for $\mathcal{D}_3$ and $\mathcal{D}_4$. Optimal bandwidths and RSS values are tabulated in Table 2.

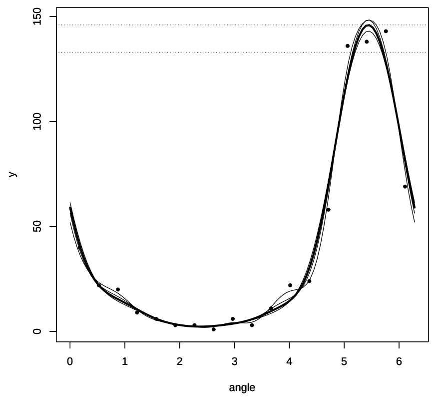
Table 2: The optimal bandwidths (15) and (16), and the minimum values of (13) and (14).

DATA	h_{LL}	h_{DS}	$RSS_{LL}(h_{LL})$	$RSS_{DS}(h_{DS})$
$\mathcal{D}_3$	0.424	0.534	14.853	14.856
$\mathcal{D}_4$	0.369	0.422	14.663	14.666

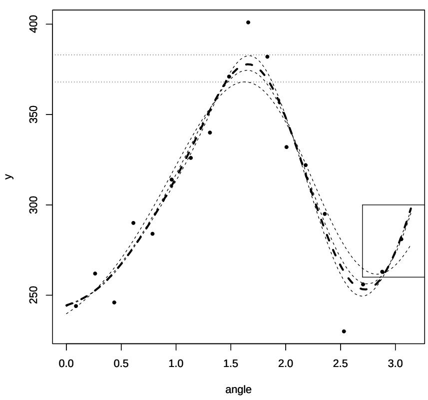
Table 2 and Figure 8 show that there is no significant difference between the LL estimator and DS



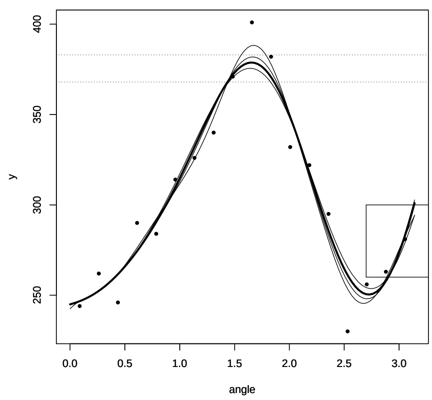
(a) LL estimators by $\mathcal{D}_1$ with $h = 0.35, 0.42, 0.55$



(b) DS estimators by $\mathcal{D}_1$ with $h = 0.35, 0.42, 0.55$

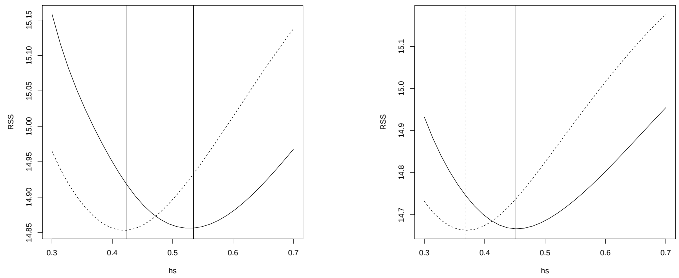


(c) LL estimators by $\mathcal{D}_2$ with $h = 0.3, 0.4, 0.5$



(d) DS estimators by $\mathcal{D}_2$ with $h = 0.3, 0.4, 0.5$

Figure 7. Comparison of the DS estimator and LL estimator with changing the bandwidth. Estimators are shown as the functions of the angle θ. The thick curve in (a), (c): the LL estimator with $h = h_{LL}$, the thick curve in (b), (d): the DS estimator with $h = h_{DS}$. The dashed solid and thin solid curves in all panels are estimators with listed values of bandwidths. The dotted points are data. The blocks in all panels are our attention area and the horizontal dotted lines are indicating the range of the maximums of the LL estimators for listed differenced bandwidths.



(a) Curves of RSS by $\mathcal{D}_3$ (b) Curves of RSS by $\mathcal{D}_4$

Figure 8. RSS as a function of bandwidth h. The solid curve is (14), the dashed curve is (13) and the vertical lines are the minimizers h_{LL} (15) and h_{DS} (16) of (13) and (14), respectively, in both panels.

estimator when considering RSS values. Note that the placement of $h_{LL} < h_{DS}$ for the data sets $\mathcal{D}_3$ and $\mathcal{D}_4$ are the same for $\mathcal{D}_1$ and $\mathcal{D}_2$.

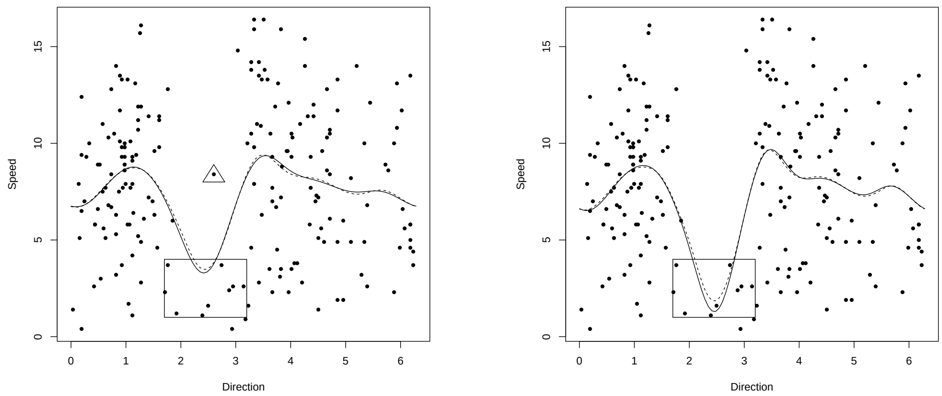
Figure 9 shows the curves of the estimator drawn by $x = \varphi(\theta)$ in $\mathbb{R}^2$. From panel (a) of Figure 9, we can observe that the data includes a sudden variation in the interval $[1.5, 3.5]$ and its variation in the interval is smaller than those outside of the interval. By closely looking at the blocked area in panel (a), the DS estimator behaves nearer the data as compared to the LL estimator, which might be understood as the effect of bias reduction.

Next, we address the result of $\mathcal{D}_4$, which is made by deleting an observation $(2.60, 8.4)$ indicated by $\triangle$ from $\mathcal{D}_3$. The estimators by optimal bandwidths are shown in panel (b) of Figure 9, where it can be observed that the DS estimator is below the LL estimator in the interval $[1.5, 3.5]$ including a sudden variation of data. Hence, we see that, due to bias reduction, the DS estimator can trace the sudden variation better than the LL estimator.

We observe from comparing panels (a) and (b) of Figure 9 that the minimum of the DS estimator is smaller than that of the LL estimator, although both estimators are pulled up by the data $\triangle$. In other words, even if the DS estimator is affected by a possible outlier, it has a tendency to trace a sudden variation. This can be understood as an effect of bias reduction.

6 Conclusion

In this paper we have proposed the DS regression estimator on an unknown low-dimensional manifold. It has been proved theoretically in Theorem 1 that the DS estimator has a reduced bias compared to the LL estimator even in the situation where the covariates are embedded in the manifold. In Sections 4 and 5, we have confirmed numerically that such a bias reduction certainly occurred for simulated data as well as some



(a) Estimators by $\mathcal{D}_3$

(b) Estimators by $\mathcal{D}_4$

Figure 9. The DS estimator and LL estimator drawn in $\mathbb{R}^2$ by $x = \varphi(\theta)$. The dashed curve is the LL estimator, the solid curve is the DS estimator, and the dotted points are data. $\triangle$ is the outlier point.

real data sets.

We have also developed asymptotic normality of the DS estimator in Theorem 2. This asymptotic normality makes it possible to construct an approximate point-wise confidence interval or confidence band of the regression function as was discussed in [13], which is our future problem.

We have not fully discussed the problem of "data-based bandwidth selection". This usually can be accomplished via the minimization of AMISE as suggested in Section 4. To do this we need a fine estimate of the unknown local chart φ and its derivatives, which is also another important issue to be tackled.

7 Proof

For a C^ℓ function $g : \mathbb{R}^p \to \mathbb{R}$, $a = (a_1, \cdots, a_p) \in \mathbb{R}^p$, $\mathbf{i} = (i_1, \cdots, i_p)$ with i_k nonnegative integer, we define $a^{\mathbf{i}} = a_1^{i_1} \cdots a_p^{i_p}, |\mathbf{i}| = \sum_{k=1}^{p} i_k$,

$$\left(\frac{\partial}{\partial x}\right)^{\mathbf{i}} g(x) = \frac{\partial^{|\mathbf{i}|}}{\partial x_1^{i_1} \cdots \partial x_p^{i_p}} g(x)$$

and

$$\begin{pmatrix} j \\ \mathbf{i} \end{pmatrix} = \frac{j!}{i_1! \cdots i_{p-1}! \left(j - \sum_{k=1}^{p-1} i_k\right)!}$$

for $|\mathbf{i}| = j$. We denote j-th differentials $(1 \le j \le \ell)$ of $g(x)$ over a segment a as

$$d_a^j g(x) = \sum_{|\mathbf{i}|=j} \begin{pmatrix} j \\ \mathbf{i} \end{pmatrix} a^{\mathbf{i}} \left(\frac{\partial}{\partial x}\right)^{\mathbf{i}} g(x).$$

Using this notation, $m(X_i)$ can be expressed as

$$\begin{aligned} m(X_i) &= m(x) + \sum_{j=1}^{3} \frac{1}{j!} d_{X_i - x}^j m(x) + \frac{1}{4!} \sum_{|\mathbf{i}|=4} \begin{pmatrix} 4 \\ \mathbf{i} \end{pmatrix} (X_i - x)^{\mathbf{i}} \left(\frac{\partial}{\partial y}\right)^{\mathbf{i}} m(y)\Big|_{y=x+t_i(X_i-x)} \\ &= m(x) + \sum_{j=1}^{4} \frac{1}{j!} d_{X_i - x}^j m(x) + \frac{1}{4!} r(x + t_i(X_i - x)) \end{aligned}$$

for some $0 < t_i < 1$, where

$$r(a) = \sum_{|\mathbf{i}|=4} \begin{pmatrix} 4 \\ \mathbf{i} \end{pmatrix} (X_i - x)^{\mathbf{i}} \left(\frac{\partial}{\partial y}\right)^{\mathbf{i}} m(y)\Big|_{y=a} - d_{X_i - x}^4 m(x).$$

We define $b(x) = E[\widehat{m}_{LL}(x,h)|X_1, \cdots, X_n] - m(x)$. Then $b(X_i)$ can be expressed as

$$b(X_i) = b(x) + \nabla b(x)^T (X_i - x) + \frac{1}{2}(X_i - x)^T \nabla^2 b(x + t_i^*(X_i - x))(X_i - x)$$

for some $0 < t_i^* < 1$. For $\alpha,\ \beta \in \mathbb{R}$, we use the notations

$$O_p(h^\alpha n^\beta 1_p) = \begin{bmatrix} O_p(h^\alpha n^\beta) \\ \vdots \\ O_p(h^\alpha n^\beta) \end{bmatrix} \in \mathbb{R}^p,$$

$O_p(h^\alpha n^\beta 1_p^T) = O_p(h^\alpha n^\beta 1_p)^T$ and

$$O_p(h^\alpha n^\beta 1_p 1_p^T) = \begin{bmatrix} O_p(h^\alpha n^\beta) & \cdots & O_p(h^\alpha n^\beta) \\ \vdots & \ddots & \vdots \\ O_p(h^\alpha n^\beta) & \cdots & O_p(h^\alpha n^\beta) \end{bmatrix} \ p \times p \text{ matrix.}$$

Similar to the notations utilized in Section 4, for smooth function $\eta(z)$, we shall use ∂_i and ∂_{ij} to denote the partial derivatives evaluated at a by variables corresponding to those indices i and j throughout in this section:

$$\frac{\partial}{\partial z_i}\eta(z)\bigg|_{z=a} = \partial_i\eta(a) \quad \text{and} \quad \frac{\partial^2}{\partial z_i \partial z_j}\eta(z)\bigg|_{z=a} = \partial_{ij}\eta(a)$$

and so on.

7.1 Proof of Theorem 1

By putting $\boldsymbol{m} = [m(X_1)\cdots m(X_n)]^T$, we have

$$\boldsymbol{m} \;=\; X_x \begin{bmatrix} m(x) \\ \nabla m(x) \end{bmatrix} + \sum_{j=2}^{4}\frac{1}{j!}\begin{bmatrix} d^j_{X_1-x}m(x) \\ \vdots \\ d^j_{X_n-x}m(x) \end{bmatrix} + \frac{1}{4!}\begin{bmatrix} r(x+t_1(X_1-x)) \\ \vdots \\ r(x+t_n(X_n-x)) \end{bmatrix}.$$

Therefore,

$$\begin{aligned}
&E[\widehat{m}_{LL}(x,h)|X_1,\cdots,X_n] \\
&= e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x\boldsymbol{m} \\
&= m(x) + e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x\left(\sum_{j=2}^{4}\frac{1}{j!}\begin{bmatrix} d^j_{X_1-x}m(x) \\ \vdots \\ d^j_{X_n-x}m(x) \end{bmatrix} + \frac{1}{4!}\begin{bmatrix} r(x+t_1(X_1-x)) \\ \vdots \\ r(x+t_n(X_n-x)) \end{bmatrix}\right) \\
&= m(x) + \sum_{j=2}^{4}\frac{1}{j!}R_j(x) + \frac{1}{4!}R(x).
\end{aligned}$$

By $E[Y_i|X_1,\cdots,X_n] = m(X_i)$ and the definition of $b(x)$, we have

$$\begin{aligned}
b(x) &= \sum_{j=2}^{4}\frac{1}{j!}R_j(x) + \frac{1}{4!}R(x), \\
E[r_i|X_1,\cdots,X_n] &= E[Y_i - \widehat{m}_{LL}(X_i,h)|X_1,\cdots,X_n] = -b(X_i).
\end{aligned}$$

By introducing $\boldsymbol{b} = [b(X_1)\cdots b(X_n)]^T$, we see that

$$\boldsymbol{b} = X_x\begin{bmatrix} b(x) \\ \nabla b(x) \end{bmatrix} + \frac{1}{2}Q_b(x) + \frac{1}{2}R_b(x),$$

where

$$
\begin{aligned}
R_b(x) &= \begin{bmatrix} (X_1-x)^T(\nabla^2 b(x+t_1^*(X_1-x)) - \nabla^2 b(x))(X_1-x) \\ \vdots \\ (X_1-x)^T(\nabla^2 b(x+t_n^*(X_1-x)) - \nabla^2 b(x))(X_1-x) \end{bmatrix}, \\
Q_b(x) &= \begin{bmatrix} (X_1-x)^T\nabla^2 b(x)(X_1-x) \\ \vdots \\ (X_1-x)^T\nabla^2 b(x)(X_1-x) \end{bmatrix}.
\end{aligned}
$$

Along the same line as the calculation of $E[\widehat{m}_{LL}(x,h)|X_1,\cdots,X_n]$,

$$
\begin{aligned}
&E[\widehat{r}_{LL}(x,h)|X_1,\cdots,X_n] \\
&= e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x \begin{bmatrix} E[r_1|X_1,\cdots,X_n] \\ \vdots \\ E[r_n|X_1,\cdots,X_n] \end{bmatrix} \\
&= -e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x\boldsymbol{b} \\
&= -e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x \left(X_x \begin{bmatrix} b(x) \\ \nabla b(x) \end{bmatrix} + \frac{1}{2}Q_b(x) + \frac{1}{2}R_b(x) \right) \\
&= -b(x) - \frac{1}{2}e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x(Q_b(x)+R_b(x)).
\end{aligned}
$$

By combining above equalities, we have

$$
\begin{aligned}
E[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n] &= E[\widehat{m}_{LL}(x,h)|X_1,\cdots,X_n] + E[\widehat{r}_{LL}(x,h)|X_1,\cdots,X_n] \\
&= m(x) + b(x) - b(x) - \frac{1}{2}e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x(Q_b(x)+R_b(x)) \\
&= m(x) - \frac{1}{2}e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x(Q_b(x)+R_b(x)).
\end{aligned}
$$

Here our focus goes to $e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xQ_b(x)$. We see that

$$
\begin{aligned}
&\frac{1}{n}X_x^TW_xQ_b(x) \\
&= \frac{1}{n}X_x^T\mathrm{diag}(K_h(X_1-x),\ \cdots,\ K_h(X_n-x)) \begin{bmatrix} (X_1-x)^T\nabla^2 b(x)(X_1-x) \\ \vdots \\ (X_n-x)^T\nabla^2 b(x)(X_n-x) \end{bmatrix}
\end{aligned}
$$

$$= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}(X_i-x)^T\nabla^2 b(x)(X_i-x)K_h(X_i-x)\\(X_i-x)^T\nabla^2 b(x)(X_i-x)K_h(X_i-x)(X_i-x)\end{bmatrix}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}(X_i-x)^T\left(\sum_{j=2}^{4}\frac{1}{j!}\nabla^2 R_j(x)+\frac{1}{4!}\nabla^2 R(x)\right)(X_i-x)K_h(X_i-x)\\(X_i-x)^T\left(\sum_{j=2}^{4}\frac{1}{j!}\nabla^2 R_j(x)+\frac{1}{4!}\nabla^2 R(x)\right)(X_i-x)K_h(X_i-x)(X_i-x)\end{bmatrix}.$$

and note that $R_2(x) = h^2 J_1(x) + o_p(h^2)$ as was shown in [2]. By assumptions in Section 3.1 and the law of large numbers,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^T\nabla^2 R_2(x)(X_i-x)K_h(X_i-x)\\&= h^2\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^T\nabla^2 J_1(x)(X_i-x)K_h(X_i-x)+o_p(h^{d-p+4})\\&= h^{d-p+4}A_1(x)B(x|J_1,\varphi)(1+o_p(1))\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^T\nabla^2 R_2(x)(X_i-x)K_h(X_i-x)(X_i-x)\\&= h^2\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^T\nabla^2 J_1(x)(X_i-x)K_h(X_i-x)(X_i-x)(1+o_p(h^2))\\&= O_p(h^{d-p+5}1_p).\end{aligned}$$

On the other hand, we have from Lemma 2 that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^T\left(\sum_{j=3}^{4}\frac{1}{j!}\nabla^2 R_j(x)+\frac{1}{4!}\nabla^2 R(x)\right)(X_i-x)K_h(X_i-x)\\&= O_p(h^{d-p+5})+O_p(h^{d-p+6})+o_p(h^{d-p+6})\\&= O_p(h^{d-p+5}),\\&\frac{1}{n}\sum_{i=1}^{n}(X_i-x)^T\left(\sum_{j=3}^{4}\frac{1}{j!}\nabla^2 R_j(x)+\frac{1}{4!}\nabla^2 R(x)\right)(X_i-x)K_h(X_i-x)(X_i-x)\\&= O_p(h^{d-p+6}1_p).\end{aligned}$$

Hence, we obtain

$$\frac{1}{n}X_x^T W_x Q_b(x) = h^{d-p+4}\left(\begin{bmatrix} A_1(x)B(x|J_1,\varphi)/2 \\ 0 \end{bmatrix} + o_p(1_{p+1})\right).$$

Using Lemma 1, we finally reach

$$\begin{aligned}
&e_1^T(X_x^T W_x X_x)^{-1}X_x^T W_x Q_b(x) \\
&= e_1^T\left(\frac{1}{n}X_x^T W_x X_x\right)^{-1}\frac{1}{n}X_x^T W_x Q_b(x) \\
&= h^{p-d}\left(\begin{bmatrix} A_1(x)^{-1} & -A_1(x)^{-1}A_2(x)^T A_3(x)^{-1}\end{bmatrix} + \begin{bmatrix} o_p(1) & o_p(1_p^T)\end{bmatrix}\right) \\
&\qquad \times h^{d-p+4}\left(\begin{bmatrix} A_1(x)B(x|J_1,\varphi)/2 \\ 0 \end{bmatrix} + o_p(1_{p+1})\right) \\
&= \frac{h^4}{2}B(x|J_1,\varphi)(1+o_p(1)),
\end{aligned}$$

which gives the bias expression in Theorem 1.

Next, we turn to the variance. We shall introduce the following notation

$$\widehat{m}_{DS}(x,h) \quad = \quad e_1^T(X_x^T W_x X_x)^{-1}X_x^T W_x(2I_n - L)\mathbf{Y},$$

where $L = [L_1 \cdots L_n]^T$ and $L_i^T = e_1^T(X_{X_i}^T W_{X_i} X_{X_i})^{-1}X_{X_i}^T W_{X_i}$. By Lemma 4 with $x = X_i$,

$$L_i = \frac{h^{p-d}}{n}A_1(X_i)^{-1}\begin{bmatrix} (1 - A_2(X_i)^T A_3(X_i)^{-1}(X_1 - X_i))K_h(X_1 - X_i) \\ \vdots \\ (1 - A_2(X_i)^T A_3(X_i)^{-1}(X_n - X_i))K_h(X_n - X_i)\end{bmatrix}(1+o_p(1)).$$

Simple but long calculations give

$$\begin{aligned}
&Var[\widehat{m}_{LL}(x,h)|X_1,\cdots,X_n] \\
&= \frac{1}{nh^d}v(x)A_1(x)^{-2}f(\varphi^{-1}(x))\int_{\mathbb{R}^d} K\left(\mathcal{J}(\varphi^{-1}(x)|\varphi)v\right)^2 dv(1+o_p(1)), && (17)\\
&e_1^T(X_x^T W_x X_x)^{-1}X_x^T W_x L V L^T W_x X_x(X_x^T W_x X_x)^{-1}e_1 \\
&= \frac{1}{nh^d}v(x)A_1(x)^{-4}f(\varphi^{-1}(x))^3\int_{\mathbb{R}^d}\{K*K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)\}^2 dv(1+o_p(1)) && (18)
\end{aligned}$$

and

$$
\begin{aligned}
&e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xVL^TW_xX_x(X_x^TW_xX_x)^{-1}e_1\\
&=\frac{1}{nh^d}v(x)A_1(x)^{-3}f(\varphi^{-1}(x))^2\\
&\quad\times\int_{\mathbb{R}^d}K\left(\mathcal{J}(\varphi^{-1}(x)|\varphi)v\right)K*K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)dv(1+o_p(1)).
\end{aligned}
\tag{19}
$$

Using (17), (18) and (19), we have

$$
\begin{aligned}
&Var[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n]\\
&=4e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xVW_xX_x(X_x^TW_xX_x)^{-1}e_1\\
&\quad-4e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xVL^TW_xX_x(X_x^TW_xX_x)^{-1}e_1\\
&\quad+e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xLVL^TW_xX_x(X_x^TW_xX_x)^{-1}e_1\\
&=\frac{1}{nh^d}v(x)A_1(x)^{-2}f(\varphi^{-1}(x))\\
&\quad\times\left(\int_{\mathbb{R}^d}4K\left(\mathcal{J}(\varphi^{-1}(x)|\varphi)v\right)^2dv+\int_{\mathbb{R}^d}A_1(x)^{-2}f(\varphi^{-1}(x))^2K*K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)^2dv\right.\\
&\qquad\left.-\int_{\mathbb{R}^d}4A_1(x)^{-1}f(\varphi^{-1}(x))K\left(\mathcal{J}(\varphi^{-1}(x)|\varphi)v\right)K*K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)dv\right)(1+o_p(1))\\
&=\frac{1}{nh^d}\frac{R(x|K^*,\varphi)}{R(x|K,\varphi)}J_2(x)(1+o_p(1)),
\end{aligned}
$$

which is the variance in Theorem 1. □

7.2 Proof of Theorem 2

We put $u(x)^T=e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_x,\boldsymbol{\varepsilon}=[\varepsilon_1\cdots\varepsilon_n]^T$ and $U_n=[u(X_1)\cdots u(X_n)]^T$. Then we have $\widehat{m}_{DS}(x,h)=u(x)^T(2I_n-U_n)\mathbf{Y}$ and $\mathbf{Y}=\boldsymbol{m}+\boldsymbol{\varepsilon}$. By $E[\mathbf{Y}|X_1,\cdots,X_n]=\boldsymbol{m}$, we have

$$
\begin{aligned}
\widehat{m}_{DS}(x,h)&=u(x)^T(2I_n-U_n)(\boldsymbol{m}+\boldsymbol{\varepsilon})\\
&=u(x)^T(2I_n-U_n)\boldsymbol{m}+u(x)^T(2I_n-U_n)\boldsymbol{\varepsilon}\\
&=E[u(x)^T(2I_n-U_n)\mathbf{Y}|X_1,\cdots,X_n]+u(x)^T(2I_n-U_n)\boldsymbol{\varepsilon}\\
&=E[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n]+u(x)^T(2I_n-U_n)\boldsymbol{\varepsilon}\\
&=m(x)+\frac{h^4}{4}B(x|J_1,\varphi)+u(x)^T(2I_n-U_n)\boldsymbol{\varepsilon}+o_p(h^4).
\end{aligned}
$$

By plugging above expressions as well as $\sqrt{nh^d}o_p(h^4) = o_p(1)$, we have

$$\sqrt{nh^d}\left(\widehat{m}_{DS}(x,h) - m(x) - \frac{h^4}{4}B(x|J_1,\varphi)\right) = \sqrt{nh^d}u(x)^T(2I_n - U_n)\boldsymbol{\varepsilon} + o_p(1).$$

Let $\nu_{ni}(x)$ be the i-th element of $\sqrt{nh^d}u(x)^T(2I_n - U_n)$, then

$$\begin{aligned}
\sqrt{nh^d}u(x)^T(2I_n - U_n)\boldsymbol{\varepsilon} &= \sum_{i=1}^{n}\nu_{ni}(x)\varepsilon_i, \\
E\left[\sum_{i=1}^{n}\nu_{ni}(x)\varepsilon_i|X_1,\cdots,X_n\right] &= 0
\end{aligned}$$

and

$$Var[\nu_{ni}(x)\varepsilon_i|X_1,\cdots,X_n] = \nu_{ni}(x)^2.$$

We would apply Liapounoff's central limit theorem to $\{\nu_{ni}(x)\varepsilon_i\}_{i=1}^n$, so that we aim to show that $\{\nu_{ni}(x)\varepsilon_i\}_{i=1}^n$ satisfy the Liapounoff's condition. The summation in the denominator is

$$\begin{aligned}
\sum_{i=1}^{n} Var\left[\nu_{ni}(x)\varepsilon_i|X_1,\cdots,X_n\right] &= Var\left[\sum_{i=1}^{n}\nu_{ni}(x)\varepsilon_i|X_1,\cdots,X_n\right] \\
&= nh^d Var[u(x)^T(2I_n - U_n)\boldsymbol{\varepsilon}|X_1,\cdots,X_n] \\
&= nh^d Var[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n] \\
&= O_p(1).
\end{aligned}$$

Let $u_i(x)$ be the i-th element of $u(x)$. By the definition of $\nu_{ni}(x)$, we get

$$\begin{aligned}
\nu_{ni}(x) &= \sqrt{nh^d}\left((2I_n - U_n)^T\right)_i u(x) \\
&= \sqrt{nh^d}\left(\begin{bmatrix} 2\mathrm{e}_1 & \cdots & 2\mathrm{e}_n \end{bmatrix} - \begin{bmatrix} u(X_1) & \cdots & u(X_n) \end{bmatrix}\right)_i u(x) \\
&= \sqrt{nh^d}\left(2\mathrm{e}_i^T - \begin{bmatrix} u_i(X_1) & \cdots & u_i(X_n) \end{bmatrix}\right) u(x) \\
&= \sqrt{nh^d}\left(2u_i(x) - u(x)^T\begin{bmatrix} u_i(X_1) \\ \vdots \\ u_i(X_n) \end{bmatrix}\right),
\end{aligned}$$

where e_i is n-dimensional vector with its i-th element 1 and 0 otherwise. Similar calculations to obtain the bias of the LL estimator based on $[u_i(X_1) \cdots u_i(X_n)]^T$ yield that

$$
\begin{aligned}
u(x)^T \begin{bmatrix} u_i(X_1) \\ \vdots \\ u_i(X_n) \end{bmatrix} &= u_i(x) + \frac{1}{2} h^2 S_i(x) + o_p\left(\frac{h^{p-d+2}}{n}\right) \\
&= u_i(x) + \frac{1}{2}\frac{h^{p-d+2}}{n}\frac{n}{h^{p-d}} S_i(x) + o_p\left(\frac{h^{p-d+2}}{n}\right),
\end{aligned}
$$

where

$$
S_i(x) = A_1(x)^{-1} \int_{\mathbb{R}^d} u^T \mathcal{J}(\varphi^{-1}(x)|\varphi)^T \nabla^2 u_i(x) \mathcal{J}(\varphi^{-1}(x)|\varphi) u K(\mathcal{J}(\varphi^{-1}(x)|\varphi)u) du.
$$

Lemma 4 implies that

$$
u_i(x) = \frac{h^{p-d}}{n} A_1(x)^{-1}(1 - A_2(x)^T A_3(x)^{-1}(X_i - x)) K_h(X_i - x)(1 + o_p(1))
$$

and $(n/n^{p-d})S_i(x) = O_p(1)$ holds for all $i (1 \leq i \leq n)$. Thus, we obtain

$$
2u_i(x) - u(x)^T \begin{bmatrix} u_i(X_1) \\ \vdots \\ u_i(X_n) \end{bmatrix} = u_i(x) + O_p\left(\frac{h^{p-d+2}}{n}\right).
$$

By Lemma 5 and assumptions in Section 3.1, we have

$$
\begin{aligned}
&\sum_{i=1}^n E\left[\{\nu_{ni}(x)\varepsilon_i\}^4 | X_1, \cdots, X_n\right] \\
&= (nh^d)^2 \frac{h^{4(p-d)}}{n^3} \frac{1}{n} \sum_{i=1}^n \left\{A_1(x)^{-1}(1 - A_2(x)^T A_3(x)^{-1}(X_i - x)) K_h(X_i - x)\right\}^4 \\
&\quad \times E\left[\varepsilon_i^4 | X_1, \cdots, X_n\right](1 + o(1)) + (nh^d)^2 O_p\left(\frac{h^{4p-4d+8}}{n^3}\right) \\
&= E[\varepsilon^4] \frac{h^{4p-2d}}{n} \frac{1}{n} \sum_{i=1}^n \left\{A_1(x)^{-1}(1 - A_2(x)^T A_3(x)^{-1}(X_i - x))\right\}^4 K_h(X_i - x)^4 \\
&\quad \times (1 + o(1)) + o_p\left(\frac{h^{4p-2d}}{n}\right) \\
&= E[\varepsilon^4] \frac{h^{4p-2d}}{n} \left(h^{-4p+d} f(\varphi^{-1}(x)) + O_p(h^{-4p+d+1})\right) + o_p\left(\frac{h^{4p-2d}}{n}\right) \\
&= \frac{1}{nh^d} E[\varepsilon^4] f(\varphi^{-1}(x)) + O_p\left(\frac{h}{nh^d}\right) + o_p\left(\frac{h^{4p-d}}{nh^d}\right) \\
&\to 0,
\end{aligned}
$$

as $n \to \infty$. Therefore, Liapounoff's condition

$$\lim_{n\to\infty} E\left[\sum_{i=1}^{n}\{\nu_{ni}(x)\varepsilon_i\}^4 \middle| X_1,\cdots,X_n\right] \Bigg/ \left(\sum_{i=1}^{n} Var[\nu_{ni}(x)\varepsilon_i|X_1,\cdots,X_n]\right)^4 = 0$$

has been confirmed by choosing $\delta = 2$, from which it follows that

$$\frac{\sum_{i=1}^{n}\nu_{ni}(x)\varepsilon_i}{\sqrt{\sum_{i=1}^{n} Var[\nu_{ni}(x)\varepsilon_i]}} = \frac{\sqrt{nh^d}u(x)(2I_n - U_n)\boldsymbol{\varepsilon}}{\sqrt{Var[\sqrt{nh^d}u(x)(2I_n - U_n)\boldsymbol{\varepsilon}|X_1,\cdots,X_n]}} \xrightarrow{D} N(0,1). \quad \square$$

References

[1] Aswani, A., Bickel, P. and Tomlin, C. (2011). Regression on manifolds: estimation of the exterior derivative, *The Annals of Statistics*, **39**, 48-81.

[2] Bickel, P. and Li, B. (2007). Local polynomial regression on unknown manifolds, *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. Institute of Mathematical Statistics Lecture Notes-Monograph Series*, **54**, 177-186.

[3] Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324-339.

[4] Cheng, M.-Y. and Wu, T. (2013). Local linear regression on manifolds and its geometric interpretation, *Journal of the American Statistical Association*, **108**, 1421-1434.

[5] Choi, E., Hall, P. and Rousson, V. (2000). Data sharpening methods for bias reduction in nonparametric regression, *The Annals of Statistics*, **28**, 1339-1355.

[6] Di Marzio, M., Panzera, A. and Taylor, C. C. (2009). Local polynomial regression for circular predictors, *Statistics and Probability Letters*, **79**, 2066-2075.

[7] Di Marzio, M., Panzera, A. and Taylor, C. C. (2014). Nonparametric regression for spherical data, *Journal of the American Statistical Association*, **109**, 748-763.

[8] Mardia, K. V. (1972). *Statistics of directional data*, John Wiley, New York.

[9] Naito, K. and Yoshizaki, M. (2009). Bandwidth selection for a data sharpening estimator in nonparametric regression, *Journal of Multivariate Analysis*, **100**, 1465-1486.

[10] R Development Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

[11] Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression, *The Annals of Statistics*, **22**, 1346-1370.

[12] Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Philippines.

[13] Yoshizaki, M. and Naito, K. (2004). Practical aspects of bias reducing estimators in nonparametric regression, *Japanese Journal of Applied Statistics*, **33**, 131-155 (in Japanese).

SUPPLEMENTARY MATERIAL

to

"Data Sharpening on Unknown Manifold"

Masaki Kudo and Kanta Naito

In this appendix we provide the following Lemmas to complete the proof of Theorems.

LEMMA 1. Let $x \in \mathcal{X}$. If the assumptions in Section 3.1 hold, then as $n \to \infty$ and $h \to 0$,

$$\left(\frac{1}{n} X_x^T W_x X_x\right)^{-1}$$

$$= h^{p-d}\left(\begin{bmatrix} A_1(x)^{-1} & -A_1(x)^{-1}A_2(x)^T A_3(x)^{-1} \\ -A_1(x)^{-1}A_3(x)^{-1}A_2(x) & h^{-2}A_3(x)^{-1} \end{bmatrix} + \begin{bmatrix} o_p(1) & o_p(1_p^T) \\ o_p(1_p) & o_p(h^{-2}1_p 1_p^T) \end{bmatrix}\right).$$

Proof. Direct but lengthy calculations yield that

$$\begin{aligned} E\left[K_h(X-x)\right] &= h^{d-p}A_1(x) + o(h^{d-p}), \\ E\left[K_h(X-x)(X-x)\right] &= h^{d-p+2}A_2(x) + o(h^{d-p+2}1_p) \end{aligned}$$

and

$$E\left[K_h(X-x)(X-x)(X-x)^T\right] = h^{d-p+2}A_3(x) + o(h^{d-p+2}1_p 1_p^T).$$

Using above results, we obtain

$$\begin{aligned} \frac{1}{n}X_x^T W_x X_x &= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} K_h(X_i-x) & K_h(X_i-x)(X_i-x) \\ K_h(X_i-x)(X_i-x)^T & K_h(X_i-x)(X_i-x)(X_i-x)^T \end{bmatrix} \\ &= \begin{bmatrix} h^{d-p}(A_1(x)+o_p(1)) & h^{2+d-p}(A_2(x)^T + o_p(1_p^T)) \\ h^{2+d-p}(A_2(x)+o_p(1_p)) & h^{2+d-p}(A_3(x)+o_p(1_p 1_p^T)) \end{bmatrix} \\ &= h^{d-p}\left(\begin{bmatrix} A_1(x) & h^2 A_2(x)^T \\ h^2 A_2(x) & h^2 A_3(x) \end{bmatrix} + \begin{bmatrix} o_p(1) & o_p(h^2 1_p^T) \\ o_p(h^2 1_p) & o_p(h^2 1_p 1_p^T) \end{bmatrix}\right). \end{aligned}$$

By using $(M + o(1_{p+1}1_{p+1}^T))^{-1} = M^{-1} + o(1_{p+1}1_{p+1}^T)$ for $(p+1)\times(p+1)$ nonsingular M,

$$\left(\frac{1}{n}X_x^T W_x X_x\right)^{-1} = h^{p-d}\left(\begin{bmatrix} A_1(x) & h^2 A_2(x)^T \\ h^2 A_2(x) & h^2 A_3(x) \end{bmatrix} + \begin{bmatrix} o_p(1) & o_p(h^2 1_p^T) \\ o_p(h^2 1_p) & o_p(h^2 1_p 1_p^T) \end{bmatrix}\right)^{-1}$$

$$= h^{p-d}\left(\begin{bmatrix} A_1(x) & h^2A_2(x)^T \\ h^2A_2(x) & h^2A_3(x)\end{bmatrix}^{-1} + \begin{bmatrix} o_p(1) & o_p(1_p^T) \\ o_p(1_p) & o_p(1_p1_p^T)\end{bmatrix}\right),$$

where

$$\begin{bmatrix} A_1(x) & h^2A_2(x)^T \\ h^2A_2(x) & h^2A_3(x)\end{bmatrix}^{-1} = \begin{bmatrix} A_1(x)^{-1} + h^4A_1(x)^{-2}A_2(x)^TS^{-1}A_2(x) & -h^2A_1(x)^{-1}A_2(x)^TS^{-1} \\ -h^2S^{-1}A_2(x)A_1(x)^{-1} & S^{-1}\end{bmatrix}$$

and $S = h^2A_3(x) - h^4A_1(x)^{-1}A_2(x)A_2(x)^T$. Using an asymptotic evaluation of S^{-1} as

$$S^{-1} = h^{-2}(A_3(x) - h^2A_1(x)^{-1}A_2(x)A_2(x)^T)^{-1} = h^{-2}(A_3(x)^{-1} + o_p(1_p1_p^T)),$$

we have

$$\begin{aligned} h^4A_1(x)^{-2}A_2(x)^TS^{-1}A_2(x) &= h^2A_1(x)^{-2}A_2(x)^T(A_3(x)^{-1} + o_p(1_p1_p^T))A_2(x) \\ &= h^2A_1(x)^{-2}A_2(x)^TA_3(x)^{-1}A_2(x) + o_p(h^21_p1_p^T)) \\ &= O_p(h^2) \end{aligned}$$

and

$$\begin{aligned} -h^2A_1(x)^{-1}A_2(x)^TS^{-1} &= -A_1(x)^{-1}A_2(x)^T(A_3(x)^{-1} + o_p(1_p1_p^T)) \\ &= -A_1(x)^{-1}A_2(x)^TA_3(x)^{-1} + o_p(1_p^T). \end{aligned}$$

Above calculations are combined into

$$\begin{aligned} &\begin{bmatrix} A_1(x)^{-1} + h^4A_1(x)^{-2}A_2(x)^TS^{-1}A_2(x) & -h^2A_1(x)^{-1}A_2(x)^TS^{-1} \\ -h^2S^{-1}A_2(x)A_1(x)^{-1} & S^{-1}\end{bmatrix} \\ &= \begin{bmatrix} A_1(x)^{-1} + O_p(h^2) & -A_1(x)^{-1}A_2(x)^TA_3(x)^{-1} + o_p(1_p^T) \\ -A_1(x)^{-1}A_3(x)^{-1}A_2(x) + o_p(1_p) & h^{-2}A_3(x)^{-1} + o_p(h^{-2}1_p1_p^T)\end{bmatrix} \\ &= \begin{bmatrix} A_1(x)^{-1} & -A_1(x)^{-1}A_2(x)^TA_3(x)^{-1} \\ -A_1(x)^{-1}A_3(x)^{-1}A_2(x) & h^{-2}A_3(x)^{-1}\end{bmatrix} + \begin{bmatrix} O_p(h^2) & o_p(1_p^T) \\ o_p(1_p) & o_p(h^{-2}1_p1_p^T)\end{bmatrix}, \end{aligned}$$

from which, we finally obtain

$$\begin{aligned} \left(\frac{1}{n}X_x^TW_xX_x\right)^{-1} = h^{p-d}&\left(\begin{bmatrix} A_1(x)^{-1} & -A_1(x)^{-1}A_2(x)^TA_3(x)^{-1} \\ -A_1(x)^{-1}A_3(x)^{-1}A_2(x) & h^{-2}A_3(x)^{-1}\end{bmatrix}\right. \\ &\left. + \begin{bmatrix} o_p(1) & o_p(1_p^T) \\ o_p(1_p) & o_p(h^{-2}1_p1_p^T)\end{bmatrix}\right). \qquad \square \end{aligned}$$

LEMMA 2. Let $x \in \mathcal{X}$. Under the assumptions in Section 3.1, we have for $j = 3, 4$,

$$e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x R_b(x) = o_p(h^4), \tag{20}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - x)^T \nabla^2 R(x)(X_i - x)K_h(X_i - x) = o_p(h^{d-p+6}), \tag{21}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - x)^T \nabla^2 R(x)(X_i - x)K_h(X_i - x)(X_i - x) = o_p(h^{d-p+7}1_p), \tag{22}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - x)^T \nabla^2 R_j(x)(X_i - x)K_h(X_i - x) = O_p(h^{d-p+j+2}), \tag{23}$$

where

$$R(x) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x \begin{bmatrix} r(x + t_1(X_1 - x)) \\ \vdots \\ r(x + t_n(X_n - x)) \end{bmatrix},$$

$$R_j(x) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x \begin{bmatrix} d^j_{X_1 - x} m(x) \\ \vdots \\ d^j_{X_n - x} m(x) \end{bmatrix}.$$

Proof. To prove (9), we start the following calculations of $X_x^T W_x R_b(x)/n$:

$$\begin{aligned}
&\frac{1}{n} X_x^T W_x R_b(x) \\
&= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} (X_i - x)^T(\nabla^2 b(x + t_i^*(X_i - x)) - \nabla^2 b(x))(X_i - x)K_h(X_i - x) \\ (X_i - x)^T(\nabla^2 b(x + t_i^*(X_i - x)) - \nabla^2 b(x))(X_i - x)K_h(X_i - x)(X_i - x) \end{bmatrix} \\
&= \sum_{\alpha,\beta=1}^{p}\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} w_{(\alpha\beta)}(X_i, t_i^*)K_h(X_i - x) \\ w_{(\alpha\beta)}(X_i, t_i^*)K_h(X_i - x)(X_i - x) \end{bmatrix},
\end{aligned}$$

where we put

$$w_{(\alpha\beta)}(y, t) = \{\partial_{\alpha\beta} b(x + t(y - x)) - \partial_{\alpha\beta} b(x)\}(y_\alpha - x_\alpha)(y_\beta - x_\beta)$$

for $y \in \mathcal{X}$ and $0 < t < 1$. To derive the result of convergence in probability, we would calculate

$$E\left[\frac{1}{n}\sum_{i=1}^{n} w_{(\alpha\beta)}(X_i, t_i^*)K_h(X_i - x)\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[w_{(\alpha\beta)}(X_i, t_i^*)K_h(X_i - x)\right]$$

and

$$E\left[\frac{1}{n}\sum_{i=1}^{n} w_{(\alpha\beta)}(X_i,t_i^*)K_h(X_i-x)(X_i-x)\right]=\frac{1}{n}\sum_{i=1}^{n} E\left[w_{(\alpha\beta)}(X_i,t_i^*)K_h(X_i-x)(X_i-x)\right].$$

We shall check that $w_{(\alpha\beta)}(y,t)$ is satisfying $|w_{(\alpha\beta)}(y,t)| \le C(1+||y||^2)$ for some $C>0$. Since the function m is C^4, the function b is also C^4 and hence we have $|\partial_{\alpha\beta}b(x+t(y-x))-\partial_{\alpha\beta}b(x)| \le M$ for the maximum value M because above first term $\partial_{\alpha\beta}b(x+t(y-x))$ has continuous function with y on the compact region $\mathcal{X}$. Using evaluations

$$|y_\alpha y_\beta| \le \frac{1}{2}(y_\alpha+y_\beta) \le \frac{1}{2}||y||^2$$

and

$$|x_\alpha y_\beta| \le |x_\alpha| \max\{||y||^2,1\} \le |x_\alpha|(1+||y||^2),$$

we observe that

$$\begin{aligned}
|w_{(\alpha\beta)}(y,t)| &= |\partial_{\alpha\beta}b(x+t(y-x))-\partial_{\alpha\beta}b(x)||(y_\alpha-x_\alpha)(y_\beta-x_\beta)| \\
&\le M(|y_\alpha y_\beta|+|x_\alpha y_\beta|+|y_\alpha x_\beta|+|x_\alpha x_\beta|) \\
&\le C(1+||y||^2)
\end{aligned}$$

for $C=M\max_{\alpha\beta}\{|x_\alpha|,|x_\alpha x_\beta|,1/2\}$. So we have just finished to check that assumption 6 in Section 3.1 holds for $w_{(\alpha\beta)}(y,t)$ and therefore we can proceed to evaluate the expectation for $w_{(\alpha\beta)}(y,t)$ in the sequel. We see that $b(x)=h^2J_1(x)+o_p(h^2)$ holds by the results in [2], and $y \in \mathcal{B}^p_{x,h^{1-\eta}}$ tends to x as h tends to zero because $||y-x||<h^{1-\eta}$ holds, hence we have $\partial_{\alpha\beta}b(x+t(y-x))-\partial_{\alpha\beta}b(x)=o(h^2)$ for $y \in \mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X}$. We obtain from evaluations above

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} E\left[w_{(\alpha\beta)}(X_i,t_i^*)K_h(X_i-x)\right] \\
&= E\left[w_{(\alpha\beta)}(X_1,t_1^*)K_h(X_1-x)1(X_1 \in \mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})\right] \\
&\quad +E\left[w_{(\alpha\beta)}(X_1,t_1^*)K_h(X_1-x)1(X_1 \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})^c)\right]
\end{aligned}$$

$$
\begin{aligned}
&= E\Bigg[\{\partial_{\alpha\beta} b(x+t_1^*(X_1-x)) - \partial_{\alpha\beta} b(x)\}(X_1-x)_\alpha (X_1-x)_\beta \\
&\qquad \times K_h(X_1-x) 1(X_1 \in \mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})\Bigg] + o(h^{d-p+4}) \\
&= h^{d-p+4} \int_{\mathbb{R}^d} \frac{\partial_{\alpha\beta} b(\varphi(0)+t_1^*(\varphi(hu)-\varphi(0))) - \partial_{\alpha\beta} b(\varphi(0))}{h^2} \frac{\varphi_\alpha(hu)-\varphi_\alpha(0)}{h} \\
&\qquad \times \frac{\varphi_\beta(hu)-\varphi_\beta(0)}{h} K\left(\frac{\varphi(hu)-\varphi(0)}{h}\right) 1(u \in \mathcal{B}^d_{0,h^{-\eta}}) f(hu) du + o(h^{d-p+4}) \\
&= o(h^{d-p+4}).
\end{aligned}
$$

Furthermore we have

$$
\frac{1}{n}\sum_{i=1}^{n} E\left[w_{(\alpha\beta)}(X_i, t_i^*) K_h(X_i-x)(X_i-x)\right] \quad = \quad o(h^{d-p+5} 1_p),
$$

holds because $|w_{(\alpha\beta)}(y,t)(y_\gamma - x_\gamma)| \leq C(1+||y||^2)|y_\gamma - x_\gamma| < C(1+||y||^2)^2$ for each $\gamma \in \{1, \cdots, p\}$. Hence, we have (20) by the following calculations:

$$
\begin{aligned}
& e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x R_b(x) \\
&= e_1^T \left(\frac{1}{n} X_x^T W_x X_x\right)^{-1} \frac{1}{n} X_x^T W_x \begin{bmatrix} (X_1-x)^T(\nabla^2 b(x+t_1^*(X_1-x)) - \nabla^2 b(x))(X_1-x) \\ \vdots \\ (X_n-x)^T(\nabla^2 b(x+t_n^*(X_n-x)) - \nabla^2 b(x))(X_n-x) \end{bmatrix} \\
&= O_p(h^{p-d} 1_{p+1}^T) \begin{bmatrix} o_p(h^{d-p+4}) \\ o_p(h^{d-p+5} 1_p) \end{bmatrix} \\
&= o_p(h^4).
\end{aligned}
$$

Next, our focus goes to

$$
R(x) \quad = \quad e_1^T \left(\frac{1}{n} X_x^T W_x X_x\right)^{-1} \frac{1}{n} X_x^T W_x \begin{bmatrix} r(x+t_1(X_1-x)) \\ \vdots \\ r(x+t_n(X_n-x)) \end{bmatrix}.
$$

It follows that

$$
\frac{1}{n} X_x^T W_x \begin{bmatrix} r(x+t_1(X_1-x)) \\ \vdots \\ r(x+t_n(X_n-x)) \end{bmatrix}
$$

$$= \frac{1}{n} X_x^T \mathrm{diag}(K_h(X_1 - x), \ \cdots, \ K_h(X_n - x)) \begin{bmatrix} r(x + t_1(X_1 - x)) \\ \vdots \\ r(x + t_n(X_n - x)) \end{bmatrix}$$

$$= \sum_{\alpha,\beta,\gamma,\delta=1}^{p} \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} r_{(\alpha\beta\gamma\delta)}(X_i, t_i) K_h(X_i - x) \\ r_{(\alpha\beta\gamma\delta)}(X_i, t_i) K_h(X_i - x)(X_i - x) \end{bmatrix},$$

where $r_{(\alpha\beta\gamma\delta)}(y,t) = D_{(\alpha\beta\gamma\delta)}(y,t)(y_\alpha - x_\alpha)(y_\beta - x_\beta)(y_\gamma - x_\gamma)(y_\delta - x_\delta)$ and $D_{(\alpha\beta\gamma\delta)}(y,t) = \partial_{\alpha\beta\gamma\delta} m(x + t(y - x)) - \partial_{\alpha\beta\gamma\delta} m(x)$ for $y \in \mathcal{X}$ and $0 < t < 1$. We shall evaluate $E[r_{(\alpha\beta\gamma\delta)}(X_1, t_1) K_h(X_1 - x)]$. By the evaluation same as $w_{(\alpha\beta)}(y,t)$, we obtain

$$|r_{(\alpha\beta\gamma\delta)}(y,t)| \le N |w_{(\alpha\beta)}(y,t) w_{(\gamma\delta)}(y,t)| \le N C^2 (1 + ||y||^2)^2$$

for some $N > 0$. By assumptions in Section 3.1, we see that

$$\begin{aligned}
& E\left[r_{(\alpha\beta\gamma\delta)}(X_1, t_1) K_h(X_1 - x)\right] \\
&= E\left[r_{(\alpha\beta\gamma\delta)}(X_1, t_1) K_h(X_1 - x) 1(X_1 \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X}))\right] \\
&\quad + E\left[r_{(\alpha\beta\gamma\delta)}(X_1, t_1) K_h(X_1 - x) 1(X_1 \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})^c)\right] \\
&= E\left[r_{(\alpha\beta\gamma\delta)}(\varphi(Z), t_1) K_h(\varphi(Z) - \varphi(0)) 1(\varphi(Z) \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X}))\right] \\
&\quad + E\left[r_{(\alpha\beta\gamma\delta)}(\varphi(Z), t_1) K_h(\varphi(Z) - \varphi(0)) 1(\varphi(Z) \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})^c)\right] \\
&= h^{-p+4} E\Bigg[D_{(\alpha\beta\gamma\delta)}(\varphi(Z), t_1) \frac{\varphi_\alpha(Z) - \varphi_\alpha(0)}{h} \frac{\varphi_\beta(Z) - \varphi_\beta(0)}{h} \frac{\varphi_\gamma(Z) - \varphi_\gamma(0)}{h} \\
&\qquad \times \frac{\varphi_\delta(Z) - \varphi_\delta(0)}{h} K\left(\frac{\varphi(Z) - \varphi(0)}{h}\right) 1(Z \in \mathcal{B}^d_{0,h^{1-\eta}})\Bigg] + o(h^{d-p+4}).
\end{aligned}$$

Since

$$D_{(\alpha\beta\gamma\delta)}(\varphi(hu), t_1) = \partial_{\alpha\beta\gamma\delta} m(x + t_1(\varphi(hu) - \varphi(0))) - \partial_{\alpha\beta\gamma\delta} m(x) = o(1),$$

as $h \to 0$ and by noting that $(\varphi_\alpha(hu) - \varphi_\alpha(0))/h = O(1)$, we obtain

$$\begin{aligned}
& E\Bigg[D_{(\alpha\beta\gamma\delta)}(\varphi(Z), t_1) \frac{\varphi_\alpha(Z) - \varphi_\alpha(0)}{h} \frac{\varphi_\beta(Z) - \varphi_\beta(0)}{h} \frac{\varphi_\gamma(Z) - \varphi_\gamma(0)}{h} \frac{\varphi_\delta(Z) - \varphi_\delta(0)}{h} \\
&\quad \times K\left(\frac{\varphi(Z) - \varphi(0)}{h}\right) 1(Z \in \mathcal{B}^d_{0,h^{1-\eta}})\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathcal{B}^d_{0,h^{1-\eta}}} D_{(\alpha\beta\gamma\delta)}(\varphi(z), t_1) K\left(\frac{\varphi(z)-\varphi(0)}{h}\right) \\
&\quad \times \frac{\varphi_\alpha(z)-\varphi_\alpha(0)}{h}\frac{\varphi_\beta(z)-\varphi_\beta(0)}{h}\frac{\varphi_\gamma(z)-\varphi_\gamma(0)}{h}\frac{\varphi_\delta(z)-\varphi_\delta(0)}{h} f(z)dz \\
&= h^d \int_{\mathcal{B}^d_{0,h^{-\eta}}} D_{(\alpha\beta\gamma\delta)}(\varphi(hu), t_1) K\left(\frac{\varphi(hu)-\varphi(0)}{h}\right) \\
&\quad \times \frac{\varphi_\alpha(hu)-\varphi_\alpha(0)}{h}\frac{\varphi_\beta(hu)-\varphi_\beta(0)}{h}\frac{\varphi_\gamma(hu)-\varphi_\gamma(0)}{h}\frac{\varphi_\delta(hu)-\varphi_\delta(0)}{h} f(hu)du \\
&= o(h^d).
\end{aligned}
$$

Thus, we have

$$
E\left[\frac{1}{n}\sum_{i=1}^n r(x+t_i(X_i-x))K_h(X_i-x)\right] = o(h^{d-p+4}).
$$

By the law of large numbers,

$$
\frac{1}{n}\sum_{i=1}^n r((x+t_i(X_i-x))K_h(X_i-x) = o_p(h^{d-p+4})
$$

holds. From the same calculations, we have

$$
\frac{1}{n}\sum_{i=1}^n r(x+t_i(X_i-x))K_h(X_i-x)(X_i-x) = o_p(h^{d-p+5}1_p)
$$

because there exists $N > 0$ such that $|r_{(\alpha\beta\gamma\delta)}(y,t)(y_\epsilon - x_\epsilon)| \le N(1+||y||^2)^3$ for all $\alpha, \beta, \gamma, \delta, \epsilon \in \{1, \cdots, p\}$. So we obtain

$$
\begin{aligned}
R(x) &= e_1^T \left(\frac{1}{n}X_x^T W_x X_x\right)^{-1} \frac{1}{n} X_x^T W_x \begin{bmatrix} r(x+t_1(X_1-x)) \\ \vdots \\ r(x+t_n(X_n-x)) \end{bmatrix} \\
&= e_1^T \left(\frac{1}{n}X_x^T W_x X_x\right)^{-1} \sum_{\alpha,\beta,\gamma,\delta=1}^{p} \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} r_{(\alpha\beta\gamma\delta)}(X_i,t_i)K_h(X_i-x) \\ r_{(\alpha\beta\gamma\delta)}(X_i,t_i)K_h(X_i-x)(X_i-x) \end{bmatrix} \\
&= e_1^T O_p(h^{p-d}1_{p+1}1_{p+1}^T) \begin{bmatrix} o_p(h^{d-p+4}) \\ o_p(h^{d-p+5}1_p) \end{bmatrix} \\
&= o_p(h^4).
\end{aligned}
$$

Therefore, we finally have (21) and (22).

To prove (23), we shall evaluate $R_j(x)$ for $j = 3, 4$. It is easy to confirm that

$$\frac{1}{n} X_x^T W_x \begin{bmatrix} d^j_{X_1 - x} m(x) \\ \vdots \\ d^j_{X_n - x} m(x) \end{bmatrix} = \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} d^j_{X_i - x} m(x) K_h(X_i - x) \\ d^j_{X_i - x} m(x) K_h(X_i - x)(X_i - x) \end{bmatrix}$$

hence, we obtain

$$\begin{aligned}
& E\left[\frac{1}{n}\sum_{i=1}^n d^j_{X_i - x} m(x) K_h(X_i - x)\right] \\
&= E\left[d^j_{X_1 - x} m(x) K_h(X_1 - x) 1(X_1 \in \mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})\right] \\
&\quad + E\left[d^j_{X_1 - x} m(x) K_h(X_1 - x) 1(X_1 \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})^c)\right] \\
&= h^{-p} E\left[d^j_{\varphi(Z) - \varphi(0)} m(x) K\left(\frac{\varphi(Z) - \varphi(0)}{h}\right) 1(Z \in \mathcal{B}^d_{0,h^{1-\eta}})\right] + o(h^{d-p+4}) \\
&= h^{d-p+j} \int_{\mathbb{R}^d} \sum_{|\mathbf{i}|=j} \begin{pmatrix} j \\ \mathbf{i} \end{pmatrix} \left(\frac{\varphi(hu) - \varphi(0)}{h}\right)^{\mathbf{i}} \left(\frac{\partial}{\partial x}\right)^{\mathbf{i}} m(x) \\
&\qquad \times K\left(\frac{\varphi(hu) - \varphi(0)}{h}\right) 1(u \in \mathcal{B}^d_{0,h^{-\eta}}) f(hu) du + o(h^{d-p+4})
\end{aligned}$$

and

$$\begin{aligned}
& E\left[\frac{1}{n}\sum_{i=1}^n d^j_{X_i - x} m(x) K_h(X_i - x)(X_i - x)\right] \\
&= h^{d-p+j+1} \int_{\mathbb{R}^d} \sum_{|\mathbf{i}|=j} \begin{pmatrix} j \\ \mathbf{i} \end{pmatrix} \left(\frac{\varphi(hu) - \varphi(0)}{h}\right)^{\mathbf{i}} \left(\frac{\partial}{\partial x}\right)^{\mathbf{i}} m(x) \\
&\qquad \times K\left(\frac{\varphi(hu) - \varphi(0)}{h}\right) 1(u \in \mathcal{B}^d_{0,h^{-\eta}}) f(hu) \frac{\varphi(hu) - \varphi(0)}{h} du + o(h^{d-p+4} 1_p),
\end{aligned}$$

where we have used the assumption 6 in Section 3.1. By using $(\varphi_\alpha(hu) - \varphi_\alpha(0))/h = O(1)$, we get

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n d^3_{X_i - x} m(x) K_h(X_i - x) &= O_p(h^{d-p+3}), \\
\frac{1}{n}\sum_{i=1}^n d^3_{X_i - x} m(x) K_h(X_i - x)(X_i - x) &= O_p(h^{d-p+4} 1_p),
\end{aligned}$$

$$\frac{1}{n}\sum_{i=1}^{n} d_{X_i-x}^4 m(x)K_h(X_i-x) \quad = \quad O_p(h^{d-p+4})$$

and

$$\frac{1}{n}\sum_{i=1}^{n} d_{X_i-x}^4 m(x)K_h(X_i-x)(X_i-x) = O_p(h^{d-p+5}1_p) + o_p(h^{d-p+4}1_p) = o_p(h^{d-p+4}1_p)$$

by the assumption 6 in Section 3.1. Thus, we have

$$\begin{aligned}
R_j(x) &= e_1^T\left(\frac{1}{n}X_x^T W_x X_x\right)^{-1}\frac{1}{n}X_x^T W_x \begin{bmatrix} d_{X_1-x}^j m(x) \\ \vdots \\ d_{X_n-x}^j m(x)\end{bmatrix} \\
&= e_1^T\left(\frac{1}{n}X_x^T W_x X_x\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} d_{X_i-x}^j m(x)K_h(X_i-x) \\ d_{X_i-x}^j m(x)K_h(X_i-x)(X_i-x)\end{bmatrix} \\
&= e_1^T O_p(h^{p-d}1_{p+1}1_{p+1}^T)\begin{bmatrix} O_p(h^{d-p+j}) \\ O_p(h^{d-p+4}1_p)\end{bmatrix} \\
&= O_p(h^j),
\end{aligned}$$

so we obtain (23). □

Lemma 3. Let $x \in \mathcal{X}$. Under the assumptions in Section 3.1, as $n \to \infty$ and $h \to 0$, we obtain

$$\begin{aligned}
&e_1^T(X_x^T W_x X_x)^{-1}X_x^T W_x \\
&= \frac{h^{p-d}}{n}A_1(x)^{-1}\begin{bmatrix}(1-A_2(x)^T A_3(x)^{-1}(X_1-x))K_h(X_1-x) \\ \vdots \\ (1-A_2(x)^T A_3(x)^{-1}(X_n-x))K_h(X_n-x)\end{bmatrix}^T (1+o_p(1)).
\end{aligned}$$

Proof. Using Lemma 1, we calculate

$$\begin{aligned}
&e_1(X_x^T W_x X_x)^{-1}X_x^T W_x \\
&\qquad = e_1^T\left(\frac{1}{n}X_x^T W_x X_x\right)^{-1}\frac{1}{n}X_x^T W_x \\
&\qquad = h^{p-d}A_1(x)^{-1}\left\{\begin{bmatrix}1 & -A_2(x)^T A_3(x)^{-1}\end{bmatrix} + o_p(1_{p+1}^T)\right\}
\end{aligned}$$

$$
\times \frac{1}{n} \begin{bmatrix} K_h(X_1 - x) & \cdots & K_h(X_n - x) \\ K_h(X_1 - x)(X_1 - x) & \cdots & K_h(X_n - x)(X_n - x) \end{bmatrix}
$$

$$
= \frac{h^{p-d}}{n} A_1(x)^{-1} \begin{bmatrix} (1 - A_2(x)^T A_3(x)^{-1}(X_1 - x))K_h(X_1 - x) \\ \vdots \\ (1 - A_2(x)^T A_3(x)^{-1}(X_n - x))K_h(X_n - x) \end{bmatrix}^T
$$

$$
\times (1 + o_p(1)). \qquad \square
$$

LEMMA 4. Let $x \in \mathcal{X}$ and $k \in \{1, 2, 3, 4\}$. If $c(x)$ is a p-dimensional vector function whose elements are continuous functions of x, then as $h \to 0$,

$$
E[(c(x)^T(X - x))^k K_h(X - x)^4] = O(h^{-4p+d+k}).
$$

Proof. By direct calculation, we obtain

$$
\begin{aligned}
&E[(c(x)^T(X - x))^k K_h(X - x)^4] \\
&= h^{-4p}\mathrm{E}\left[(c(x)^T(X - x))^k K\left(\frac{X - x}{h}\right)^4 1(X \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X}))\right] \\
&\quad + h^{-4p}\mathrm{E}\left[(c(x)^T(X - x))^k K\left(\frac{X - x}{h}\right)^4 1(X \in (\mathcal{B}^p_{x,h^{1-\eta}} \cap \mathcal{X})^c)\right] \\
&= h^{-4p}\mathrm{E}\left[(c(x)^T(\varphi(Z) - \varphi(0)))^k K\left(\frac{\varphi(Z) - \varphi(0)}{h}\right)^4 1(Z \in \mathcal{B}^d_{0,h^{1-\eta}})\right] + o(h^{d-4p+4}) \\
&= h^{-4p}\int_{\mathcal{B}^d_{0,h^{1-\eta}}} (c(x)^T(\varphi(z) - \varphi(0)))^k K\left(\frac{\varphi(z) - \varphi(0)}{h}\right)^4 f(z)dz + o(h^{d-4p+4}) \\
&= h^{-4p}\int_{\mathcal{B}^d_{0,h^{-\eta}}} (c(x)^T(\varphi(hu) - \varphi(0)))^k K\left(\frac{\varphi(hu) - \varphi(0)}{h}\right)^4 f(hu)h^d du + o(h^{d-4p+4}) \\
&= h^{d-4p+k}\int_{\mathcal{B}^d_{0,h^{-\eta}}} \left(c(x)^T\frac{\varphi(hu) - \varphi(0)}{h}\right)^k K\left(\frac{\varphi(hu) - \varphi(0)}{h}\right)^4 f(hu)du + o(h^{d-4p+4}) \\
&= h^{d-4p+k}f(0)\int_{\mathbb{R}^d} \left(c(x)^T\mathcal{J}(0|\varphi)u\right)^k K(\mathcal{J}(0|\varphi)u)^4 du + o(h^{d-4p+k}) + o(h^{d-4p+4}). \qquad \square
\end{aligned}
$$

LEMMA 5. Let $x \in \mathcal{X}$.

1. If function $w^*: \mathcal{X} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous on $\overline{\mathcal{B}^p_{x,\mu}} \times \overline{\mathcal{B}^p_{x,\mu}} \times \overline{\mathcal{B}^p_{x,\mu}}$, then as $h \to 0$,

$$E\left[w^*(X_1, X_2, X_3)K\left(\frac{X_3 - X_2}{h}\right)K\left(\frac{X_3 - X_1}{h}\right)K\left(\frac{X_2 - x}{h}\right)K\left(\frac{X_1 - x}{h}\right)(X_1 - x)\right]$$
$$= O(h^{3d+1}1_p), \tag{24}$$
$$E\left[w^*(X_1, X_2, X_3)K\left(\frac{X_3 - X_2}{h}\right)K\left(\frac{X_3 - X_1}{h}\right)K\left(\frac{X_2 - x}{h}\right)K\left(\frac{X_1 - x}{h}\right)(X_1 - x)(X_2 - x)^T\right]$$
$$= O(h^{3d+2}1_p1_p^T). \tag{25}$$

2. If function $w': \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous on $\overline{\mathcal{B}^p_{x,\mu}} \times \overline{\mathcal{B}^p_{x,\mu}}$, then as $h \to 0$, for $i, j \in \{1, 2\}$,

$$E\left[w'(X_1, X_2)K\left(\frac{X_1 - X_2}{h}\right)K\left(\frac{X_i - x}{h}\right)K\left(\frac{X_j - x}{h}\right)(X_1 - x)\right] = O(h^{2d+1}1_p), \tag{26}$$
$$E\left[w'(X_1, X_2)K\left(\frac{X_1 - X_2}{h}\right)K\left(\frac{X_i - x}{h}\right)K\left(\frac{X_j - x}{h}\right)(X_1 - x)(X_2 - x)^T\right]$$
$$= O(h^{2d+2}1_p1_p^T). \tag{27}$$

3. If function $w: \mathcal{X} \to \mathbb{R}$ is continuous on $\overline{\mathcal{B}^p_{x,\mu}}$, then as $h \to 0$,

$$E\left[w(X_1)K\left(\frac{X_1 - x}{h}\right)^2(X_1 - x)\right] = O(h^{d+1}1_p), \tag{28}$$
$$E\left[w(X_1)K\left(\frac{X_1 - x}{h}\right)^2(X_1 - x)(X_1 - x)^T\right] = O(h^{d+2}1_p1_p^T). \tag{29}$$

Proof. We introduce the following calculations before starting the proof of Lemma 5. By applying Taylor's theorem for each $\varphi_\alpha(hu)$ with variate h, we get

$$\varphi_\alpha(hu) = \varphi_\alpha(0) + h\nabla\varphi_\alpha(0)^T u + \frac{h^2}{2}u^T\nabla^2\varphi_\alpha(0)u + \frac{h^3}{6}\sum_{i,j,k=1}^{d} u_iu_ju_k \left.\frac{\partial^3}{\partial z_i \partial z_j \partial z_k}\varphi_\alpha(z)\right|_{z=h_\alpha u}$$

for some $0 < h_\alpha < h$. We put

$$\xi_\alpha(u, h) = \frac{1}{2}u^T\nabla^2\varphi_\alpha(0)u + \frac{h}{6}\sum_{i,j,k=1}^{d} u_iu_ju_k \left.\frac{\partial^3}{\partial z_i \partial z_j \partial z_k}\varphi_\alpha(z)\right|_{z=h_\alpha u}.$$

By collecting $\varphi_1(hu) \cdots \varphi_p(hu)$, we obtain

$$\frac{\varphi(hu) - \varphi(0)}{h} = \mathcal{J}(0|\varphi)u + h\begin{bmatrix}\xi_1(u, h)\\ \vdots \\ \xi_p(u, h)\end{bmatrix}.$$

Using above expansion and multivariate Taylor's theorem, we have

$$K\left(\frac{\varphi(hu)-\varphi(0)}{h}\right) = K\left(\mathcal{J}(0|\varphi)u + h\begin{bmatrix}\xi_1(u,h)\\ \vdots \\ \xi_p(u,h)\end{bmatrix}\right) = K(\mathcal{J}(0|\varphi)u) + \kappa(0,u,h), \quad (30)$$

where

$$\kappa(0,u,h) \;=\; h\sum_{\alpha=1}^{p}\xi_\alpha(u,h)\partial_\alpha K(\mathcal{J}(0|\varphi)u) + \frac{h^2}{2}\sum_{\alpha,\beta=1}^{p}\xi_\alpha(u,h)\xi_\beta(u,h)\partial_{\alpha\beta}K(u_0)$$

and

$$u_0 = \mathcal{J}(0|\varphi)u + s_0 h\begin{bmatrix}\xi_1(u,h)\\ \vdots \\ \xi_p(u,h)\end{bmatrix}$$

for some $0 < s_0 < 1$. Additionally we have

$$\frac{\varphi(hu)-\varphi(hv)}{h} = \mathcal{J}(0|\varphi)(u-v) + h\begin{bmatrix}\xi_1(u,h)-\xi_1(v,h)\\ \vdots \\ \xi_p(u,h)-\xi_p(v,h)\end{bmatrix}.$$

So we get

$$K\left(\frac{\varphi(hu)-\varphi(hv)}{h}\right) \;=\; K(\mathcal{J}(0|\varphi)(u-v)) + \lambda(0,u,v,h), \quad (31)$$

where

$$\begin{aligned}\lambda(0,u,v,h) \;=\; & h\sum_{\alpha=1}^{p}\left\{\xi_\alpha(u,h)-\xi_\alpha(v,h)\right\}\partial_\alpha K(\mathcal{J}(0|\varphi)(u-v)) \\ & +\frac{h^2}{2}\sum_{\alpha,\beta=1}^{p}\left\{\xi_\alpha(u,h)-\xi_\alpha(v,h)\right\}\left\{\xi_\alpha(u,h)-\xi_\alpha(v,h)\right\}\partial_{\alpha\beta}K(w_0)\end{aligned}$$

and

$$w_0 = \mathcal{J}(0|\varphi)(u-v) + t_0 h\begin{bmatrix}\xi_1(u,h)-\xi_1(v,h)\\ \vdots \\ \xi_p(u,h)-\xi_p(v,h)\end{bmatrix}$$

for some $0 < t_0 < 1$.

We shall start the proof of Lemma 5. By direct calculation,

$$E\left[w^*(X_1,X_2,X_3)K\left(\frac{X_3-X_2}{h}\right)K\left(\frac{X_3-X_1}{h}\right)K\left(\frac{X_2-x}{h}\right)K\left(\frac{X_1-x}{h}\right)(X_1-x)\right]$$

$$= E\Bigg[w^*(X_1,X_2,X_3)K\left(\frac{X_3-X_2}{h}\right)K\left(\frac{X_3-X_1}{h}\right)K\left(\frac{X_2-x}{h}\right)K\left(\frac{X_1-x}{h}\right)$$

$$\times 1(X_1\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})1(X_2\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})1(X_3\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})(X_1-x)\Bigg]$$

$$+E\Bigg[w^*(X_1,X_2,X_3)K\left(\frac{X_3-X_2}{h}\right)K\left(\frac{X_3-X_1}{h}\right)K\left(\frac{X_2-x}{h}\right)K\left(\frac{X_1-x}{h}\right)$$

$$\times 1(X_1\in(\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})^c)1(X_2\in(\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})^c)1(X_3\in(\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})^c)(X_1-x)\Bigg].$$

Above second term is

$$E\Bigg[w^*(X_1,X_2,X_3)K\left(\frac{X_3-X_2}{h}\right)K\left(\frac{X_3-X_1}{h}\right)K\left(\frac{X_2-x}{h}\right)K\left(\frac{X_1-x}{h}\right)(X_1-x)$$

$$\times 1(X_1\in(\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})^c)1(X_2\in(\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})^c)1(X_3\in(\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})^c)\Bigg] = o(h^{3d+1}1_p)$$

because there exists C such that

$$|w^*(y,z,w)(y_\alpha-x_\alpha)|\le C(1+||y||^2)$$

for all $y,z,w\in\mathcal{X}$. By using (30) and (31), the first term can be calculated as

$$E\Bigg[w^*(\varphi(Z_1),\varphi(Z_2),\varphi(Z_3))K\left(\frac{\varphi(Z_3)-\varphi(Z_2)}{h}\right)K\left(\frac{\varphi(Z_3)-\varphi(Z_1)}{h}\right)$$

$$\times K\left(\frac{\varphi(Z_2)-\varphi(0)}{h}\right)K\left(\frac{\varphi(Z_1)-\varphi(0)}{h}\right)(\varphi_\beta(Z_1)-\varphi_\beta(0))$$

$$\times 1(\varphi(Z_1)\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})1(\varphi(Z_2)\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})1(\varphi(Z_3)\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})\Bigg]$$

$$=\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} w^*(\varphi(z_1),\varphi(z_2),\varphi(z_3))K\left(\frac{\varphi(z_3)-\varphi(z_2)}{h}\right)K\left(\frac{\varphi(z_3)-\varphi(z_1)}{h}\right)$$

$$\times K\left(\frac{\varphi(z_2)-\varphi(0)}{h}\right)K\left(\frac{\varphi(z_1)-\varphi(0)}{h}\right)(\varphi_\beta(z_1)-\varphi_\beta(0))f(z_1)f(z_2)f(z_3)$$

$$\times 1(\varphi(z_1) \in \mathcal{B}^p_{x,h^{1-\eta^*}} \cap \mathcal{X}) 1(\varphi(z_2) \in \mathcal{B}^p_{x,h^{1-\eta^*}} \cap \mathcal{X}) 1(\varphi(z_3) \in \mathcal{B}^p_{x,h^{1-\eta^*}} \cap \mathcal{X}) dz_1 dz_2 dz_3$$

$$= \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} w^*(\varphi(z_1), \varphi(z_2), \varphi(z_3)) K\left(\frac{\varphi(z_3) - \varphi(z_2)}{h}\right) K\left(\frac{\varphi(z_3) - \varphi(z_1)}{h}\right)$$

$$\times K\left(\frac{\varphi(z_2) - \varphi(0)}{h}\right) K\left(\frac{\varphi(z_1) - \varphi(0)}{h}\right) f(z_1) f(z_2) f(z_3) (\varphi_\beta(z_1) - \varphi_\beta(0))$$

$$\times 1(z_1 \in \mathcal{B}^d_{0,h^{1-\eta^*}}) 1(z_2 \in \mathcal{B}^d_{0,h^{1-\eta^*}}) 1(z_2 \in \mathcal{B}^d_{0,h^{1-\eta^*}}) dz_1 dz_2 dz_3$$

$$= \int_{\mathcal{B}^d_{0,h^{1-\eta^*}} \times \mathcal{B}^d_{0,h^{1-\eta^*}} \times \mathcal{B}^d_{0,h^{1-\eta^*}}} w^*(\varphi(z_1), \varphi(z_2), \varphi(z_3)) K\left(\frac{\varphi(z_3) - \varphi(z_2)}{h}\right)$$

$$\times K\left(\frac{\varphi(z_3) - \varphi(z_1)}{h}\right) K\left(\frac{\varphi(z_2) - \varphi(0)}{h}\right) K\left(\frac{\varphi(z_1) - \varphi(0)}{h}\right)$$

$$\times f(z_1) f(z_2) f(z_3) (\varphi_\beta(z_1) - \varphi_\beta(0)) dz_1 dz_2 dz_3$$

$$= h^{3d} \int_{\mathcal{B}^d_{0,h^{-\eta^*}} \times \mathcal{B}^d_{0,h^{-\eta^*}} \times \mathcal{B}^d_{0,h^{-\eta^*}}} w^*(\varphi(hu_1), \varphi(hu_2), \varphi(hu_3)) K\left(\frac{\varphi(hu_3) - \varphi(hu_2)}{h}\right)$$

$$\times K\left(\frac{\varphi(hu_3) - \varphi(hu_1)}{h}\right) K\left(\frac{\varphi(hu_2) - \varphi(0)}{h}\right) K\left(\frac{\varphi(hu_1) - \varphi(0)}{h}\right)$$

$$\times f(hu_1) f(hu_2) f(hu_3) (\varphi_\beta(hu_1) - \varphi_\beta(0)) du_1 du_2 du_3$$

$$= h^{3d} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} w^*(\varphi(hu_1), \varphi(hu_2), \varphi(hu_3)) f(hu_1) f(hu_2) f(hu_3)$$

$$\times K\left(\frac{\varphi(hu_3) - \varphi(hu_2)}{h}\right) K\left(\frac{\varphi(hu_3) - \varphi(hu_1)}{h}\right) K\left(\frac{\varphi(hu_2) - \varphi(0)}{h}\right) K\left(\frac{\varphi(hu_1) - \varphi(0)}{h}\right)$$

$$\times 1(u_1 \in \mathcal{B}^d_{0,h^{-\eta^*}}) 1(u_2 \in \mathcal{B}^d_{0,h^{-\eta^*}}) 1(u_3 \in \mathcal{B}^d_{0,h^{-\eta^*}}) (\varphi_\beta(hu_1) - \varphi_\beta(0)) du_1 du_2 du_3$$

$$= h^{3d} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} w^*(\varphi(hu_1), \varphi(hu_2), \varphi(hu_3))$$

$$\times \left(f(0) + hu_1^T \nabla f(0) + \frac{h^2}{2} u_1^T \nabla^2 f(h_0 u_1) u_1\right)$$

$$\times \left(f(0) + hu_2^T \nabla f(0) + \frac{h^2}{2} u_2^T \nabla^2 f(h_0 u_2) u_2\right)$$

$$\times \left(f(0) + hu_3^T \nabla f(0) + \frac{h^2}{2} u_3^T \nabla^2 f(h_0 u_3) u_3\right)$$

$$\times \left(K\left(\mathcal{J}(0|\varphi)(u_3 - u_2)\right) + \lambda(0, u_3.u_2, h)\right)$$

$$\times \left(K\left(\mathcal{J}(0|\varphi)(u_3-u_1)\right)+\lambda(0,u_3.u_1,h)\right)$$

$$\times \left\{K\left(\mathcal{J}(0|\varphi)u_2\right)+\kappa(0,u_2,h)\right\}$$

$$\times \left\{K\left(\mathcal{J}(0|\varphi)u_1\right)+\kappa(0,u_1,h)\right\}$$

$$\times 1(u_1 \in \mathcal{B}^d_{0,h^{-\eta^*}})1(u_2 \in \mathcal{B}^d_{0,h^{-\eta^*}})1(u_3 \in \mathcal{B}^d_{0,h^{-\eta^*}})(\varphi_\beta(hu_1)-\varphi_\beta(0))du_1du_2du_3,$$

of which the leading term is

$$\begin{aligned}
&h^{3d}f(0)^3\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} K\left(\mathcal{J}(0|\varphi)(u_3-u_2)\right)K\left(\mathcal{J}(0|\varphi)(u_3-u_1)\right)K\left(\mathcal{J}(0|\varphi)u_2\right)\\
&\qquad \times K\left(\mathcal{J}(0|\varphi)u_1\right)1(u_1 \in \mathcal{B}^d_{0,h^{-\eta^*}})1(u_2 \in \mathcal{B}^d_{0,h^{-\eta^*}})1(u_3 \in \mathcal{B}^d_{0,h^{-\eta^*}})\\
&\qquad \times(\varphi_\beta(hu_1)-\varphi_\beta(0))du_1du_2du_3\\
&= h^{3d+1}f(0)^3\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} K\left(\mathcal{J}(0|\varphi)(u_3-u_2)\right)K\left(\mathcal{J}(0|\varphi)(u_3-u_1)\right)K\left(\mathcal{J}(0|\varphi)u_2\right)\\
&\qquad \times K\left(\mathcal{J}(0|\varphi)u_1\right)1(u_1 \in \mathcal{B}^d_{0,h^{-\eta^*}})1(u_2 \in \mathcal{B}^d_{0,h^{-\eta^*}})1(u_3 \in \mathcal{B}^d_{0,h^{-\eta^*}})\\
&\qquad \times\frac{\varphi_\beta(hu_1)-\varphi_\beta(0)}{h}du_1du_2du_3\\
&= h^{3d+1}f(0)^3\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} K(\mathcal{J}(0|\varphi)(u_3-u_2))K(\mathcal{J}(0|\varphi)(u_3-u_1))\\
&\qquad \times K(\mathcal{J}(0|\varphi)u_2)K(\mathcal{J}(0|\varphi)u_1)\nabla\varphi_\beta(0)u_1du_1du_2du_3(1+o(1)).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&E\Bigg[w^*(\varphi(Z_1),\varphi(Z_2),\varphi(Z_3))K\left(\frac{\varphi(Z_3)-\varphi(Z_2)}{h}\right)K\left(\frac{\varphi(Z_3)-\varphi(Z_1)}{h}\right)\\
&\quad\times K\left(\frac{\varphi(Z_2)-\varphi(0)}{h}\right)K\left(\frac{\varphi(Z_1)-\varphi(0)}{h}\right)(\varphi_\beta(Z_1)-\varphi_\beta(0))\\
&\quad\times 1(\varphi(Z_1)\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})1(\varphi(Z_2)\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})1(\varphi(Z_3)\in\mathcal{B}^p_{x,h^{1-\eta^*}}\cap\mathcal{X})\Bigg]\\
&= O(h^{3d+1}1_p),
\end{aligned}$$

which is in (24). The proof of (25) is same as (24).

Next we prove (26). We have

$$E\left[w'(X_1,X_2)K\left(\frac{X_1-X_2}{h}\right)K\left(\frac{X_2-x}{h}\right)K\left(\frac{X_1-x}{h}\right)(X_1-x)\right]$$

$$= E\left[w'(X_1,X_2)K\left(\frac{X_1-X_2}{h}\right)K\left(\frac{X_2-x}{h}\right)K\left(\frac{X_1-x}{h}\right)\right.$$

$$\left.\times 1(X_1\in\mathcal{B}^p_{x,h^{1-\eta'}}\cap\mathcal{X})1(X_2\in\mathcal{B}^p_{x,h^{1-\eta'}}\cap\mathcal{X})(X_1-x)\right]$$

$$+E\left[w'(X_1,X_2)K\left(\frac{X_1-X_2}{h}\right)K\left(\frac{X_2-x}{h}\right)K\left(\frac{X_1-x}{h}\right)\right.$$

$$\left.\times 1(X_1\in(\mathcal{B}^p_{x,h^{1-\eta'}}\cap\mathcal{X})^c)1(X_2\in(\mathcal{B}^p_{x,h^{1-\eta'}}\cap\mathcal{X})^c)(X_1-x)\right]$$

$$= E\left[w'(\varphi(Z_1),\varphi(Z_2))K\left(\frac{\varphi(Z_1)-\varphi(Z_2)}{h}\right)K\left(\frac{\varphi(Z_3)-\varphi(Z_2)}{h}\right)K\left(\frac{\varphi(Z_2)-\varphi(0)}{h}\right)\right.$$

$$\left.\times K\left(\frac{\varphi(Z_1)-\varphi(0)}{h}\right)1(Z_1\in\mathcal{B}^d_{0,h^{1-\eta'}})1(Z_2\in\mathcal{B}^d_{0,h^{1-\eta'}})(\varphi(Z_1)-\varphi(0))\right]$$

$$+o(h^{2d+1}1_p).$$

The above leading term is

$$E\left[w'(\varphi(Z_1),\varphi(Z_2))K\left(\frac{\varphi(Z_1)-\varphi(Z_2)}{h}\right)K\left(\frac{\varphi(Z_2)-\varphi(0)}{h}\right)K\left(\frac{\varphi(Z_1)-\varphi(0)}{h}\right)\right.$$

$$\left.\times 1(Z_1\in\mathcal{B}^d_{0,h^{1-\eta'}})1(Z_2\in\mathcal{B}^d_{0,h^{1-\eta'}})(\varphi_\beta(Z_1)-\varphi_\beta(0))\right]$$

$$=\int_{\mathbb{R}^d\times\mathbb{R}^d} w'(\varphi(z_1),\varphi(z_2))K\left(\frac{\varphi(z_1)-\varphi(z_2)}{h}\right)K\left(\frac{\varphi(z_2)-\varphi(0)}{h}\right)K\left(\frac{\varphi(z_1)-\varphi(0)}{h}\right)$$

$$\times 1(\varphi(z_1)\in\mathcal{B}^p_{x,h^{1-\eta'}}\cap\mathcal{X})1(\varphi(z_2)\in\mathcal{B}^p_{x,h^{1-\eta'}}\cap\mathcal{X})f(z_1)f(z_2)(\varphi_\beta(z_1)-\varphi_\beta(0))dz_1dz_2$$

$$=\int_{\mathcal{B}^d_{0,h^{1-\eta'}}\times\mathcal{B}^d_{0,h^{1-\eta'}}} w'(\varphi(z_1),\varphi(z_2))f(z_1)f(z_2)K\left(\frac{\varphi(z_1)-\varphi(z_2)}{h}\right)$$

$$\times K\left(\frac{\varphi(z_2)-\varphi(0)}{h}\right)K\left(\frac{\varphi(z_1)-\varphi(0)}{h}\right)(\varphi_\beta(z_1)-\varphi_\beta(0))dz_1dz_2$$

$$=h^{2d+1}\int_{\mathcal{B}^d_{0,h^{-\eta'}}\times\mathcal{B}^d_{0,h^{-\eta'}}} w'(\varphi(hu_1),\varphi(hu_2))f(hu_1)f(hu_2)K\left(\frac{\varphi(hu_1)-\varphi(hu_2)}{h}\right)$$

$$\times K\left(\frac{\varphi(hu_2)-\varphi(0)}{h}\right)K\left(\frac{\varphi(hu_1)-\varphi(0)}{h}\right)\frac{\varphi_\beta(hu_1)-\varphi_\beta(0)}{h}du_1du_2$$

$$= h^{2d+1}\int_{\mathbb{R}^d\times\mathbb{R}^d}\{w'(\varphi(0),\varphi(0))+(w'(\varphi(hu_1),\varphi(hu_2))-w'(\varphi(0),\varphi(0)))\}$$

$$\times\left(f(0)+hu_1^T\nabla f(0)+\frac{h^2}{2}u_1^T\nabla^2 f(h_0u_1)u_1\right)$$

$$\times\left(f(0)+hu_2^T\nabla f(0)+\frac{h^2}{2}u_2^T\nabla^2 f(h_0u_2)u_2\right)$$

$$\times\left(K\left(\mathcal{J}(0|\varphi)(u_1-u_2)\right)+\lambda(0,u_1,u_2,h)\right)$$

$$\times\left(K\left(\mathcal{J}(0|\varphi)u_2\right)+\kappa(0,u_2,h)\right)\left(K\left(\mathcal{J}(0|\varphi)u_1\right)+\kappa(0,u_1,h)\right)$$

$$\times 1(u_1\in\mathcal{B}^d_{0,h^{-\eta'}})1(u_2\in\mathcal{B}^d_{0,h^{-\eta'}})\frac{\varphi_\beta(hu_1)-\varphi_\beta(0)}{h}du_1du_2du_3$$

$$= h^{2d+1}w'(\varphi(0),\varphi(0))f(0)^2$$

$$\times\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}K(\mathcal{J}(0|\varphi)(u_1-u_2))K(\mathcal{J}(0|\varphi)u_2)du_2K(\mathcal{J}(0|\varphi)u_1)\nabla\varphi_\beta(0)u_1du_1(1+o(1)).$$

Thus we have

$$E\left[w'(X_1,X_2)K\left(\frac{X_1-X_2}{h}\right)K\left(\frac{X_2-x}{h}\right)K\left(\frac{X_1-x}{h}\right)(X_1-x)\right]=O(h^{2d+1}1_p),$$

which completes to derive (26).

Finally, (28) can be derived as

$$E\left[w(X_1)K\left(\frac{X_1-x}{h}\right)^2(X_1-x)\right]$$

$$= h^{d+1}w(0)f(0)\int_{\mathbb{R}^d}K\left(\mathcal{J}(0|\varphi)u\right)^2\mathcal{J}(0|\varphi)udu(1+o(1)).$$

The proof of (29) is same as (28). □

Lemma 6. Let $x\in\mathcal{X}$. Under the assumptions in Section 3.1, as $n\to\infty$ and $h\to 0$, we obtain

$$e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xLVL^TW_xX_x(X_x^TW_xX_x)^{-1}e_1$$

$$=\frac{1}{nh^d}v(x)A_1(x)^{-4}f(\varphi^{-1}(x))^3\int_{\mathbb{R}^d}K*K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)^2dv(1+o_p(1)),$$

$$e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xVL^TW_xX_x(X_x^TW_xX_x)^{-1}e_1$$

$$= \frac{1}{nh^d}v(x)A_1(x)^{-1}f(\varphi^{-1}(x))^2\int_{\mathbb{R}^d}K\left(\mathcal{J}(\varphi^{-1}(x)|\varphi)v\right)K*K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)dv$$

$$\times(1+o_p(1)).$$

Proof. Lemma 1 implies that $e_1^T(1/nX_x^TW_xX_x)^{-1} = O_p(h^{p-d}1_{p+1}^T)$ and we note that

$$LVL^T = [L_1\cdots L_n]^TV[L_1\cdots L_n] = \begin{bmatrix} L_1^T \\ \vdots \\ L_n^T \end{bmatrix} V[L_1\cdots L_n] = \begin{bmatrix} L_1^TVL_1 & \cdots & L_1^TVL_n \\ \vdots & \ddots & \vdots \\ L_n^TVL_1 & \cdots & L_n^TVL_n \end{bmatrix}.$$

We have by using blockwise calculation that

$$\frac{1}{n^2}X_x^TW_xLVL^TW_xX_x \equiv \begin{bmatrix} \mathscr{M}_{11} & \mathscr{M}_{21}^T \\ \mathscr{M}_{21} & \mathscr{M}_{22} \end{bmatrix},$$

where

$$\begin{aligned}
\mathscr{M}_{11} &= \frac{1}{n^2}\sum_{i,j=1}^{n} L_i^TVL_jK_h(X_i-x)K_h(X_j-x) \\
&= \frac{h^{d-2p}}{n}v(x)A_1(x)^{-2}f(\varphi^{-1}(x))^3\int_{\mathbb{R}^d}K*K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)^2dv(1+o(1)), \quad (32) \\
\mathscr{M}_{21} &= \frac{1}{n^2}\sum_{i,j=1}^{n} L_i^TVL_jK_h(X_i-x)K_h(X_j-x)(X_i-x) \\
&= O_p\left(\frac{h^{d-2p+1}}{n}1_p\right), \quad (33) \\
\mathscr{M}_{22} &= \frac{1}{n^2}\sum_{i,j=1}^{n} L_i^TVL_jK_h(X_i-x)K_h(X_j-x)(X_i-x)(X_j-x)^T \\
&= O_p\left(\frac{h^{d-2p+2}}{n}1_p1_p^T\right). \quad (34)
\end{aligned}$$

Using (32), (33) and (34), we obtain

$$e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xLVL^TW_xX_x(X_x^TW_xX_x)^{-1}e_1$$

$$= e_1^T\left(\frac{1}{n}X_x^TW_xX_x\right)^{-1}\frac{1}{n^2}X_x^TW_xLVL^TW_xX_x\left(\frac{1}{n}X_x^TW_xX_x\right)^{-1}e_1$$

$$= e_1^T \left(\frac{1}{n} X_x^T W_x X_x\right)^{-1} \begin{bmatrix} \mathscr{M}_{11} & \mathscr{M}_{21}^T \\ \mathscr{M}_{21} & \mathscr{M}_{22} \end{bmatrix} \left(\frac{1}{n} X_x^T W_x X_x\right)^{-1} e_1$$

$$= h^{p-d} A_1(x)^{-1} \left(\begin{bmatrix} 1 & -A_2(x)^T A_3(x)^{-1} \end{bmatrix} + o_p(1_{p+1}^T)\right)$$

$$\times \begin{bmatrix} \mathscr{M}_{11} & \mathscr{M}_{21}^T \\ \mathscr{M}_{21} & \mathscr{M}_{22} \end{bmatrix} h^{p-d} A_1(x)^{-1} \left(\begin{bmatrix} 1 \\ -A_3(x)^{-1} A_2(x) \end{bmatrix} + o_p(1_{p+1})\right)$$

$$= h^{2p-2d} A_1(x)^{-2} \mathscr{M}_{11}(1 + o_p(1)) - 2h^{2p-2d} A_1(x)^{-2} \mathscr{M}_{21} A_3(x)^{-1} A_2(x)(1 + o_p(1))$$

$$+ h^{2p-2d} A_1(x)^{-2} A_2(x)^T A_3(x)^{-1} \mathscr{M}_{22} A_3(x)^{-1} A_2(x)(1 + o_p(1))$$

$$= \frac{1}{nh^d} v(x) A_1(x)^{-4} f(\varphi^{-1}(x))^3 \int_{\mathbb{R}^d} K * K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)^2 dv(1 + o_p(1)).$$

Next we shall start the proof of (17). Notice that e_i is a n-dimensional and e_1 is a $(p+1)$-dimensional.

$$I_n V L^T = [\mathrm{e}_1 \cdots \mathrm{e_n}]^T V[L_1 \cdots L_n] = \begin{bmatrix} \mathrm{e}_1^T \\ \vdots \\ \mathrm{e}_n^T \end{bmatrix} V[L_1 \cdots L_n] = \begin{bmatrix} \mathrm{e}_1^T V L_1 & \cdots & \mathrm{e}_1^T V L_n \\ \vdots & \ddots & \vdots \\ \mathrm{e}_n^T V L_1 & \cdots & \mathrm{e}_n^T V L_n \end{bmatrix}.$$

We have by using blockwise calculation that

$$\frac{1}{n^2} X_x^T W_x V L^T W_x X_x = \frac{1}{n^2} X_x^T W_x I_n V L^T W_x X_x \equiv \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{21}^T \\ \mathcal{S}_{21} & \mathcal{S}_{22} \end{bmatrix},$$

where

$$\mathcal{S}_{11} = \frac{1}{n^2} \sum_{i,j=1}^n \mathrm{e}_i^T V L_j K_h(X_i - x) K_h(X_j - x)$$

$$= \frac{h^{d-2p}}{n} v(x) A_1(x)^{-1} f(\varphi^{-1}(x))^2$$

$$\times \int_{\mathbb{R}^d} K\left(\mathcal{J}(\varphi^{-1}(x)|\varphi)v\right) K * K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v) dv(1 + o_p(1)), \tag{35}$$

$$\mathcal{S}_{21} = \frac{1}{n^2} \sum_{i,j=1}^n \mathrm{e}_i^T V L_j K_h(X_i - x) K_h(X_j - x)(X_i - x)$$

$$= O_p\left(\frac{h^{d-2p+1}}{n} 1_p\right), \tag{36}$$

$$\mathcal{S}_{22} = \frac{1}{n^2}\sum_{i,j=1}^{n} \mathrm{e}_i^T V L_j K_h(X_i - x) K_h(X_j - x)(X_i - x)(X_j - x)^T$$
$$= O_p\left(\frac{h^{d-2p+2}}{n} 1_p 1_p^T\right). \tag{37}$$

Using (35), (36) and (37), we obtain

$$e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x V L^T W_x X_x (X_x^T W_x X_x)^{-1} e_1$$

$$= e_1^T \left(\frac{1}{n} X_x^T W_x X_x\right)^{-1} \frac{1}{n^2} X_x^T W_x I_n V L^T W_x X_x \left(\frac{1}{n} X_x^T W_x X_x\right)^{-1} e_1$$

$$= e_1^T \left(\frac{1}{n} X_x^T W_x X_x\right)^{-1} \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{21}^T \\ \mathcal{S}_{21} & \mathcal{S}_{22} \end{bmatrix} \left(\frac{1}{n} X_x^T W_x X_x\right)^{-1} e_1$$

$$= h^{p-d} A_1(x)^{-1} \left(\begin{bmatrix} 1 & -A_2(x)^T A_3(x)^{-1} \end{bmatrix} + o_p(1_{p+1}^T)\right)$$

$$\times \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{21}^T \\ \mathcal{S}_{21} & \mathcal{S}_{22} \end{bmatrix} h^{p-d} A_1(x)^{-1} \left(\begin{bmatrix} 1 \\ -A_3(x)^{-1} A_2(x) \end{bmatrix} + o_p(1_{p+1})\right)$$

$$= h^{2p-2d} A_1(x)^{-2} \mathcal{S}_{11}(1 + o_p(1)) - 2h^{2p-2d} A_1(x)^{-2} \mathcal{S}_{21} A_3(x)^{-1} A_2(x)(1 + o_p(1))$$

$$+ h^{2p-2d} A_1(x)^{-2} A_2(x)^T A_3(x)^{-1} \mathcal{S}_{22} A_3(x)^{-1} A_2(x)(1 + o_p(1))$$

$$= \frac{1}{nh^d} v(x) A_1(x)^{-1} f(\varphi^{-1}(x))^2 \int_{\mathbb{R}^d} K\left(\mathcal{J}(\varphi^{-1}(x)|\varphi)v\right) K * K(\mathcal{J}(\varphi^{-1}(x)|\varphi)v) dv$$

$$\times (1 + o_p(1)). \qquad \square$$