GRADIENT BASED SMOOTHING PARAMETER SELECTION FOR NONPARAMETRIC REGRESSION ESTIMATION

DANIEL J. HENDERSON, QI LI, CHRISTOPHER F. PARMETER, AND SHUANG YAO

Abstract. Uncovering gradients is of crucial importance across a broad range of economic environments. Here we consider data-driven bandwidth selection based on the gradient of an unknown regression function. The procedure developed here is automatic and does not require initial estimation of unknown functions with pilot bandwidths. We prove that it delivers bandwidths which have the optimal rate of convergence for the gradient. Both simulated and empirical examples showcase the finite sample attraction of this new mechanism.

1. Overview

The success of nonparametric estimation hinges critically on the level of smoothing exerted on the unknown surface. Given this importance, a large literature has developed focusing on appropriate selection of the smoothing parameter(s) for both density and conditional mean settings. However, methods developed for recovering optimal smoothness levels for the density or conditional mean are not necessarily the proper surrogates when interest instead hinges on the derivative of the unknown function. Economic applications which require gradient estimation include estimates of heterogenous individual attitudes toward risk (Chiapporis, Gandhi, Salanié & Salanié 2009) and marginal willingness to pay within a two-stage hedonic regression (Bajari & Kahn 2005, Heckman, Matzkin & Nesheim 2010) to name a few.

The importance of appropriate smoothness selection for derivatives was illustrated by Wahba & Wang (1990) who showed in the smoothing spline setting that the ideal smoothing parameter depends on the derivative. Yet, standard selection methods are designed strictly around the fit of the function. A small strand of literature has developed focusing attention on smoothing parameter selection when interest hinges on the derivative. Within this literature there exist several different..
approaches for construction of the optimal bandwidth. The first approach is the factor method, which adjusts a bandwidth selected for the conditional mean by a known constant (depending upon the kernel) to obtain a bandwidth for estimation of the derivative. The second approach is an analog to traditional data-driven methods (such as cross-validation) which uses empirically constructed noise-corrupted derivatives to replace the oracle inside the criterion function. The third approach is to create plug-in bandwidths.

To develop the intuition for existing approaches consider a $d$-dimensional nonparametric regression model

$$y_j = g(x_j) + u_j = g(x_{j1}, ..., x_{jd}) + u_j, \quad j = 1, ..., n. \quad (1)$$

Rice (1986) proposed a method for selecting a smoothing parameter optimal for construction of the derivative of $g(x)$. Rice's (1986) focus was univariate in nature. He suggested the use of a differencing operator though it is not formally defined and the criterion proposed is a nearly unbiased estimator of the mean integrated squared error (MISE) between the estimated derivative and the oracle. Building on the insight of Rice (1986), Müller, Stadtmüller & Schmitt (1987) used his noise-corrupted suggestion to select the bandwidth based on the natural extension of least-squares cross-validation (LSCV). Müller et al. (1987) also formally proposed a differencing operator for calculating noise-corrupted observations of the gradients. Noting that the differencing operator deployed by Müller et al. (1987) possessed a high variance, Charnigo, Hall & Srinivasan (2011) proposed a differencing operator with more desirable variance properties as well as a generalized criterion to be used for selecting the optimal smoothing parameter.

As an alternative to noise-corrupted observations of the desired gradients, Müller et al. (1987) proposed a simpler approach by adjusting a bandwidth selected for $g(x)$ to account for the fact that the bandwidth for the gradient estimate needs to converge slower. The interesting aspect of the factor method is that, in the univariate setting, the ratio between the asymptotically optimal bandwidths for estimation of $g(x)$ and its derivative depend on the kernel. Using this fact, Müller et al. (1987) recovered an optimal bandwidth for the derivative eschewing difference quotients. Fan & Gijbels (1995) used this insight to first construct a plug-in estimator for the conditional mean and then adjust this bandwidth to have an optimal bandwidth for the derivative of the conditional mean.

Beyond the factor method, Fan & Gijbels (1995) also proposed a two-step bandwidth which consists of using the factor method, plug-in bandwidth to then construct empirical measures of the bias and conditional variance of the local-polynomial estimator. The unknown terms within the bias and variance are replaced with estimates using the factor-method bandwidth. Once these measures are constructed, the final bandwidth, termed the refined bandwidth, is found by minimizing MISE. Fan, Gijbels, Hu & Huang (1996) showed that this bandwidth selection mechanism has desirable properties both theoretically as well as in simulated settings.

In a separate approach, Ruppert (1997) developed empirical-bias bandwidth selection. A key difference from Ruppert's (1997) approach is that instead of fitting a local-polynomial to obtain
estimates for the unknown components in the bias expansion for the gradient, he instead estimates
the gradient for several different bandwidths and then uses least-squares to fit a Taylor expansion to
the estimated unknown components of the bias. A benefit of this approach over the aforementioned
methods is that it requires estimation of fewer components in practice.

Each of the existing methods leaves something to be desired in a multivariate setting. The
factor method requires bandwidth selection on the conditional mean followed by calculation of a
scaling factor dependent upon the kernel function which can be tedious. The calculation of noise-
corrupted derivatives also requires computing the number of neighboring observations to construct
the estimate prior to minimizing the criterion function. In high dimensional settings this may not
be feasible. Lastly, plug-in approaches, while having desirable theoretical properties, require the
calculation of numerous unknown quantities, neutering the ability of having a completely automatic
procedure. All plug-in approaches require estimation of unknown functions and their derivatives
prior to the formal selection of the bandwidth. Moreover, the plug-in formula for the optimal
bandwidths can become quite complicated in high dimensional settings. The framework laid out
here does not require adjustment, calculation of noise-corrupted derivatives or unknown quantities
related to the underlying data generating process. The method also does not hinge on a pilot
bandwidth or set of estimates being supplied to the criterion function, streamlining the process.

We begin with the oracle LSCV setup for the gradient as in Müller et al. (1987), with a local-
linear estimator. We then derive a simplification of this expression where unknowns are replaced
with their local-constant counterparts. This yields an expression which can be minimized and
depends on a single set of bandwidths. This dramatically decreases the computational burden over
similar procedures with multivariate data. Furthermore, the framework here is firmly entrenched
in the criterion based, data-driven arena as opposed to plug-in approaches.

The remainder of the paper is as follows. Section 2 gives a cursory discussion of our cross-
validation procedure through the lens of a simple univariate example. It further provides the formal
details of our new cross-validation procedure and the asymptotic justification for our proposed
method. Section 3 contains a set of small sample simulations to show the performance of our
bandwidth selection method relative to LSCV for estimation of derivative functions. Section 4
provides a traditional, applied econometric application of our bandwidth selection mechanism to
the study of the public/private capital productivity puzzle while Section 5 concludes. Proofs of
the theorems and the requisite lemmata appear in Appendices A and B with Appendix B available
from the authors upon request.

2. THE GRADIENT BASED CROSS-VALIDATION METHOD AND ITS ASYMPTOTIC BEHAVIOR

2.1. CURSORY DISCUSSION OF CROSS-VALIDATION AND AN ILLUSTRATIVE EXAMPLE. In this section
we provide an illustrative example to show the conventional LSCV method supplies bandwidths
which often lead to estimated regression curves that display too much variation in the estimated
gradients. Consider a univariate nonparametric regression model
\[ y_j = g(x_j) + u_j, \quad j = 1, \ldots, n. \]  
(2)

To operationalize a nonparametric kernel estimator for the conditional mean function \( g(\cdot) \), a smoothing parameter must be selected. The LSCV approach aims to minimize the sample analogue of estimated mean square error, i.e., select \( h \) to minimize \( C_1(h) = n^{-1} \sum_{i=1}^{n} [g(x_i) - \hat{g}_h(x_i)]^2 \), where \( \hat{g}_h(x_i) \) is a kernel estimator of \( g(x_i) \) that depends on the smoothing parameter \( h \). In practice, the unknown function \( g(x_i) \) is replaced by \( y_i \), and to avoid over-fitting, \( \hat{g}_h(x_i) \) is replaced by a leave-one-out version of it, say \( \hat{g}_{-i,h}(x_i) \). Then \( C_2(h) = n^{-1} \sum_{i=1}^{n} [y_i - \hat{g}_{-i,h}(x_i)]^2 \) is minimized. It can be shown that \( C_1(h) \) and \( C_2(h) \) have the same leading term, so asymptotically, the LSCV method selects a smoothing parameter that minimizes the asymptotic estimated mean square error. Approaches like this are designed to choose smoothing parameters which yield the best fit without any regard for the gradient(s). Hence, minor changes in the fit can often lead to major changes in the estimated gradient(s). Let \( \beta(x) = \partial g(x) / \partial x \), the derivative function of \( g(x) \). In this paper we want to select \( h \) to minimize \( C_3(h) = n^{-1} \sum_{i=1}^{n} [\beta(x_i) - \hat{\beta}_h(x_i)]^2 \), where \( \hat{\beta}_h(x_i) \) is a kernel estimator of \( \beta(x_i) \). Our problem, which we term gradient based cross-validation (GBCV), poses more difficulty than the conventional LSCV problem as there is no obvious substitute for \( \beta(x_i) \) in \( C_3(h) \). We will show in Section 2.2 how we can resolve this difficulty.

An area of interest for the gradients is the hedonic price literature. For structural policy analysis to be conducted, the gradients of the hedonic price function need to be recovered to assess how they change across markets based on utility. For this simple illustration we use a subset of the hedonic housing price data of Anglin & Gençay (1996), which was studied in a nonparametric context by Parmeter, Henderson & Kumbhakar (2007).\(^1\) Focusing on the relationship between the change in housing price (price is measured in 100,000 Canadian Dollars) and the size of the lot the house resides on (in 10,000 feet squared), we expect to see a monotonically increasing relationship, implying that the estimated gradient should be everywhere positive. Panel (a) of Figure 1 provides the conditional mean estimates and panel (b) gives the derivative plots using the two competing data-driven approaches. The LSCV bandwidth produces a conditional mean estimate which is non-monotonic and an estimated gradient curve which is wiggly throughout the range of the data and repeatedly changes sign. It is difficult to give a meaningful economic interpretation of the many ups and downs of the gradient. In contrast, our GBCV method produces a smooth (monotonic) conditional mean and a (nearly constant) gradient.\(^2\) This estimated relationship is consistent with prior expectations.

While limited in scope, this simple example serves to underscore that while bandwidths obtained via LSCV possess well known theoretical properties and are widely used by applied researchers, it is not uncommon to arrive at bandwidths which turn the focus away from the unknown relationship

\(^1\)We consider houses which have less than 15,000 square foot lots.

\(^2\)The scale of the figure precludes a more detailed understanding of the GBCV estimated gradient, however the estimated gradient using the GBCV bandwidths declines smoothly from 0.72 to 0.64.
of interest and more on explaining the rapid fluctuations produced by the ‘selected’ smoothing parameter.

**Figure 1.** Hedonic housing price data estimated with local-linear least-squares. Each curve is constructed using a second-order Epanechnikov kernel. The solid line uses a smoothing parameter of \( h = 0.0641 \) obtained via LSCV while the dashed line uses a smoothing parameter of \( h = 1.1754 \) obtained via GBCV.

2.2. **The Gradient Based Cross-Validation Function.** In this section we describe our gradient based cross-validation method. Consider a \( d \)-dimensional nonparametric regression model

\[
y_j = g(x_j) + u_j = g(x_{j1}, \ldots, x_{jd}) + u_j, \quad j = 1, \ldots, n. \tag{3}
\]

Denote the \( d \times 1 \) vector of the first derivatives of \( g(x) \) by \( \beta(x) \):

\[
\beta(x) \overset{\text{def}}{=} \frac{\partial g(x)}{\partial x} = \begin{pmatrix}
\frac{\partial g(x)}{\partial x_1} \\
\vdots \\
\frac{\partial g(x)}{\partial x_d}
\end{pmatrix}
\]

and define a \( (d + 1) \times 1 \) vector \( \delta(x) \) by

\[
\delta(x) = \begin{pmatrix}
g(x) \\
\beta(x)
\end{pmatrix},
\]

where the first component of \( \delta(x) \) is \( g(x) \) and the remaining \( d \) components are the first derivatives of \( g(x) \). Taking a Taylor series expansion of \( g(x_j) \) at \( x_i \), we get

\[
g(x_j) = g(x_i) + (x_j - x_i)^T \beta(x_i) + R_{ji}, \tag{4}
\]

where the superscript \( T \) denotes the transpose of a matrix, and \( R_{ji} = g(x_j) - g(x_i) - (x_j - x_i)^T \beta(x_i) \). Using (4) we can re-write (3) as

\[
y_j = g(x_i) + (x_j - x_i)^T \beta(x_i) + R_{ji} + u_j = (1, (x_j - x_i)^T) \delta(x_i) + R_{ji} + u_j.
\]
The local-linear estimator of \( \delta(x) = (g(x), \beta(x)^T)^T \) is obtained by choosing \((a, b)^T \in \mathbb{R}^{d+1}\) to minimize the following objective function

\[
\min_{a,b} \sum_{j=1}^{n} [y_j - a - (x_j - x)^T b]^2 W_{h,jx},
\]

where \( W_{h,jx} = \prod_{s=1}^{d} h^{s-1}_s w\left( \frac{x_j - x_s}{h_s} \right) \) is a product kernel function, \( w(\cdot) \) is a univariate kernel function, \( x_j \) and \( x_s \) are the \( s^{th} \) components of \( x_j \) and \( x \), respectively, and \( h_s \) is the smoothing parameter associated with the \( s^{th} \) component of \( x \), \( s = 1, \ldots, d \).

The first-order condition (normal equations) to the minimization problem (5) is:

\[
\sum_{j=1}^{n} [y_j - a - (x_j - x)^T b] \begin{pmatrix} 1 \\ (x_j - x) \end{pmatrix} W_{h,jx} = 0,
\]

which leads to the closed form solution of \( \hat{\delta}(x) = (\hat{a}, \hat{b}^T)^T \equiv (\hat{g}(x), \hat{\beta}(x)^T)^T \) given by

\[
\hat{\delta}(x) = \begin{pmatrix} \hat{g}(x) \\ \hat{\beta}(x) \end{pmatrix} = \left[ \sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ (x_j - x) \end{pmatrix} \right]^{-1} \left[ \sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ (x_j - x) \end{pmatrix} y_j \right].
\]

A leave-one-out local-linear kernel estimator of \( \delta(x_i) \) is obtained by replacing \( x \) with \( x_i \) and replacing \( \sum_{j=1}^{n} \) by \( \sum_{j \neq i}^{n} \).

\[
\hat{\delta}_{-i}(x_i) = \begin{pmatrix} \hat{g}_{-i}(x_i) \\ \hat{\beta}_{-i}(x_i) \end{pmatrix} = \left[ \sum_{j \neq i}^{n} W_{h,ji} \begin{pmatrix} 1 \\ (x_j - x_i) \end{pmatrix} \right]^{-1} \left[ \sum_{j \neq i}^{n} W_{h,ji} \begin{pmatrix} 1 \\ (x_j - x_i) \end{pmatrix} y_j \right],
\]

where \( W_{h,ji} = \prod_{s=1}^{d} h^{s-1}_s w((x_j - x_i)/h_s) \).

Define two column vectors \( e_g = (1, 0_{1 \times d})^T \) and \( e_\beta = (0, 1)^T \), where \( 0_{1 \times d} \) is a \( 1 \times d \) row vector of zeros, and \( 1 = (1, \ldots, 1)^T \) is a \( d \times 1 \) vector of ones. The leave-one-out kernel estimators of \( g(x_i) \) and \( \beta(x_i) \) are given by \( \hat{g}_{-i}(x_i) = e_g^T \hat{\delta}_{-i}(x_i) \) and \( \hat{\beta}_{-i}(x_i) = e_\beta^T \hat{\delta}_{-i}(x_i) \), respectively.

We will choose \( h = (h_1, \ldots, h_d) \) to minimize a feasible version of the following (infeasible) cross-validation objective function:

\[
C_{0,\beta}(h) = \frac{1}{n} \sum_{i=1}^{n} \| \hat{\beta}_{-i}(x_i) - \beta(x_i) \|^2 M(x_i),
\]

where \( \| A \|^2 = A^T A \) for a \( d \times 1 \) vector \( A \) and \( M(\cdot) \) is a compactly supported weight function that trims out observations near the boundary of the support of \( x_i \). The objective function \( C_{0,\beta}(h) \) defined in (9) is infeasible because \( \beta(x_i) \) is unobservable. Below we will derive a feasible quantity that mimics (9).
2.3. A Feasible Cross-Validation Function. We first re-write (8) in an equivalent form. Inserting the identity matrix $I_{d+1} = G_n^{-1}G_n$ into the middle of (8), where $G_n = \begin{pmatrix} 1, & 0 \\ 0, & D_h^{-2} \end{pmatrix}$, $D_h^{-2} = \text{diag}(h_s^{-2})$ (a $d \times d$ diagonal matrix where the $s^{th}$ diagonal equals $h_s^{-2}$), and note that $A_n^{-1}B_n = A_n^{-1}G_n^{-1}G_nB_n = (G_nA_n)^{-1}G_nB_n$, we obtain

$$\delta_{-i}(x_i) = \left[ \sum_{j \neq i} W_{h,ji}G_n \left( \frac{1}{n} (x_j - x_i) \right) \right]^{-1} \sum_{j \neq i} W_{h,ji}G_n \left( \frac{1}{n} (x_j - x_i) \right) y_j$$

$$= \left[ \sum_{j \neq i} W_{h,ji} \left( D_h^{-2}(x_j - x_i) \right) \right]^{-1} \sum_{j \neq i} W_{h,ji} \left( D_h^{-2}(x_j - x_i) \right) y_j.$$  

(10)

The advantage of using (10) is that, conditional on $x_i$,

$$n^{-1} \sum_{j \neq i} W_{h,ji} \left( \frac{1}{D_h^{-2}(x_j - x_i)} \right)$$

converges in probability to a non-singular matrix. Hence, we can analyze the denominator and numerator of (10) separately and this simplifies the derivations substantially.

Recall that $R_{ji} = g(x_j) - g(x_i) - (x_j - x_i)\beta(x_i)$. We can write $y_j$ as

$$y_j = g(x_j) + u_j = g(x_i) + (x_j - x_i)^T\beta(x_i) + R_{ji} + u_j$$

$$= \left( 1, (x_j - x_i)^T \right) \begin{pmatrix} g(x_i) \\ \beta(x_i) \end{pmatrix} + R_{ji} + u_j.$$  

(11)

Substituting $y_j$ in (10) with (11), leads to

$$\delta_{-i}(x_i) = \delta(x_i) + A_2^{-1}A_1i,$$

where

$$A_1i = \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( \frac{1}{D_h^{-2}(x_j - x_i)} \right) [R_{ji} + u_j],$$

(12)

and

$$A_2i = \left( \frac{\hat{f}_i}{D_h^{-2}B_1i}, \frac{B_1^T_i}{D_h^{-2}B_2i} \right),$$

(13)

where $\hat{f}_i = n^{-1} \sum_{j \neq i} W_{h,ji}$, $B_1i = n^{-1} \sum_{j \neq i} W_{h,ji}(x_j - x_i)$, and $B_2i = n^{-1} \sum_{j \neq i} W_{h,ji}(x_j - x_i)(x_j - x_i)^T$. 
Recall that \( e_\beta = (0, i^T)^T \), is a \((d+1)\) column vector whose first element is zero with all other elements equal to one. Therefore we have \( \hat{\beta}_{-i}(x_i) = e_\beta \odot \hat{\delta}_{-i}(x_i) = \beta(x_i) + e_\beta \odot A_{2i}^{-1} A_{1i} \). Rearranging terms we obtain

\[
\hat{\beta}_{-i}(x_i) - \beta(x_i) = e_\beta \odot A_{2i}^{-1} A_{1i}. \tag{14}
\]

\( A_{2i} \) is observable but \( A_{1i} \) is not. Below we will find a consistent estimate of \( A_{1i} \). From (12) we know that we need a consistent estimate of \( R_{ji} + u_j \). Notice that

\[
R_{ji} + u_j \equiv g(x_j) - g(x_i) - (x_j - x_i)^T \beta(x_i) + u_j = y_j - g(x_i) - (x_j - x_i)^T \beta(x_i). \tag{15}
\]

The above quantity can be consistently estimated by nonparametric (leave-one-out) estimators of \( g(x_i) \) and \( \beta(x_i) \), say \( \tilde{g}_{-i}(x_i) \) and \( \tilde{\beta}_{-i}(x_i) \). Then we can estimate \( A_{1i} \) by

\[
\hat{A}_{1i} = \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( \frac{1}{D_h^{-2}(x_j - x_i)} \right) \left[ y_j - \tilde{g}_{-i}(x_i) - (x_j - x_i)^T \tilde{\beta}_{-i}(x_i) \right]. \tag{16}
\]

It may be tempting to use the local-linear estimators to substitute the nonparametric estimators \( \tilde{g}_{-i}(x_i) \) and \( \tilde{\beta}_{-i}(x_i) \) in (16). However, this will not work as \( \hat{A}_{1i} \) as defined in (16) represents the vector of normal equations for the local-linear estimator, see (6). Hence, (16) becomes identically zero if we replace \( \tilde{g}_{-i}(x_i) \) and \( \tilde{\beta}_{-i}(x_i) \) in (16) by the local-linear estimators \( \hat{g}_{-i}(x_i) \) and \( \hat{\beta}_{-i}(x_i) \). In this paper we propose to use the local-constant estimates to substitute \( \tilde{g}_{-i}(x_i) \) and \( \tilde{\beta}_{-i}(x_i) \) in (16), i.e.,

\[
\hat{g}_{-i}(x_i) = \frac{n^{-1} \sum_{j \neq i} y_j W_{h,ji}}{n^{-1} \sum_{j \neq i} W_{h,ji}}, \tag{17}
\]

\[
\hat{\beta}_{-i}(x_i) = \left[ \frac{n^{-1} \sum_{j \neq i} y_j W'_{h,ji}}{n^{-1} \sum_{j \neq i} W_{h,ix}} - \frac{(n^{-1} \sum_{j \neq i} y_l W_{h,li})(n^{-1} \sum_{j \neq i} W'_{h,ji})}{(n^{-1} \sum_{j \neq i} W_{h,ji})^2} \right] = \frac{n^{-2} \sum_{j \neq i} \sum_{l \neq i,j} (y_j - y_l) W'_{h,ji} W_{h,li}}{(n^{-1} \sum_{j \neq i} W_{li})^2}, \tag{18}
\]

where \( W'_{h,ji} \) is a \( d \times 1 \) vector, its \( s^{th} \) element is given by \((s = 1, \ldots, d)\)

\[
W'_{h,ji,s} = \frac{\partial}{\partial x_{is}} W_{h,ji} = - \left[ \prod_{l \neq s} h_l^{-1} w \left( \frac{x_jt - x_it}{h_t} \right) \right] \frac{1}{h_s^2} \frac{\partial w(u)}{\partial u} \bigg|_{u=(x_is-x_{is})/h_s}.
\]

Hence, substituting (14) into (9), and replacing \( A_{1i} \) by \( \hat{A}_{1i} \) defined in (16) with \( \hat{g}_{-i}(x_i) \) and \( \hat{\beta}_{-i}(x_i) \) defined in (17) and (18), we obtain a feasible objective function:

\[
C_\beta(h) = \frac{1}{n} \sum_{i=1}^n \| e_\beta A_{2i}^{-1} \hat{A}_{1i} \|^2 M(x_i). \tag{19}
\]

Our GBCV method selects \( h \) that minimizes \( C_\beta(h) \) defined in (19). We will use \( \hat{h} \) to denote the GBCV selected \( h \). The asymptotic behavior of \( \hat{h} \) is the subject of the following discussion.
2.4. Theoretical Properties. We begin our discussion of the theoretical properties of our bandwidth selection mechanism by listing our assumptions. These are as follows:

Assumption 2.1. (i) The data \( \{x_i, y_i\}_{i=1}^n \) are independent and identically distributed (i.i.d.), \( x_i \) admits a density function \( f(\cdot) \). (ii) Let \( g(x_i) = E(y_i|x_i) \). \( g(x) \) has continuous partial derivative functions up to fourth-order on \( x \in \mathcal{M} \), where \( \mathcal{M} \) is the support of the trimming function (\( \mathcal{M} \) a compact subset of \( \mathbb{R}^d \)). (iii) \( h(x) \) has continuous partial derivatives up to second-order on \( x \in \mathcal{M} \).

Assumption 2.2. (i) Let \( u_i = y_i - g(x_i) \). Then \( \sigma^2(x) = E(u_i^2|x_i = x) \) is a continuous function on \( x \in \mathcal{M} \). (ii) Define \( \mu_m(x_i) = E(u_i^m|x_i), \mu_m(x) \) is bounded on \( x \in \mathcal{M} \) for all finite positive \( m \).

Assumption 2.3. (i) The kernel function is a non-negative, bounded, differentiable even density function \( (w(v) = w(-v)) \); (ii) \( w'(v) = dv(v)/dv \) is a continuous and bounded function; (iii) \( \int w(v)v^6 \) and \( \int |w'(v)|v^5dv \) are both finite.

Assumption 2.4. \( (h_1, ..., h_d) \in H \), where \( H = \{h \in \mathcal{R}^d : c_1n^{-1/(d+\delta_1)} \leq h_s \leq c_2n^{-1/(d+6+\delta_2)} \}, \) for some small positive constant \( \delta_1 > 0 \), and large positive constant \( \delta_2 > 0 \), where \( c_1 \) and \( c_2 \) are positive constants.

We only consider i.i.d. data in this paper although it is well known that most nonparametric estimation asymptotic results remain valid when the independent data assumption is replaced by some weakly dependent data processes such as \( \alpha \) or \( \beta \) mixing processes. Assumption 2.1 imposes some standard smoothness conditions on \( g(x) \) and \( f(x) \). Assumption 2.2 (i) imposes continuity on \( \sigma^2(x) \). Assumption 2.2 (ii) is a common assumption in the literature and is also used in Hall, Li & Racine (2007). Assumption 2.3 is quite standard except that we also assume that the kernel function is differentiable. Note that Assumption 2.4 basically requires that \( \max_{1 \leq s \leq d} h_s \to 0 \) and \( nh_1...h_d \to \infty \) as \( n \to \infty \), which are needed for the estimation bias and variance to converge to 0 as \( n \) gets large.

For expository simplicity our proofs of the main result will focus on the scalar \( x \) case. The following theorem deals with the scalar \( x \) case.

Theorem 2.1. Assuming that \( d = 1 \) and under assumptions 2.1 to 2.4, we have

\[
(i) \quad C_\beta(h) = h^4 \int B(x)M(x)f(x)dx + \frac{1}{nh^3} \zeta_0 \int \sigma^2(x)M(x)dx + o_p \left( h^4 + (nh^3)^{-1} \right)
\]

uniformly in \( h \in H \),

\[
(ii) \quad \hat{h} = c_0 n^{-1/7} + o_p \left( n^{-1/7} \right),
\]

where \( B(x) = (B_1(x) - B_2(x))^2 \), \( B_1(x) = \left( \frac{\mu_x - \mu_2^2}{2\mu_2} \right) \frac{g''(x)f'(x)}{f(x)} + \frac{\mu_2 g''(x)}{\mu_2} \), \( B_2(x) = \mu_2 g''(x)f'(x)f(x) + 2g'(x)f''(x)f(x)\right) + (1/2)g''(x)f(x)^2 - g'(x)f'(x) f(x)^2, \zeta_0 = \kappa_0 + \nu_2/\mu_2^2 - \nu_0/\mu_2, \kappa_0 = \int (w'(v))^2dv, \mu_j = \int w(v)v^jdv, \nu_j = \int w(v)^2v^jdv, \sigma^2(x) = E(u_i^2|x_i = x), c_0 = \left[ \frac{3\kappa_0}{4} \frac{\sigma^2(x)M(x)dx}{B(x)M(x)f(x)dx} \right]^{1/7}, m'(-), m''(-) \) and \( m'''(-) \) denote the first, second and third derivative functions of \( m(\cdot) \) with \( m = g \) or \( m = f \).
The leading terms of $C_{\beta}(h)$ have an interesting connection to the underlying estimators. The first component is related to a weighted version of the estimated MSE of $\hat{\beta}(x)$, while the second piece is related to a weighted version of the estimated MSE of $\tilde{\beta}(x)$, the local-constant estimator of $\beta(x)$. Indeed it is well established that

$$\sqrt{nh^3} \left[ \hat{\beta}(x) - \beta(x) - h^2 B_1(x) \right] \xrightarrow{d} N \left( 0, \frac{\nu_2}{\mu_2} \frac{\sigma^2(x)}{f(x)} \right),$$

see Cai, Fan & Yao (2000), and in the supplement Appendix B we show that

$$\sqrt{nh^3} \left[ \tilde{\beta}(x) - \beta(x) - h^2 B_2(x) \right] \xrightarrow{d} N \left( 0, \frac{\kappa_0 \sigma^2(x)}{f(x)} \right).$$

It can also be shown that

$$\sqrt{nh^3} \text{Cov}(\hat{\beta}(x), \tilde{\beta}(x)) \rightarrow \frac{\nu_0 \sigma^2(x)}{2\mu_2 f(x)}. \tag{22}$$

If we denote the point-wise (leading terms) estimated MSE of the local-linear and local-constant derivative estimators by $MSE[\hat{\beta}(x)] = h^4 B_1(x)^2 + \frac{\nu_2}{nh^3 \mu_2} \frac{\sigma^2(x)}{f(x)}$, and $MSE[\tilde{\beta}(x)] = h^4 B_2(x)^2 + \frac{\kappa_0 \sigma^2(x)}{f(x)}$, respectively. Then Theorem 2.1 states that the leading term of the cross-validation function is

$$C_{\beta}(h) = \int \text{MSE} \left( \hat{\beta}(x) - \tilde{\beta}(x) \right) f(x)M(x)dx + o_p \left( h^4 + (nh^3)^{-1} \right)$$

$$= \int \left\{ \left[ \text{Bias} \left( \hat{\beta}(x) - \tilde{\beta}(x) \right) \right]^2 + \text{Var} \left( \hat{\beta}(x) - \tilde{\beta}(x) \right) \right\} f(x)M(x)dx + o_p \left( h^4 + (nh^3)^{-1} \right). \tag{23}$$

Note that it well known that the local-constant estimate has large bias at the boundary of the data support. Therefore, it is important to use the trimming function $M(\cdot)$ to remove observations near the boundary in constructing the objective function $C_{\beta}(h)$.

The next Theorem deals with the general multivariate $x$ case. For the general $d$-dimensional $x$, we denote $m_x(x) = \frac{\partial m(x)}{\partial x}, m_{tx}(x) = \frac{\partial^2 m(x)}{\partial x \partial x}$, and $m_{ss}(x) = \frac{\partial^3 m(x)}{\partial x^3}$, where $m(x) = g(x)$ or $m(x) = f(x)$. Then we have the following result:

**Theorem 2.2.** Under assumptions 2.1 to 2.4, we have

(i) $C_{\beta}(h) = \int B_h(x)M(x)f(x)dx + \frac{\nu_0^{d-1} c_0}{nh_1...h_d} \left( \sum_{s=1}^d \frac{1}{h_s^2} \right) \int \sigma^2(x)M(x)dx + o_p \left( \eta_n \right)$,

uniformly in $h \in H_n$, where $\eta_n = ||h||^4 + \frac{1}{nh_1...h_d ||h||^2}, ||h||^2 = \sum_{s=1}^d h_s^2$,

(ii) $h_s = c_{0,s} n^{-1/(d+6)} + o_p \left( n^{-1/(d+6)} \right)$,
where $\zeta_0 = [\nu_2/\mu_2^2 + \kappa_0 - \nu_0/\mu_2]$, $B_h(x) = \sum_{s=1}^d (B_{1s,h}(x) - B_{2s,h}(x))^2$ with

\[
B_{1s,h}(x) = \frac{h_s^2 \mu_4 g_{sss}(x)}{6 \mu_2} + \frac{h_s^2 g_{ss}(x) f_s(x) \mu_4}{2 \mu_2 f(x)} - \frac{\mu_2 f_s(x) \sum_{t=1}^d g_{tt}(x) h_t^2}{2 f(x)} + \frac{1}{2} \frac{f_s(x) \sum_{t \neq s} h_t^2 g_{ts}(x)}{2 f(x)} + \mu_2 \sum_{t \neq s} h_t^2 f_t(x) g_{ts}(x)/f(x) + (\mu_2/2) \sum_{t \neq s} h_t^2 g_{ts}(x),
\]

\[
B_{2s,h}(x) = \frac{\mu_2}{f(x)} \left\{ \frac{1}{2} g_{sss}(x) f(x) h_s^2 + \frac{f_s(x)}{f(x)} \sum_{t=1}^d g_{ts}(x) f_t(x) h_t^2 + g_s(x) \sum_{t=1}^d f_{tt}(x) h_t^2 + \sum_{t=1}^d [f_t(x) g_{ts}(x) + g_t(x) f_{ts}(x)] h_t^2 \right\}
\]

and $c_{0,s}$ are positive constants, $s = 1, \ldots, d$.

We can see from Theorem 2.2 that the leading squared bias term is complicated as it involves partial smoothing derivatives up to the 3rd order. If we were to use a plug-in method to select optimal smoothing parameter values need to be selected and estimates for all the related partial derivative functions of $g$ and $f$ are required. In the multivariate regression case, this can be a daunting task. In contrast, our GBCV method delivers optimally selected smoothing parameters in a fully automatic, data-driven procedure.

The proof of Theorem 2.2 is similar, but much more tedious than the proof of Theorem 2.1. Although we do not provide a detailed proof for Theorem 2.2, we provide a sketch of a proof for one leading term of $CV(\cdot)$ (the cross term) in lemma A.4 of Appendix A. The results for other leading terms can be similarly proved. Also, in the supplement Appendix B (which is available from the authors upon request) we derive the leading terms of $MSE(\hat{\beta}(x))$ and $MSE(\tilde{\beta}(x))$ in lemmas B.6 and B.7, respectively. By comparing the result of Theorem 2.2 and those from lemmas B.6 and B.7, we observe that

\[
C_\beta(h) = \int \text{Tr} \left[ \text{MSE} \left( \hat{\beta}(x) - \tilde{\beta}(x) \right) \right] f(x) M(x) dx + o_p \left( ||h||^4 + \frac{1}{n h_1 \ldots h_d ||h||^2} \right),
\]

where $\text{Tr}(\cdot)$ denotes the trace of a matrix. That is, similar to the scalar $x$ case, we still have that the leading term of $C_\beta(h)$ equals (the trace of) the sum of a weighted integrated MSE of $\hat{\beta}(\cdot) - \tilde{\beta}(\cdot)$.

3. Simulation Study

In this section we perform a small set of Monte Carlo simulations to assess the performance of our bandwidth selection procedure relative to the standard LSCV approach. We consider a nonparametric model with heteroskedastic error

$$y_j = g(x_j) + \sigma(x_j) u_j, \quad j = 1, 2, \ldots, n.$$  

We investigate three function specifications for $g(x)$:

**DGP1**: $g(x) = 2 + \sin(1.5x)$
Table 1. Summary of simulation results at the median [10th percentile, 90th percentile] for the gradient.

<table>
<thead>
<tr>
<th></th>
<th>DGP1</th>
<th>DGP2</th>
<th>DGP3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LSCV</td>
<td>GBCV</td>
<td>LSCV</td>
</tr>
<tr>
<td>(n = 50)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>1.5545</td>
<td>0.8894</td>
<td>0.2794</td>
</tr>
<tr>
<td></td>
<td>[0.7705, 23.0919]</td>
<td>[0.4376, 1.9354]</td>
<td>[0.0798, 22.4479]</td>
</tr>
<tr>
<td>(n = 100)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>1.2831</td>
<td>0.6235</td>
<td>0.1734</td>
</tr>
<tr>
<td></td>
<td>[0.4420, 21.1051]</td>
<td>[0.3235, 1.3433]</td>
<td>[0.0673, 19.3226]</td>
</tr>
<tr>
<td>(n = 200)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>0.9885</td>
<td>0.4647</td>
<td>0.1307</td>
</tr>
<tr>
<td></td>
<td>[0.2278, 16.2692]</td>
<td>[0.2369, 0.9595]</td>
<td>[0.0499, 11.9774]</td>
</tr>
</tbody>
</table>

DGP2: \(g(x) = 2 + \frac{x^3}{1 + e^{-3x}}\);  
DGP3: \(g(x) = x + 2e^{-16x^2}\).

We use sample sizes of \(n = 50, 100,\) and \(200\) with 1000 replications per experiment. Our covariate \(x\) is generated from the uniform\([-2, 2]\), \(u\) is distributed standard normal and we set \(\sigma(x) = \sqrt{x^4 + 1}\). Our simulations employ smooth functions with heteroskedastic errors as most economic data display various degrees of heteroskedasticity.

To determine the performance of the proposed bandwidth selection mechanism against the rate adjusted LSCV bandwidths we compare the estimated MSE of \(\beta(x)\) for each bandwidth selector. The rate adjusted bandwidth selector is constructed by multiplying the bandwidths found via LSCV by \(n^{2/35}\) so that the rate is optimal \((n^{1/7})\). We calculate the estimated MSE of \(\beta(x)\) using both bandwidth selection methods over 1000 replications. We then examine the median as well as the 10th and 90th percentiles of these values over the 1000 replications. We use a second-order Epanechnikov kernel for all simulations. The results are presented in Table 1 for each cross-validation procedure. We find improvement for our cross-validation method versus rate-adjusted LSCV in terms of estimated MSE (for all but two median values in DGP2). Moreover, we see drastic improvements at the upper percentile. For example, when \(n = 200\), the 90th percentile of the estimated MSE of the estimated gradient of DGP2 using rate adjusted bandwidths selected via LSCV is over 30 times larger than that of the estimated MSE of the gradients estimated with the GBCV bandwidths. This showcases the major benefit of our approach. In the cases where LSCV severely undersmooths (even after rate-adjustment), GBCV bandwidths still perform well. This likely represents a case where LSCV is modeling noise while GBCV continues to reliably estimate the smooth function.

What we have shown in our simulations is that our bandwidth selector generally performs better in terms of MSE. Many of these improvements come at the higher percentiles (of our simulations) where the standard rate adjusted LSCV bandwidth estimator often undersmooths substantially. In fact, if we were to plot the distribution of the MSE terms from the replications, we would see that the popular LSCV method is associated with a long tail.

Our method is even more useful in multivariate settings. Limited (unreported) simulations show that, with additional regressors, larger amounts of data are needed to properly estimate the gradient.
of the conditional mean function and the standard LSCV method tends to overfit the data relative to GBCV. Given that economic data often have a relatively high noise-to-signal ratio, as well as the fact that samples are often small or moderate with relatively large numbers of covariates, we believe that our method will often lead to improved performance in practice. We give a prime multivariate example of such a situation in the following section.

4. Empirical Application: Public/Private Capital Productivity Puzzle

The Monte Carlo results in the previous section showcase the finite sample performance of GBCV. We also saw in Section 2.1 how this approach works well in a simple univariate regression model. However, this type of simple problem is uncommon in economics and we feel it necessary to provide an economic application with multivariate data. In this section we apply the aforementioned procedures to the well known public capital productivity puzzle debate. Baltagi & Pinnoi (1995) use the following production function

$$y_{jt} = \alpha + \beta_1 k_{gjt} + \beta_2 k_{pjt} + \beta_3 e_{mpjt} + \beta_4 u_{nemjt} + \varepsilon_{jt}$$  (24)

to study the public capital productivity puzzle. Here, $y_{jt}$ denotes the gross state product of state $j$ ($j = 1, ..., 48$) in period $t$ ($t = 1970, ..., 1986$). Covariates include public capital ($k_g$) which aggregates highways and streets, water and sewer facilities, and other public buildings and structures, $k_p$ is the Bureau of Economic Analysis' private capital stock estimates, and labor ($e_{mp}$) is employment in non-agricultural payrolls. Each of these variables are measured in logarithms. Following Baltagi & Pinnoi (1995), we also use the unemployment rate ($u_{nem}$) to control for business cycle effects. Details on these variables can be found in Munnell (1990) as well as Baltagi & Pinnoi (1995). $\varepsilon_{jt}$ is our mean zero additive error term.3

<table>
<thead>
<tr>
<th></th>
<th>Bandwidths</th>
<th>Scale Factors</th>
<th>Ratio GBCV/LSCV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LSCV</td>
<td>GBCV</td>
<td>LSCV</td>
</tr>
<tr>
<td>ln($k_g$)</td>
<td>0.0825</td>
<td>0.6018</td>
<td>0.2023</td>
</tr>
<tr>
<td>ln($k_p$)</td>
<td>0.1086</td>
<td>0.5618</td>
<td>0.2716</td>
</tr>
<tr>
<td>ln($e_{mp}$)</td>
<td>0.1336</td>
<td>0.4893</td>
<td>0.3032</td>
</tr>
<tr>
<td>$u_{nem}$</td>
<td>2.9516</td>
<td>1.2352</td>
<td>3.0555</td>
</tr>
</tbody>
</table>

We estimate nonparametric versions of Equation (24) using local-linear least-squares with second-order Epanechnikov kernels and bandwidths selected by both LSCV and GBCV. Table 2 provides the estimated bandwidths using both methods. We see that while we expect larger bandwidths

3 Although we have only considered the independent data case in our theoretical analysis, there is no doubt the asymptotic analysis can be extended to weakly dependent data and panel data cases. We leave this theoretical investigation as a future research topic.
from GBCV relative to LSCV, the bandwidth for unemployment actually decreases (the other three bandwidths increase as expected). The ratio of the estimated bandwidths reveal that indeed the two methods can provide substantially different bandwidths in applied settings and that the GBCV bandwidths do not necessarily rise. Further, these ratios show that we do not just get a generic percentage or absolute increase. Since we cannot know the ‘optimal’ bandwidth for an actual dataset without prior information on the DGP, we resort to analyzing the economic implications of the two sets of bandwidths.

Table 3 provides the estimated gradients (elasticities) from local-linear estimation using both LSCV and GBCV bandwidths. We present the 10th and 90th percentile point estimates, $D_{10}$ and $D_{90}$, respectively, along with the median point estimates (Median) as well as the standard deviation (SD) across the point estimates for each regressor. We see that for all four estimated gradients, the GBCV estimated gradients lie in narrow ranges relative to the LSCV gradients estimates. More importantly, from an economic standpoint, the estimated LSCV gradients for the logarithm of public capital straddles zero. Its $t$-ratio at the median value is around one, which is statistically insignificant, suggesting that public capital does not have a significant role in explaining gross state product. This is the so-called public capital productivity ‘puzzle’. In contrast, our GBCV elasticities are predominantly positively signed. The median $t$-ratio is greater than five, suggesting that public capital has a significant positive impact on gross state product.

Figure 2 provides kernel density estimates for the estimated gradients with respect to ‘public capital’ ($kg$) and ‘private capital’ ($kp$) using the LSCV and GBCV bandwidths. Consistent with Table 3, we see that our GBCV estimated densities (dashed line) are inside the range of the estimated LSCV densities. This figure shows that this is not simply a tails problem. Large percentages of the elasticities produced with the LSCV bandwidths are in regions which are economically unreasonable. For example, we are unsure how to interpret an elasticity greater than one for physical or public capital.

Of equal interest is the effect on the standard errors of the gradient estimates. We use a wild bootstrap with 1,000 replications in order to calculate each standard error. For the LSCV bandwidths, many elasticities are estimated with a lower level of precision. On the other hand, the standard errors for the gradients produced with GBCV bandwidths are much smaller. Here we find many more cases of statistical significance.

To further highlight our bandwidth selection mechanism, we present Figure 3 which shows three-dimensional plots of the estimated conditional means for public capital as public capital and employment vary (we hold private capital and unemployment fixed at their respective medians) for each bandwidth procedure. Panel (a) uses the bandwidths from LSCV to construct this grid of estimates. It is difficult to provide an intuitive description of this figure. The shapes of the curves do not reflect the assumed relationship between these variables and they are quite bumpy. On the

---

4Negative elasticities for public capital are not necessarily incorrect. We examined those states with negative elasticities and found that these states have large relative investments in highways. These particular states (primarily plains states) have major highways running through them (designed to transport goods through their respective states) while at the same time their gross state products are relatively small.
Table 3. Summary statistics of the estimated gradients for local-linear least-squares estimation using the bandwidths in Table 2. $D_{10}$ and $D_{90}$ refer to the 10th and 90th deciles of the estimates while Median refers to the median. Standard errors are given below each estimate (1,000 wild bootstraps). SD refers to the standard deviation across the estimated gradients for a given regressor.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$D_{10}$</th>
<th>Median</th>
<th>$D_{90}$</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_{kg}$ (LSCV)</td>
<td>-0.4014</td>
<td>0.1829</td>
<td>0.5391</td>
<td>0.9772</td>
</tr>
<tr>
<td></td>
<td>0.0738</td>
<td>0.1485</td>
<td>0.1999</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{kg}$ (GBCV)</td>
<td>-0.0377</td>
<td>0.1521</td>
<td>0.2520</td>
<td>0.1442</td>
</tr>
<tr>
<td></td>
<td>0.0389</td>
<td>0.0289</td>
<td>0.0215</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{kp}$ (LSCV)</td>
<td>-0.0964</td>
<td>0.1878</td>
<td>0.6997</td>
<td>0.3867</td>
</tr>
<tr>
<td></td>
<td>0.1206</td>
<td>0.0818</td>
<td>0.0912</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{kp}$ (GBCV)</td>
<td>0.2229</td>
<td>0.2941</td>
<td>0.3761</td>
<td>0.0696</td>
</tr>
<tr>
<td></td>
<td>0.0338</td>
<td>0.0417</td>
<td>0.0264</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{emp}$ (LSCV)</td>
<td>0.0040</td>
<td>0.7454</td>
<td>1.0764</td>
<td>0.4391</td>
</tr>
<tr>
<td></td>
<td>0.0673</td>
<td>0.2332</td>
<td>0.1878</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{emp}$ (GBCV)</td>
<td>0.5282</td>
<td>0.6432</td>
<td>0.7052</td>
<td>0.0793</td>
</tr>
<tr>
<td></td>
<td>0.0590</td>
<td>0.0615</td>
<td>0.0422</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{unem}$ (LSCV)</td>
<td>-0.0201</td>
<td>-0.0059</td>
<td>0.0060</td>
<td>0.0118</td>
</tr>
<tr>
<td></td>
<td>0.0049</td>
<td>0.0036</td>
<td>0.0050</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{unem}$ (GBCV)</td>
<td>-0.0143</td>
<td>-0.0048</td>
<td>0.0039</td>
<td>0.0118</td>
</tr>
<tr>
<td></td>
<td>0.0041</td>
<td>0.0039</td>
<td>0.0059</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Comparison of estimated gradients across bandwidths obtained via LSCV (solid line) versus GBCV (dashed line). These curves are the estimated density of the gradient estimates using a second-order Gaussian kernel with the Silverman rule-of-thumb bandwidth.
other hand, Panel (b) shows that the estimated conditional mean is quite smooth and we do not see any of the large upward or downward spikes.

The result from the GBCV bandwidths are in line with conventional wisdom. We see that as the levels of private and public capital rise, we expect to have higher values of gross state product. Further, the nonlinear relationship shows that the return to public capital is lower than that of private capital. Although an optimizing model would suggest equal returns to each type of capital, finding a political process to allocate government capital in such an optimal manner is difficult. This result is intuitive.

In summary, in this particular application, GBCV has produced a set of bandwidths which lead to clearer insights than those using the traditional LSCV bandwidths. The bandwidths estimated using GBVC produced an estimator that was much smoother than the estimator using bandwidths obtained by LSCV. We saw less variation in the point estimates. Further, a large percentage of the elasticities produced with the LSCV bandwidths were economically infeasible. The elasticities from our proposed bandwidth selection criteria are in line with conventional wisdom. It is easy to see where this could lead to improved policy analysis. We expect an assortment of applied studies could also benefit from this new approach.

5. Concluding Remarks

In this paper we propose a novel approach to select bandwidths in nonparametric kernel regression. In contrast to previous research on bandwidth selection focusing on the unknown conditional
mean, we are primarily concerned with estimation of the gradient function. Estimation of gradients is often of more interest as studying ‘marginal effects’ is a cornerstone of microeconomics. Uncovering gradients nonparametrically is important in many areas of economics such as determining risk premium or recovering distributions of individual preferences. Our procedure is shown to deliver bandwidths with the optimal rate for estimation of the gradients. Our simulations show large improvements in performance when estimating the gradient function. When we applied our method to empirical data sets, we found that our procedure gave estimates in line with conventional wisdom while the standard cross-validation procedure produced estimates which displayed substantial variation.

There exist many possible extensions of our proposed method. For example, we can extend our method to the case of selecting smoothing parameters that are optimal for estimating higher-order derivatives. Also, we only consider the case of independent data with continuous covariates. The result of this paper can be extended to the weakly dependent data case, and to the mixture of continuous and discrete covariates case. Finally, given that a multivariate nonparametric regression model suffers from the ‘curse of dimensionality’, it will be useful to extend our result to various semiparametric models such as the partially linear or varying coefficient models. We leave these research problems as future research topics.
References


APPENDIX A. PROOF OF THEOREMS 2.1 AND 2.2

The proof of Theorem 2.1 is quite tedious. Therefore, it is necessary to introduce some shorthand notation and preliminary manipulations in order to simplify the derivations that follow. For reader’s convenience we list most of the notations used in the appendices below.

1. We will use the short hand notation: $g_i = g(x_i)$, $\beta_i = \beta(x_i)$, $f_i = f(x_i)$, $\hat{f}_i = \hat{f}(x_i)$, $\bar{g}_i = \bar{g}_i(x_i)$, $\bar{\beta}_i = \bar{\beta}_i(x_i)$, etc. Also, we will often omit the weight function $M(x_i)$ to save space.

2. We define $\sum_i = \sum_{i=1}^n$, $\sum_{j \neq i} = \sum_{j=1,j \neq i}^n$, $\sum_{i,j \neq l} = \sum_{i=1,j \neq l}^n$, $\sum_{i,j \neq l,k \neq m} = \sum_{i=1,j \neq l,k \neq m}^n$. Similarly, $\sum_{i,j \neq l,k \neq m}^n$ means that all the five summation indices $i,j,l,k,m$ are different from each other.

3. We write $A_n = B_n + (s.o.)$ to denote the fact that $B_n$ is the leading term of $A_n$, where (s.o.) denotes terms that have probability orders smaller than that of $B_n$. $A_i = B_i + (s.o.)$ means that $n^{-1} \sum_i A_i = n^{-1} \sum_i B_i + (s.o.)$, or $n^{-1} \sum_i A_i D_i = n^{-1} \sum_i B_i D_i + (s.o.)$ for some $D_i$, i.e., $B_i$ is the leading term of $A_i$, replacing $A_i$ by $B_i$ will not affect the asymptotic result. Also, we write $A_n \sim B_n$ to mean that $A_n$ and $B_n$ have the same probability order.

4. We often ignore the difference among $\frac{1}{n}$, $\frac{1}{n-1}$ and $\frac{1}{n-2}$ simply because this will have no effect on the asymptotic analysis.

Proof of Theorem 2.1: We first analyze $\hat{A}_{1i}$. By adding and subtracting terms in (16), using $y_j = g_j + u_j$, $R_{ji} = g_j - g_i - (x_j - x_i)\beta_i$ and noticing that $D_h^{-2} = h^{-2}$ (since $x$ is a scalar), we get

$$
\hat{A}_{1i} = \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( \frac{1}{h^2 (x_j - x_i)} \right) \left\{ g_j + u_j - g_i - (x_j - x_i)\beta_i - [g_j - g_i - (x_j - x_i)\beta_i] \right\} \\
= \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( \frac{1}{h^2 (x_j - x_i)} \right) \left\{ R_{ji} + u_j - [g_i - g_j] - (x_j - x_i)[\beta_i - \beta_j] \right\} \\
= A_{1i} - \Delta_i
$$

(A.1)

where $A_{1i}$ is defined in (12), and

$$
\Delta_i = \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( \frac{1}{h^2 (x_j - x_i)} \right) \left\{ [g_i - g_j] + (x_j - x_i)[\beta_i - \beta_j] \right\}. \quad \text{(A.2)}
$$

Substituting (A.1) into (19), we obtain

$$
C_\beta(h) = \frac{1}{n} \sum_{i=1}^n \left[ e_\beta^T A_{2i}^{-1} A_{1i} - e_\beta^T A_{2i}^{-1} \Delta_i \right]^2 \\
= \frac{1}{n} \sum_{i=1}^n \left[ (e_\beta^T A_{2i}^{-1} A_{1i})^2 + (e_\beta^T A_{2i}^{-1} \Delta_i)^2 - 2(e_\beta^T A_{2i}^{-1} A_{1i})(e_\beta^T A_{2i}^{-1} \Delta_i) \right] \\
= C_1 + C_2 - 2C_3 \quad \text{(A.3)}
$$

where $C_1 = n^{-1} \sum_{i=1}^n (e_\beta^T A_{2i}^{-1} A_{1i})^2$, $C_2 = n^{-1} \sum_{i=1}^n (e_\beta^T A_{2i}^{-1} \Delta_i)^2$ and $C_3 = n^{-1} \sum_{i=1}^n (e_\beta^T A_{2i}^{-1} A_{1i})(e_\beta^T A_{2i}^{-1} \Delta_i)$. 


In Lemmas A.1 to A.3 below we show that
\[
C_1 = h^4 \int B_1(x)^2 M(x) f(x) dx + \frac{1}{n h^3 \mu_2^2} \int \sigma(x)^2 M(x) dx + (s.o.),
\]
\[
C_2 = h^4 \int B_2(x)^2 M(x) f(x) dx + \frac{1}{n h^3 \kappa_0} \int \sigma(x)^2 M(x) dx + (s.o.),
\]
\[
C_3 = h^4 \int B_1(x) B_2(x) M(x) f(x) dx + \frac{1}{n h^3 \mu_2^2} \int \sigma^2(x) M(x) dx + (s.o.),
\]
uniformly in \( h \in H_n \).

Note that \( B(x) = (B_1(x) - B_2(x))^2 \) and \( \kappa_0 = \kappa_0 + \nu_2/\mu_2 - \nu_0/\mu_2 \). This proves Theorem 2.1 (i). Theorem 2.1 (ii) follows from Theorem 2.1 (i).

\[\Box\]

**Proof of Theorem 2.2**

For vector case, we have
\[
C_{\beta}(h) = C_1 + C_2 - 2C_3,
\]
where \( h = (h_1, ..., h_d) \), \( C_1 = n^{-1} \sum_i (e_{i}^T A_{2i}^{-1} A_{1i})^T e_{i}^T A_{2i}^{-1} A_{1i}, \)
\( C_2 = n^{-1} \sum_i (e_{i}^T A_{2i}^{-1} \delta_i)^T e_{i}^T A_{2i}^{-1} \Delta_{1i} \)
and \( C_3 = n^{-1} \sum_i (e_{i}^T A_{2i}^{-1} A_{1i})^T e_{i}^T A_{2i}^{-1} \Delta_{1i} \). In Lemma A.4 we show that
\[
C_3 = \int \sum_{s=1}^{d} B_{1s,h}(x) B_{2s,h}(x) M(x) f(x) dx + \frac{\nu_0}{\mu_2} \frac{1}{n h_{1...h_d}} \sum_{s=1}^{d} h_{s}^{-2} + o_p(\eta_n), \tag{A.4}
\]
where \( \eta_n = ||h||^4 + (n h_{1...h_d} ||h||^2)^{-1} \). Similarly, one can show that
\[
C_1 = \int \sum_{s=1}^{d} B_{1s,h}(x)^2 M(x) f(x) dx + \frac{\nu_0}{\mu_2} \frac{1}{n h_{1...h_d}} \sum_{s=1}^{d} h_{s}^{-2} + o_p(\eta_n) \tag{A.5}
\]
\[
C_2 = \int \sum_{s=1}^{d} B_{2s,h}(x)^2 M(x) f(x) dx + \frac{\nu_0}{\mu_2} \frac{1}{n h_{1...h_d}} \sum_{s=1}^{d} h_{s}^{-2} + o_p(\eta_n). \tag{A.6}
\]

Theorem 2.2 follows from (A.4), (A.5) and (A.6).

**Lemma A.1.** \( C_1 = h^4 \int B_1(x)^2 M(x) f(x) dx + \frac{\nu_0}{n h^3 \mu_2^2} \int \sigma(x)^2 M(x) dx + o_p(h^4 + (nh^3)^{-1}) \) uniformly in \( h \in H_n \), where \( \mu_2 = \int w(v) v^2 dv \).

**Proof of Lemma A.1:** Using the standard kernel estimation uniform convergence result, we have
\[
A_{2i} = \begin{pmatrix} f(x_i) & 0 \\ \mu_2 f'(x_i) & \mu_2 f(x_i) \end{pmatrix} + O_p \left( h^2 + \frac{(ln(n))^{1/2}}{\sqrt{nh}} \right),
\]
uniformly in \( x_i \in M \), where \( M \) is the support of the trimming function \( M(\cdot) \).

Using the partitioned inverse, we get
\[
A_{2i}^{-1} = \begin{pmatrix} 1/f(x_i) & 0 \\ -C_{1i} & C_{2i} \end{pmatrix} + O_p \left( h^2 + \frac{(ln(n))^{1/2}}{\sqrt{nh}} \right),
\]
where \( C_{1i} = f'(x_i)/f^2(x_i) \) and \( C_{2i} = 1/(\mu_2 f(x_i)) \).

Recall that \( e_\beta = (0,1)^T \), we have

\[
e_\beta^TA^{-1}_{2i}(A.7)
\]

Combining (A.7) and (12) lead to

\[
e_\beta^TA^{-1}_{2i}A_{1i} = n^{-1} \sum_{j \neq i} W_{h,j}(R_{ji} + u_j)[C_{2i}h^{-2}(x_j - x_i) - C_{1i}] + (s.o.)
\]

\[
= n^{-1} \sum_{j \neq i} P_{ji}(R_{ji} + u_j) + (s.o.),
\]

where

\[
P_{ji} = W_{h,j}(C_{2i}h^{-2}(x_j - x_i) - C_{1i}).
\]

Substituting the above result into \( C_1 \) gives,

\[
C_1 = n^{-1} \sum_{i=1}^{n} (e_\beta^TA_{2i}^{-1}A_{1i})^2
\]

\[
= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq j} P_{ji}P_{li}(R_{ji} + u_j)(R_{li} + u_l) + (s.o.)
\]

\[
= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq j} P_{ji}^2(R_{ji} + u_j)^2 + n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq j} P_{ji}P_{li}(R_{ji} + u_j)(R_{li} + u_l) + (s.o.)
\]

\[
= C_{1,1} + C_{1,2} + (s.o.)
\]

where \( C_{1,1} = n^{-3} \sum_i \sum_{j \neq i} P_{ji}^2(R_{ji} + u_j)^2 \) and \( C_{1,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq j} P_{ji}P_{li}(R_{ji} + u_j)(R_{li} + u_l). \)

From \( (R_{ji} + u_j)^2 = R_{ji}^2 + 2u_jR_{ji} + u_j^2 \), it is easy to see that the leading term of \( C_{1,1} \) is associated with the term \( u_j^2 \), we use \( C_0^{1,1} \) to denote this leading term.

\[
C_0^{1,1} = n^{-3} \sum_i \sum_{j \neq i} P_{ji}^2u_j^2 = n^{-1}E[\sigma_j^2P_{ji}^2] + (s.o.) = n^{-1}E[\sigma_j^2P_{ji}^2|x_i] + (s.o.)
\]

\[
= \frac{\nu_2}{nh^3\mu_2^2} \int \sigma(x)^2dx + (s.o.)
\]

where the second equality follows from the H-decomposition, and the last equality holds because

\[
E\{E[\sigma_j^2P_{ji}^2|x_i]\} = E\{E[W_{h,j}^2u_j^2(C_{1i} - C_{2i}h^{-2}(x_j - x_i))^2|x_i]\}
\]

\[
= E\{E[W_{h,j}^2\sigma(x_j)^2(C_{2i}h^{-4}(x_j - x_i))^2 + C_{1i}^2 - 2C_{2i}h^{-2}(x_j - x_i)C_{1i}]|x_i]\}
\]

\[
= E\left[\frac{\nu_2}{h^3}\sigma(x_i)^2C_{2i}^2f(x_i) + \frac{\nu_0}{h}\sigma(x_i)^2C_{1i}^2f(x_i) - \frac{2\nu_2}{h^3}\sigma(x_i)^2C_{1i}C_{2i}f'(x_i)\right]
\]

\[
= \frac{\nu_2}{h^3\mu_2^2} \int \sigma(x)^2dx + O(h^{-1}),
\]

where \( \sigma^2(x_i) = E(u_i^2|x_i) \) and we used \( C_{2i} = 1/(\mu_2 f(x_i)) \) and \( \nu_2 = \int W(v)v^2dv. \)
Recall that we omit the weight function for notational simplicity, so the above leading term should be \( \frac{\nu_2}{nh^3\mu_2^2} \int \sigma(x)^2 M(x) dx \), which is finite since \( M(\cdot) \) has a bounded support. Hence, we have shown that
\[
C_{1,1} = C_{1,1}^0 + (s.o.) = \frac{\nu_2}{nh^3\mu_2^2} \int \sigma(x)^2 M(x) dx + (s.o.).
\] (A.10)

Next, we consider \( C_{1,2} = n^{-3} \sum_{i} \sum_{j \neq i} \sum_{l \neq j \neq i} P_{ji} R_{li} (R_{ji} R_{li} + u_j u_l + 2 R_{ji} u_l) = C_{1,2,a} + C_{1,2,b} + 2C_{1,2,c} \). \( C_{1,2,a} = n^{-3} \sum_{l \neq j \neq i} P_{ji} P_{li} R_{ji} R_{li} \), which can be written as a third-order U-statistic, whose leading term is its expectation:
\[
E(C_{1,2,a}) = \left\{ E \left( E \left[ P_{ji} R_{ji} | x_i \right] \right) \right\}^2 = \left\{ E \left( E \left[ W_{h,j} (h^{-2} C_{2i} (x_j - x_i) - C_{1i}) R_{ji} | x_i \right] \right) \right\}^2
= h^4 \int B_1(x)^2 f(x) dx + O(h^6),
\]
by lemma B.1 (iii), where \( B_1(x) = \left( \frac{\mu_4 - \mu_2^2}{2\mu_2^2} \right) g''(x) f'(x) + \frac{\mu_4 g'''(x)}{6\mu_2^2} \), \( \mu_l = \int W(v) v^l dv \) (\( l = 2, 4 \)).

Obviously, \( E(C_{1,2,b}) = 0 \) and \( E(C_{1,2,b}^2) = n^{-6} O(n^4 h^{-5}) = O(n^2 h^5)^{-1} = o((nh^3)^{-2}) \). Hence, \( C_{1,2,b} = o_p((nh^3)^{-1}) \).

Similarly, \( E(C_{1,2,c}) = 0 \) and \( E(C_{1,2,c}^2) = n^{-6} [O(n^5 h^6) + O(n^4 h^4)] = O(h^6/n) = o(h^6) \). Hence, \( C_{1,2,c} = o_p(h^4) \).

Summarizing the above we have shown that (adding back the trimming function \( M(\cdot) \))
\[
C_1 = h^4 \int B_1(x)^2 f(x) M(x) dx + \frac{\nu_2}{nh^3\mu_2^2} \int \sigma(x)^2 M(x) dx + o_p(h^4 + (nh^3)^{-1}).
\] (A.11)

Moreover, by using Rosenthal’s and Markov’s inequalities, one can show that (A.11) holds true uniformly in \( h \in H_n \). This completes the proof of lemma A.1. □

**Lemma A.2.** Under the conditions given in Theorem 1, we have, uniformly in \( h \in H_n \),
\[
C_2 = h^4 \int B_2(x)^2 M(x) dx + \frac{\kappa_0}{nh^3} \int \sigma^2(x) M(x) dx + (s.o.).
\]

**Proof of Lemma A.2:** Recall that \( P_{ji} = W_{h,j} [C_{2i} h^{-2} (x_j - x_i) - C_{1i}] \). Then By (A.2) and (A.7), we have
\[
\epsilon_i^T A_{2i}^{-1} \Delta_i = n^{-1} \sum_{j \neq i} P_{ji} \left\{ [\tilde{g}_i - g_i] + (x_j - x_i) [\tilde{\beta}_i - \beta_i] \right\} + (s.o.)
= D_{1i} + D_{2i} + (s.o.),
\] (A.12)
where \( D_{1i} = n^{-1} \sum_{j \neq i} P_{ji} [\tilde{g}_i - g_i] \) and \( D_{2i} = n^{-1} \sum_{j \neq i} P_{ji} (x_j - x_i) [\tilde{\beta}_i - \beta_i] \).

Substituting (A.12) into \( C_2 \) we get
\[
C_2 = \sum_i D_{1i}^2 + \sum_i D_{2i}^2 + 2n^{-1} \sum_i D_{1i} D_{2i} + (s.o.)
= C_{2,1} + C_{2,2} + 2C_{2,3} + (s.o.),
\]
where $C_{2,1} = n^{-1} \sum_i D_{1i}^2$, $C_{2,2} = n^{-1} \sum_i D_{2i}^2$ and $C_{2,3} = n^{-1} \sum_i D_{1i} D_{2i}$.

We first consider $C_{2,1}$. Using $1/\hat{f}_i = 1/f_i + (s.o.)$, we have

$$\tilde{g}_i - g_i = (\tilde{g}_i - g_i) \hat{f}_i/ \hat{f}_i = (\tilde{g}_i - g_i) \hat{f}_i/ f_i + (s.o.) = \tilde{m}_i/ f_i + (s.o.), \quad (A.13)$$

where

$$\tilde{m}_i = (\tilde{g}_i - g_i) \hat{f}_i = \frac{1}{n} \sum_{k \neq i} [g_k + u_k - g_i] W_{h, k}. \quad (A.14)$$

Hence, we can replace $(\tilde{g}_i - g_i)^2$ by $\tilde{m}_i^2/ f_i^2$ to obtain the leading term of $C_{2,1}$, i.e., Using (A.14) we get

$$C_{2,1} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji} P_{li} \tilde{m}_i^2/ f_i^2 + (s.o.)$$

$$= n^{-3} \sum_i \sum_{j \neq i} P_{ji}^2 \tilde{m}_i^2/ f_i^2 + n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji} P_{li} \tilde{m}_i^2/ f_i^2 + (s.o.)$$

$$= D_3 + D_4 + (s.o.),$$

where the definitions of $D_3$ and $D_4$ should be obvious.

$$D_3 \leq \sup_{1 \leq i \leq n} (\tilde{m}_i^2/ f_i) \sum_i \sum_{j \neq i} P_{ji}^2 = O_p \left( h^4 + \frac{\ln(n)}{nh} \right) O_p((nh)^{-1})$$

by lemma B.1 (vi) and the fact that $\sup_{1 \leq i \leq n, x_i \in M} (\tilde{m}_i^2/ f_i^2) = O_p \left( h^4 + \frac{\ln(n)}{nh} \right)$.

To evaluate $D_4$, define $e_{ji} = P_{ji} - E(P_{ji}|x_i)$ so that $P_{ji} = E(P_{ji}|x_i) + e_{ji}$. By Lemma B.1 we know that $E(P_{ji}|x_i) = h^2 G_1(x_i) + O(h^4)$. Using $P_{ji} P_{li} = [h^2 G_1(x_i) + e_{ji} + e_{li}] [h^2 G_1(x_i) + e_{ji} + e_{li}] + (s.o.)$, we can replace $P_{ji} P_{li}$ by $h^4 G_1(x_i)^2 + e_{ji} e_{li} + h^2 G_1(x_i)(e_{ji} + e_{li})$ to obtain

$$D_4 = D_{4,1} + D_{4,2} + 2D_{4,3} + (s.o.),$$

where

$$D_{4,1} = n^{-3} h^4 \sum_i \sum_{j \neq i} \sum_{l \neq i} G_1(x_i)^2 \tilde{m}_i^2/ f_i^2 = h^4 \left[ n^{-1} \sum_i G_1(x_i)^2 \tilde{m}_i^2/ f_i^2 \right]$$

$$= h^4 O_p(h^4 + (nh)^{-1}) \quad (A.15)$$

by lemma B.3 and by noting that $G(x_i) = G_1(x_i) f(x_i)^{-2}$ when applying lemma B.3.

$$D_{4,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} e_{ji} e_{li} \tilde{m}_i^2/ f_i^2 = o_p((nh)^{-1})$$

by lemma B.4.

$$|D_{4,3}| \leq \sqrt{D_{4,1} D_{4,2}} = o_p(h^4 (nh)^{-1/2} + h^2 (nh)^{-1}).$$

Hence, we have shown that

$$C_{2,1} = D_3 + D_4 + (s.o.) = o_p(h^4 + (nh)^{-1}). \quad (A.16)$$
Next, we consider $C_{2,2}$. Define $F_{ji} = (x_j - x_i)P_{ji}$. We have
\[
C_{2,2} = n^{-3} \sum_i \sum_{j \neq i} F_{ji}^2(\tilde{\beta}_i - \beta_i)^2 + n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} F_{ji}F_{il}(\tilde{\beta}_i - \beta_i)^2
\]
\[
= D_5 + D_6,
\]
where the definitions of $D_5$ and $D_6$ should be obvious.
\[
D_5 \leq n^{-1} \left[ \sup_{1 \leq i \leq n} (\tilde{\beta}_i - \beta_i)^2 \right] \left[ n^{-3} \sum_i \sum_{j \neq i} F_{ji}^2 \right]
\]
\[
= n^{-1} O_p \left( h^4 + \frac{\ln(n)}{nh^3} \right) O_p(h^{-1}) = O_p \left( (nh)^{-1} \left( h^4 + \frac{\ln(n)}{nh^3} \right) \right)
\]
by lemma B.1 (v) and the fact that $\sup_{1 \leq i \leq n} (\tilde{\beta}_i - \beta_i)^2 = O_p \left( h^4 + \frac{\ln(n)}{nh^3} \right)$.

Define $\eta_{ji} = F_{ji} - E(F_{ji}|x_i)$, and replacing $F_{ji}$ by $F_{ji} = E(F_{ji}|x_i) + \eta_{ji}$ in $D_6$ we obtain.
\[
D_6 = D_{6,1} + D_{6,2} + 2D_{6,3},
\]
where $D_{6,1} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} E(F_{ji}|x_i)E(F_{il}|x_i)(\tilde{\beta}_i - \beta_i)^2$, $D_{6,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \eta_{ji}\eta_{li}(\tilde{\beta}_i - \beta_i)^2$ and $D_{6,3} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} E(F_{ji}|x_i)\eta_{ji}(\tilde{\beta}_i - \beta_i)^2$.

We first consider $D_{6,1}$. By lemma B.1 (i) we know that $E(F_{ji}|x_i) = 1 - G_2(x_i)h^2 + o(h^2)$. Hence, the leading term of $D_{6,1}$ is obtained by replacing $F_{ji}$ by 1 in $D_{6,1}$, i.e.,
\[
D_{6,1} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} (\tilde{\beta}_i - \beta_i)^2 + (s.o.) = n^{-1} \sum_i (\tilde{\beta}_i - \beta_i)^2 + (s.o.).
\]

By lemma B.2 we know that $\tilde{\beta}_i - \beta_i = h^2 B_{2i} + J_{3i} + (s.o.)$, where $B_{2i} = B_2(x_i)$ and $J_{3i} = n^{-1} \sum_{j \neq i} u_j W_{h,ji}/f_i$. We have
\[
D_{6,1} = n^{-1} \sum_i (h^2 B_{2i} + J_{3i})^2 + (s.o.) = n^{-1} \sum_i (h^4 B_{2i}^2 + J_{3i}^2 + 2h^2 B_{2i}J_{3i}) + (s.o.)
\]
\[
= D_7 + D_8 + 2D_9 + (s.o.),
\]
where $D_7 = n^{-1} h^4 \sum_i B_{2i}^2$, $D_8 = n^{-1} \sum_i J_{3i}^2$ and $D_9 = n^{-1} h^2 \sum_i B_{2i}J_{3i}$.

It is easy to see that
\[
D_7 = n^{-1} h^4 \sum_i B_{2i}(x_i)^2 + (s.o.) = h^4 E[B_2(x_i)^2] + (s.o.).
\]

Next, we consider $D_8$. Using $J_{3i} = n^{-1} \sum_{j \neq i} u_j W'_{h,ji}/f_i$.
\[
D_8 = \frac{1}{n} \sum_i J_{3i}^2 = \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} u_j W'_{h,ji}u_l W'_{h,li}/f_i^2
\]
\[
= \frac{1}{n^3} \sum_i \sum_{j \neq i} u_j^2 (W_{h,ji}')^2/f_i^2 + \frac{1}{n^3} \sum_i \sum_{l \neq i} \sum_{j \neq i} u_j W'_{h,ji}u_l W'_{h,li}/f_i^2
\]
\[
= V_1 + V_2,
\]
where the definitions of $V_2$ and $V_3$ should be obvious.

By the H-decomposition we get

$$
V_1 = E(V_1) + \text{(s.o.)} = \frac{1}{n} E[u_i^2(W'_{h,ji})^2/f_i^2]
$$

$$
= \frac{1}{nh^4} \int \sigma^2(x_j) \left( W' \left( \frac{x_j - x_i}{h} \right) \right)^2 f_i^{-1} f_j dx_i dx_j + \text{(s.o.)}
$$

$$
= \frac{\kappa_0}{nh^3} \int \sigma^2(x) dx + \text{(s.o.)},
$$

where $\kappa_0 = \int (W'(v))^2 dv$.

Next, we consider $V_2 = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i,j} u_j u_l W'_{h,ji} W'_{h,li}/f_i^2$. By H-decomposition the leading term of $V_2$ can be written as a second order degenerate U-statistic. We need to compute $E[W'_{h,ji} W'_{h,li}/f_i^2] = h^{-4} \int W'(x_j-x_i) W'(x_l-x_i) f_i^{-1} dx_i = -h^{-3} \int W'(u) W'(x_l-x_j+u) f(x_j-hu)^{-1} du = -h^{-3} \int W'_j f_j^{-1} + \text{(s.o.)},$ where $W'_j = W'(x_l-x_i)$ with $W'(v) = \int W'(u) W'(v+u) du$.

Hence, we have

$$
V_2 \sim \frac{1}{nh^3} \sum_i \sum_{j > i} u_j u_l \tilde{W}'_{jl}(f_j^{-1} + f_i^{-1}).
$$

It is easy to see that $E(V_2^2) = 2(n^2 h^6) E[\sigma_j^2 \sigma_l^2 (\tilde{W}'_{jl})^2 (f_j^{-1} + f_i^{-1})^2] = (n^2 h^6)^{-1} O(h) = O((n^2 h^5)^{-1})$. Hence,

$$
V_2 = O_p((nh^{5/2})^{-1}) = O_p((nh^3)^{-1}).
$$

Finally, $D_9 = (h^2/n^2) \sum_i \sum_{j \neq i} u_j W'_{h,ji} B_2(x_i)/f_i$. It is easy to see that $E(D_9^2) = (h^4/n^4) [O(n^3 h^{-2}) + O(n^2 h^{-3})] = O(h^2/n)$. Hence, $D_9 = O_p(h^2/n^{1/2}) = O_p(h^4)$.

Summarizing the above we have shown that

$$
D_{6,1} = D_7 + V_1 + \text{(s.o.)} = h^4 \int B_2^2(x) M(x) dx + \frac{1}{nh^3} \kappa_0 \int \sigma^2(x) M(x) dx + \text{(s.o.)} \quad (A.17)
$$

$$
D_{6,2} = o_p(h^4 + (nh^3)^{-1}) \text{ and } D_{6,3} = o_p(h^4 + (nh^3)^{-1}) \text{ by lemma B.5. Hence, we have shown that}
$$

$$
(\text{adding back the weight function } M(\cdot))
$$

$$
C_{2,2} = D_5 + D_6 = D_{6,1} + \text{(s.o.)}
$$

$$
= h^4 \int B_2^2(x) M(x) dx + \frac{1}{nh^3} \kappa_0 \int \sigma^2(x) M(x) dx + \text{(s.o.)}. \quad (A.18)
$$

Finally, by (A.16), (A.18) and Cauchy-Schwartz’s inequality, we have

$$
|C_{2,3}| \leq \sqrt{C_{2,1} C_{2,2}} = o_p(h^4 + (nh^3)^{-1}).
$$

Summarizing the above we have proved that

$$
C_2 = C_{2,2} + \text{(s.o.)} = h^4 \int B_2(x)^2 f(x) M(x) dx + \frac{1}{nh^3} \kappa_0 \int \sigma^2(x) M(x) dx + \text{(s.o.)}.
$$
The above result holds true uniformly in $h \in H_n$ by using Rosenthal’s and Markov’s inequalities. This completes the proof of Lemma A.2.

\[ \]

Lemma A.3. \[ C_3 = h^4 \int B_1(x)B_2(x)M(x)f(x)dx + \frac{1}{2nh} \int \sigma^2(x)M(x)dx + (s.o.) \] uniformly in $h \in H_n$.

Proof of Lemma A.3: Using similar arguments as we did in the proofs of lemmas A.1 and A.2, we have

\[
C_3 = n^{-1} \sum_i e_\beta^T (A_{2i}^{-1} A_{1i})(e_\beta^T A_{2i}^{-1} \Delta_i)
\]

\[
= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}(R_{ji} + u_j) \left[ P_{li}(\tilde{\beta}_i - g_i) + F_{li}(\tilde{\beta}_l - \beta_l) \right] + (s.o.)
\]

\[
= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}(R_{ji} + u_j)F_{li}(\tilde{\beta}_l - \beta_l) + o_p(h^4 + (nh^3)^{-1}) + (s.o.)
\]

\[
= n^{-3} h^2 \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}R_{ji}F_{li}(\tilde{\beta}_l - \beta_l) + n^{-3} h^2 \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}u_j F_{li}(\tilde{\beta}_l - \beta_l) + (s.o.)
\]

\[
= C_{3,1} + C_{3,2} + (s.o.),
\]

where $C_{3,1} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}R_{ji}F_{li}(\tilde{\beta}_l - \beta_l)$ and $C_{3,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}u_j F_{li}(\tilde{\beta}_l - \beta_l)$. Using $\tilde{\beta}_i - \beta_i = h^2 B_2(x_i) + J_{3i} + (s.o.)$, where $J_{3i} = n^{-1} \sum_{k \neq i} u_k W_{h_{ki}}/f_i$, one can show that

\[
C_{3,1} = \frac{h^2}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}R_{ji}F_{li}B_{2i} + o_p(h^4 + (nh^3)^{-1})
\]

\[
= h^2 E\{E(P_{ji}R_{ji}|x_i)E(F_{li}|x_i)B_{2i}\} + o_p(h^4 + (nh^3)^{-1})
\]

\[
= h^2 E\{E(P_{ji}R_{ji}|x_i)B_{2i}\} + o_p(h^4 + (nh^3)^{-1})
\]

\[
= h^4 E[B_1(x_i)/f_iB_2(x_i)] + o_p(h^4 + (nh^3)^{-1})
\]

\[
= h^4 \int B_1(x)B_2(x)dx + o_p(h^4 + (nh^3)^{-1}),
\]
and
\[
C_{3,2} = \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji} u_j F_{li} J_{3i} + o_p(h^4 + (nh^3)^{-1}) \\
= \frac{1}{n^3} \sum_i \sum_{j \neq i} P_{ji} u_j J_{3i} + o_p(h^4 + (nh^3)^{-1}) + o_p(h^4 + (nh^3)^{-1}) \\
= \frac{1}{n^3} \sum_i \sum_{j \neq i} W_{h,ji}(x_j - x_i)u_j u_i W_{h,ji}/f_i + o_p(h^4 + (nh^3)^{-1}) \\
\]
\[
= \frac{1}{n^3} \sum_i \sum_{j \neq i} W_{h,ji}(x_j - x_i)u_j^2 W_{h,ji}/f_i + o_p(h^4 + (nh^3)^{-1}) \\
= \frac{1}{n} E[W_{h,ji}(x_j - x_i)u_j^2 W_{h,ji}/f_i] + o_p(h^4 + (nh^3)^{-1}) \\
= \frac{1}{n} E\{E[W_{h,ji}(x_j - x_i)\sigma^2(x_j)W_{h,ji}/f_i|x_i]\} + o_p(h^4 + (nh^3)^{-1}) \\
= -\frac{1}{nh^3 \mu_2} E[\sigma^2(x_i)/f_i]\int W(v)W'(v)vdv + o_p(h^4 + (nh^3)^{-1}) \\
= \frac{1}{2nh^3 \mu_2} \int \sigma^2(x)dx + o_p(h^4 + (nh^3)^{-1}),
\]
where we have used \(\int W(v)W'(v)vdv = W(v^2)v^\infty_{v=-\infty} - [\int W(v)W'(v)vdv + \int W(v)^2dv]\), which gives \(\int W(v)W'(v)vdv = -\nu_0/2\).

Summarizing the above we have shown that, we have proved that (by adding back the trimming function \(M(\cdot)\))
\[
C_3 = h^4 \int B_1(x)B_2(x)f(x)M(x)dx + \frac{1}{2nh^3 \mu_2} \int \sigma^2(x)M(x)dx + (s.o.).
\]
This completes the proof of Lemma A.3. \(\square\)

Our final lemma in Appendix A examines the cross term in the \(CV(\cdot)\) function for the vector \(x\) case.

**Lemma A.4.** Let \(C_3 \overset{df}{=} n^{-1} \sum_i (e_\beta^T A_{2i}^{-1} A_{1i})^T e_\beta^T A_{2i} \Delta_i M_i\), where \(M_i = M(x_i)\). Then we have
\[
C_3 = \int \sum_{s=1}^d B_{1s,h}(x)B_{2s,h}(x)M(x)f(x)dx + \frac{\nu_0}{\mu_2} \sum_{s=1}^d \frac{1}{nh_1...h_d} \sum_{s=1}^d h_s^{-2} + o_p(||h||^4 + (nh_1...h_d||h||^2)^{-1}).
\]

Proof: Using partitioned inverse that
\[
\begin{pmatrix}
D_{11} & D_{12} \\
D_{21} & D_{22}
\end{pmatrix}^{-1} = \begin{pmatrix}
F^{-1} & -F^{-1}D_{12}D_{22}^{-1} \\
-D_{22}^{-1}D_{22}F^{-1} & D_{22}^{-1} + D_{22}^{-1}D_{21}F^{-1}D_{12}D_{22}^{-1}
\end{pmatrix}
\]
where \(F = D_{11} - D_{12}D_{22}^{-1}D_{21}\). We apply the above partitioned inverse formula to our \(A_{2i} = \begin{pmatrix}
\bar{f}_i & B_{1i}^T \\
D_{h^2}^{-2}B_{1i} & D_{h^2}^{-1}B_{2i}
\end{pmatrix}\). Straightforward calculations show that
\[
e_\beta^T A_{2i}^{-1} = (-C_{1i}, C_{2i}J_d) + O_p(n_1^{1/2}),
\]
where \( C_{1i} = \beta_i / f_i^2 \) (\( \beta_i = \frac{\partial g(v)}{\partial v} \)|\( v=x_i \)), \( C_{2i} = 1/(\mu_2 f_i) \) and \( \eta_n = \| h \|^4 + (nh_1...h_d\| h \|^2)^{-1} \).

Hence,
\[
e^T \beta A_{2i}^{-1} A_{1i} = (-C_{1i}, C_{2i} I_d) \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( \frac{1}{D_h^{-2}(x_j - x_i)} \right) (R_{ji} + u_j) + (s.o.)
\]
\[
= \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left[ -C_{1i} + D_h^{-2}(x_j - x_i)C_{2i} \right] (R_{ji} + u_j) + (s.o.)
\]
\[
= \frac{1}{n} \sum_{j \neq i} P_{ji}(R_{ji} + u_j) + (s.o.),
\]

where \( P_{ji} = W_{h,ji}[-C_{1i} + D_h^{-2}(x_j - x_i)C_{2i}] \).

Similarly, one can show that
\[
e^T \beta A_{2i}^{-1} \Delta_i = (-C_{1i}, C_{2i} I_d) \frac{1}{n} \sum_{l \neq i} W_{h,li} \left( \frac{1}{D_h^{-2}(x_l - x_i)} \right) [(x_j - x_i)^T(\tilde{\beta}_i - \beta_i) + (\tilde{g}_i - g_i)] + (s.o.)
\]
\[
= (-C_{1i}, C_{2i} I_d) \frac{1}{n} \sum_{l \neq i} W_{h,li} \left( \frac{1}{D_h^{-2}(x_l - x_i)} \right) [(x_l - x_i)^T(\tilde{\beta}_i - \beta_i)] + (s.o.)
\]
\[
= \frac{1}{n} \sum_{l \neq i} W_{h,li} \left[ -C_{1i} + D_h^{-2}(x_l - x_i)C_{2i} \right] (x_j - x_i)^T(\tilde{\beta}_i - \beta_i) + (s.o.)
\]
\[
= \frac{1}{n} \sum_{l \neq i} F_{li}(\tilde{\beta}_i - \beta_i) + (s.o.),
\]

where \( F_{li} = P_{li}(x_l - x_i)^T \).

Using the above results we obtain
\[
C_3 = \frac{1}{n} \sum_i (e^T \beta A_{2i}^{-1} A_{1i})^T e^T \beta A_{2i}^{-1} \Delta_i
\]
\[
= \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}^T(R_{ji} + u_j)F_{li}(\tilde{\beta}_i - \beta_i) + (s.o.)
\]
\[
= \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}^T(R_{ji} + u_j)F_{li}(B_{2,hi} + J_{3i}) + (s.o.)
\]
\[
= \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}^TR_{ji}F_{li}B_{2,hi} + \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}^Tu_jF_{li}J_{3i} + (s.o.)
\]
\[
\equiv P_1 + P_2 + (s.o.), \tag{A.19}
\]

where \( B_{2,hi} \) is a \( d \times 1 \) vector with the \( s^{th} \) component given by \( B_{2,s,hi}(x_i) \) as defined in Theorem 2.2, \( J_{3i} = n^{-1} \sum_{k \neq i} u_k W_{h,ki} / f_i \), \( P_1 = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}^TR_{ji}F_{li}B_{2,hi} \) and \( P_2 = \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji}^Tu_jF_{li}J_{3i} \).

Similar to the proof of Lemma B.1 (ii) and (iii), one can show that \( E(F_{li}|x_i) = 1 + O_p(\| h \|^2) \) and \( E(P_{ji}^TR_{ji}|x_i) = (B_{11,hi}, ..., B_{1d,hi}) \), where \( B_{1s,hi} = B_{1s,hi}(x_i) \) as defined in Theorem 2.2 for
\( s = 1, \ldots, d \). Then it follows that

\[
P_1 = \sum_{s=1}^{d} \int B_{1s,h}(x)B_{2s,h}(x)f(x)dx + \text{(s.o.)},
\]

(A.20)

and

\[
P_2 = \frac{1}{n^3} \sum_{i} \sum_{j \neq i} \sum_{l \neq i} P_{ji}^{T} u_j f_{li} J_{3i}
\]

\[
= \frac{1}{n^2} \sum_{i} \sum_{j \neq i} P_{ji}^{T} u_j J_{3i} + \text{(s.o.)}
\]

\[
= \frac{1}{n^3} \sum_{i} \sum_{j \neq i} \sum_{l \neq i} P_{ji} u_j u_l W_{h,li}^'/f_i + \text{(s.o.)}
\]

\[
= \frac{1}{n^3} \sum_{i} \sum_{j \neq i} P_{ji}^{T} u_j^2 W_{h,ji}^'/f_i + \text{(s.o.)}
\]

\[
= n^{-1} E[P_{ji}^{T} u_j^2 W_{h,ji}^'/f_i] + \text{(s.o.)}
\]

\[
= \nu_0^d \frac{1}{2 \mu_2 n h_1 \ldots h_d} \left( \frac{1}{h_1^2} + \ldots + \frac{1}{h_d^2} \right) + \text{(s.o.)}.
\]

(A.21)

Lemma A.4 follows from (A.19), (A.20) and (A.21) (by also adding back the weight function \( M(\cdot) \)).
Appendix B is for referees’ convenience, not for publication. This appendix is also available to readers from the authors upon request.

**APPENDIX B. SOME USEFUL LEMMAS**

Since we used the U-statistic H-decomposition several times in Appendix, we state it below for readers’ convenience.

**Theorem B.1** (The U-Statistic H-decomposition). Let \( C_n^k = n!/[k!(n-k)!] \) denote the number of combinations obtained by choosing \( k \) items from \( n \) (distinct) items. Then a general \( k^{th} \) order U-statistic \( U(k) \) is defined by

\[
U(k) = \frac{1}{C_n^k} \sum_{1 \leq i_1 < \cdots < i_k} H_n(x_{i_1}, \ldots, x_{i_k}),
\]

where \( H_n(x_{i_1}, \ldots, x_{i_k}) \) is symmetric in its arguments and \( E[H_n^2(x_{i_1}, \ldots, x_{i_k})] < \infty \).

The U-statistic H-decomposition for a general \( k^{th} \) order U-statistic is given as follows

\[
U(k) = \sum_{j=0}^{k} C_n^k H_n^{(j)} = C_n^0 H_n^{(0)} + C_n^1 H_n^{(1)} + C_n^2 H_n^{(2)} + \cdots + C_n^k H_n^{(k)},
\]

where \( H_n^{(0)} = \theta \equiv E[H_n(x_{i_1}, \ldots, x_{i_k})] \), \( H_n^{(1)} = n^{-1} \sum_{i=1}^{n} (H_{ni} - \theta) \) with \( H_{ni} = E[H_n(x_{i_1}, \ldots, x_{i_k})|x_{i}] \), \( H_n^{(2)} = \frac{2}{n(n-1)} \sum_{i,j \neq i} (H_{n,ji} - H_{ni} - H_{nj} + \theta) \) with \( H_{n,ji} = E[H_n(x_{i_1}, \ldots, x_{i_k})|x_{i_1}, x_{i_2}] \), etc., see Lee (1990, page 26) for a detailed derivation of the above H-decomposition.

When \( \theta \neq 0 \) and \( H_n^{(0)} \neq 0 \), usually the partial sum \( H_n^{(1)} \) becomes the leading term of \( H(k) \). If both \( \theta = 0 \) and \( H_n^{(0)} = 0 \), the next non-zero term, \( H_n^{(2)} \), (a second order degenerate U-statistic) usually becomes the leading term for the U-statistic \( U(k) \).

**Lemma B.1.** Recall that \( P_{ji} = W_{h,ji}[C_2 h^{-2}(x_j - x_i) - C_1] \) and \( F_{ji} = (x_j - x_i) P_{ji} \).

\( i ) E(F_{ji}|x_i) = 1 - h^2 G_1(x_i) + O(h^3) \) uniformly in \( x_i \in \mathcal{M} \), where \( G_1(x_i) = \mu_2 f'(|x_i|^2) / f(x_i)^2 \).

\( ii ) E(P_{ji}|x_i) = G_2(x_i) h^2 + O(h^3) \) uniformly in \( x_i \in \mathcal{M} \), where \( G_2(x_i) = \mu_4 h f''(|x_i|^2) + \mu_2 f''(x_i) f'(x_i) \).

\( iii ) E \left( P_{ji} R_{ji}|x_i \right) = h^2 B_1(x_i) + O(h^4) \).

\( iv ) L_{1n} \stackrel{def}{=} n^{-2} \sum_i \sum_{j \neq i} P_{ji}^2 = O_p(h^{-3}) \).

\( v ) L_{2n} \stackrel{def}{=} n^{-2} \sum_i \sum_{j \neq i} F_{ji}^2 = O_p(h^{-1}) \).
**Proof of (i):** It is easy to check that $E[W_{h,ji}(x_j-x_i)^2|x_i] = h^2 \mu_2 f(x_i) + O(h^4)$, $E[(x_j-x_i)W_{h,ji}|x_i] = h^2 \mu_2 f'(x_i) + O_p(h^4)$, $C_1i = f'(x_i)/f(x_i)^2$ and $C_2i = 1/(\mu_2 f(x_i))$. Using these results we get

$$E(F_{ji}|x_i) = E\{W_{h,ji}(x_j-x_i)[C_2i h^{-2}(x_j-x_i) - C_1i]|x_i\}$$

$$= C_2i h^{-2} E[W_{h,ji}(x_j-x_i)^2|x_i] - C_1i E[(x_j-x_i)W_{h,ji}|x_i]$$

$$= 1 - h^2 \mu_2 f'(x_i)^2/f(x_i)^2 + O(h^3)$$

uniformly in $x_i \in \mathcal{M}$.

**Proof of (ii):** By noting that $C_1i = f'(x_i)/f(x_i)^2$ and $C_2i = 1/(\mu_2 f(x_i))$, we have

$$E(W_{h,ji}|x_i) = f(x_i) + \frac{1}{2} h^2 \mu_2 f''(x_i) + O_p(h^4),$$

$$h^{-2} E[(x_j-x_i)W_{h,ji}|x_i] = h^{-2} \left[ \frac{3}{2} h^2 f'(x_i) \mu_2 + 0 + \frac{1}{6} h^4 \mu_4 f''''(x_i) + O(h^6) \right]$$

$$= f'(x_i) \mu_2 + \frac{1}{6} h^2 \mu_4 f''''(x_i) + O(h^4).$$

Hence, we have

$$E(P_{ji}|x_i) = E[W_{h,ji}(C_2i h^{-2}(x_j-x_i) - C_1i)|x_i]$$

$$= C_2i h^{-2} E[W_{h,ji}(x_j-x_i)]|x_i] - C_1i E[W_{h,ji}|x_i]$$

$$= f'(x_i) + \frac{h^2 \mu_4 f''''(x_i)}{6 \mu_2} - \frac{f'(x_i)}{2} + f'(x_i) + \frac{h^2 \mu_2 f''''(x_i)f'(x_i)}{2} + O_p(h^3)$$

$$= G_2(x_i) h^2 + O(h^3) \quad \text{(B.3)}$$

uniformly in $x_i \in \mathcal{M}$ because the two leading terms cancel each other.

**Proof of (iii):** It is easy to show that $L_{1n} \sim L_{1n,1}$, where $L_{1n,1}$ is obtained from $L_{1n}$ by removing the $C_{1i}$ term because $C_{1i}$ is dominated by $h^{-2} C_{2i}(x_j-x_i)$, i.e., $L_{1n,1} = n^{-2} \sum_i \sum_{j \neq i} W_{h,ji}^2 h^{-4} (x_j-x_i)^2$ Also, since $\sup_{x_i \in \mathcal{M}} C_{2i}^2 \leq C$, where $C > 0$ is a finite positive constant, we only need to evaluate the order of $J_{1n,2} = n^{-2} \sum_i \sum_{j \neq i} W_{h,ji}^2 h^{-4} (x_j-x_i)^2$. By the U-statistic H-decomposition, it is easy to check that $J_{1n,2} = E(J_{1n,2}) + (s.o.) = h^{-3} [\int W^2(v) v^2 dv] E[f(x_i)] + O_p(h^{-3}) = O_p(h^{-3})$.

**Proof of (vi):** It is easy to see that $J_{2n} = J_{2n,1} + (s.o.)$, where $J_{2n,1} = n^{-2} h^{-4} \sum_i \sum_{j \neq i} W_{h,ji}^2 (x_j-x_i)^4 C_{2i}^2$ because $C_{1j}$ is dominated by $h^{-2} C_{2i}(x_j-x_i)$. Also, since $C_{2i}^2$ is bounded on $\mathcal{M}$, we can replace $C_{2i}$ by 1 and evaluate the order of $J_{4n,2} = n^{-2} h^{-4} \sum_i \sum_{j \neq i} W_{h,ji}^4 (x_j-x_i)^4$. It is easy to see that $J_{2n,2} = E(J_{2n,2}) + (s.o.) = h^{-1} [\int W^2(v) v^4 dv] E[f(x_i)] + o_p(h^{-1}) = O_p(h^{-1})$. This completes the proof of Lemma B.1.
Below we provide asymptotic result for \( \hat{\beta}(x) \), the local constant of \( \beta(x) \). The local constant estimate of \( g(x) \) is given by
\[
\hat{g}(x) = \frac{n^{-1} \sum_j y_j W_{h,jx}}{n^{-1} \sum_j W_{h,jx}} ,
\]
where \( W_{h,jx} = h^{-1} W((x_j - x)/h), \) \( y_j = g_j + u_j, \) \( g_j = g(x_j) \).

For \( W(x_j-x)/h \), let \( u = (x_j-x)/h \), we have
\[
W'_{h,jx} = \frac{\partial W_h((x_j-x)/h)}{\partial x} = \frac{1}{h} \frac{\partial W(u)}{\partial u} \frac{x_j-x}{h} = -\frac{1}{h^2} W'(x_j-x/h), \tag{B.4}
\]
where \( W'(x_j-x/h) = W'(u)|_{u=(x_j-x)/h} \) and \( W'(u) = \partial W(u)/\partial u \).

Using (B.4) we obtain the local constant first derivative estimator of \( \beta(x) \) given by
\[
\hat{\beta}(x) = \hat{g}(x) + \frac{1}{n^2} \sum_{j \neq l} (g_j - g_l) W'_{h,jx} W_{h,tx}/f(x)^2 ,
\]
\[
\hat{J}_1(x) = \frac{1}{n^2} \sum_{j \neq l} (g_j - g_l) W'_{h,jx} W_{h,tx}/f(x)^2 ,
\]
\[
\hat{J}_2(x) = \frac{1}{n^2} \sum_{j \neq l} (u_j - u_l) W'_{h,jx} W_{h,tx}/f(x)^2 .
\]

**Lemma B.2.** (i) \( \hat{\beta}(x) = \beta(x) + h^2 B_2(x) + J_3(x) + O_p(h^3 + (nh^2)^{-1/2}) \), where \( B_2(x) = \mu_2[g''(x) f''(x)f(x) + 2g'(x)f''(x)f(x) + (1/2)g'''(x)f'(x)^2 - g''(x)(f'(x))^2]/f(x)^2 \) and \( J_3(x) = n^{-1} \sum_j u_j W'_{h,jx}/f(x) \).

(ii) \( \hat{J}_1(x) = \beta(x) + h^2 B_2i + J_{3i} + O_p(h^3 + (nh^2)^{-1/2}) \), \( B_2i = B_2(x_i) \) and \( J_{3i} = n^{-1} \sum_{j \neq i} u_j W'_{h,ji}/f_i \).

**Proof of Lemma B.2:** We will only prove (i) as (ii) follows by the same arguments. We first evaluate \( E[J_1(x)] \). Using \( W'_{h,ji} = -h^{-2} W'(x_j-x/h) \), we get
\[
E[J_1(x)] = \frac{1}{h^3 f(x)^2} \int \int [g(x_l) - g(x_j)] W(x_j-x/h) W'(x_j-x/h) f(x) f(x_l) dx_j dx_l
\]
\[
= \frac{1}{h f(x)^2} \int \int [g(x + hv) - g(x + huv)] W'(u) W(v) f(x + hv) f(x + hv) du dv
\]
\[
= \frac{1}{h f(x)^2} \int \int [(g'(x)hv + \frac{1}{2} g''(x)h^2 v^2 - g'(x)hu - \frac{1}{2} g''(x)h^2 u^2 - \frac{1}{6} g'''(x)h^3 u^3]
\]
\[
[ f(x) + f'(x)hu + \frac{1}{2} f''(x)h^2 u^2 ] [ f(x) + f'(x)hv + \frac{1}{2} f''(x)h^2 v^2 ] W'(u) W(v) du dv + O(h^3)
\]
\[
= g'(x) + B_2(x) h^2 + O(h^3),
\]
where $B_2(x)$ is defined in the beginning of lemma B.2. An easy to verify the expression of $B_2(x)$ is to realize that $\int \int W'(u)W(v)u^{p}v^{q} = 0$ if $p$ is an even integer, or $q$ is an odd integer, so the only non-zero terms coming from $p$ is odd and $q$ is even. Also note that by integration by parts, $\int uW'(u)du = uW(u)|_{-\infty}^{\infty} - \int W(u)du = -1$ and $\int u^{3}W(u)du = u^{3}W'(u)|_{-\infty}^{\infty} - 3 \int u^{2}W(u)du = -3\mu_2$.

It is easy to show that $Var(J_1(x)) = O((nh^2)^{-1})$. Hence, we have shown that

$$J_1(x) = g'(x) + h^2B_2(x) + O_p(h^3 + (nh^2)^{-1/2}). \tag{B.6}$$

Next, we consider $J_2(x)$. Define $J_{2,1} = J_2(x)f(x)^2$. Then $J_{2,1}(x) = (2/n^2)\sum\sum_{l>j}(1/2)[(u_j - u_l)W_{h,jx}W_{h,tx} + (u_l - u_j)W_{h,tx}W_{h,jx}] = (2/n^2)\sum\sum_{l\not=j}H_n(z_j, z_l)$, which is a second-order U-statistic, where $H_n(z_j, z_l) = (1/2)[(u_j - u_l)W_{h,jx}W_{h,tx} + (u_l - u_j)W_{h,tx}W_{h,jx}]$ with $z_j = (u_j, x_j)$.

By the U-statistic II-decomposition we know that the leading term of $J_{2,1}(x)$ is $\frac{2}{n}\sum_j H_{nj},$ where

$$H_{nj} = E[H_n(z_j, z_l)|z_l]$$

$$= (1/2)u_j[W_{h,jx}E(W_{h,tx}) - W_{h,tx}E(W_{h,jx})]$$

$$= (1/2)u_j\{W_{h,jx}[f(x) + O_p(h^2)] - [W_{h,tx}\mu_2f'(x) + O_p(h^2)]\}$$

$$= (1/2)u_jW_{h,jx}f(x) + (s.o.)$$

the last equality follows because $W_{h,jx} = -h^{-2}W(\frac{x_j - x}{h})$ has an order larger (by a factor of $h^{-1}$) than that of $W_{h,tx} = h^{-1}W(\frac{x_j - x}{h})$.

Hence, the leading term of $J_2(x)$ is the leading term of $J_{2,1}(x)/f(x)^2$, which is a partial sum given by

$$J_2(x) = \frac{2}{nf(x)^2}\sum_j H_{nj} + (s.o.) = \frac{1}{nf(x)}\sum_j u_jW_{h,jx} + (s.o.) = J_3(x) + (s.o.),$$

where $J_3(x) = \frac{1}{nf(x)}\sum_j u_jW_{h,jx}$.

Next, we show that $J_3(x) = O_p((nh^3)^{-1/2})$. Obviously, $J_3(x)$ has zero mean, its variance is given by

$$E[J_3(x)^2] = \frac{1}{nf(x)^2}E[\sigma_j^2W_{h,jx}^2]$$

$$= \frac{1}{nh^4f(x)^2}\int \sigma_j^2(x_j)\left[W'(\frac{x_j - x}{h})\right]^2f(x_j)dx_j$$

$$= \frac{1}{nh^4f(x)^2}h\int \sigma_j^2(x + hv)[W'(v)]^2f(x + hv)dv$$

$$= \frac{\kappa_0}{nh^3f(x)}\sigma_j^2(x) + O((nh)^{-1}),$$

where $\kappa_0 = \int [W'(v)]^2dv$ and $\sigma_j^2 = E(u_j^2|x_j)$. Hence, $J_3(x) = O_p((nh^3)^{-1/2})$.  


Applying Lyapunov’s central limit theorem, we have the following asymptotic distribution result for $\hat{\beta}(x)$ (with $h \to 0$, $nh^3 \to \infty$ and $nh^9 \to 0$ as $n \to \infty$):

$$\sqrt{nh^3}[\hat{\beta}(x) - \beta(x) - h^2B_2(x)] \xrightarrow{d} N\left(0, \frac{\kappa_0\sigma^2(x)}{f(x)}\right).$$

\[\square\]

Lemma B.3. Let $G(x_i)$ be a bounded function on $\mathcal{M}$. Define $\mathcal{D} = n^{-1} \sum_{i=1}^{n} G(x_i)\tilde{m}_i^2$, where $\tilde{m}_i = (\bar{g}_i - g_i)f_i$. Then

$$\mathcal{D} = h^4E[G(x_i)C(x_i)^2] + \frac{1}{nh}n_0E[G(x_i)\sigma^2(x_i)f(x_i)] + (s.o.),$$

where $C(x_i) = \frac{1}{2}\mu_2[g''(x_i)f(x_i) + 2g'(x_i)f'(x_i)]$, $\mu_2 = \int v^2W(v)dv$ and $n_0 = \int W^2(v)dv$.

Proof of Lemma B.3: Using (A.14), we have

$$\mathcal{D} = \frac{1}{n^3} \sum_{i} \sum_{j \neq i} \sum_{l \neq i,j} G(x_i)(g_j - g_i + u_j)W_{h,ji}(g_l - g_i + u_j)W_{h,li}$$

$$= \frac{1}{n^3} \sum_{i} \sum_{j \neq i} G(x_i)(g_j - g_i + u_j)^2W_{h,ji}^2$$

$$+ \frac{1}{n^3} \sum_{i} \sum_{j \neq i} \sum_{l \neq i,j} G(x_i)(g_j - g_i + u_j)W_{h,ji}(g_l - g_i + u_j)W_{h,li}$$

$$= \mathcal{D}_1 + \mathcal{D}_2,$$

where the definitions of $\mathcal{D}_1$ should be apparent ($l = 1, 2$).

It is easy to see that the leading term of $\mathcal{D}_1$ is associated with $u_j^2$. So we have

$$\mathcal{D}_1 = \frac{1}{n^3} \sum_{i} \sum_{j \neq i} G(x_i)u_j^2W_{h,ji}^2 + (s.o.)$$

$$= n^{-1}E[G(x_i)\sigma^2(x_j)W_{h,ji}^2] + (s.o.)$$

$$= n^{-1}h^{-1}n_0 \left[ \int G(x_i)\sigma^2(x_i)f(x_i)dx_i \right] + (s.o.)$$

$$= \frac{1}{nh}n_0E[G(x_i)\sigma^2(x_i)f(x_i)] + (s.o.),$$

where the second equality follows from the U-statistic H-decomposition, and $n_0 = \int W^2(v)dv$.

It can be shown that the leading term of $\mathcal{D}_2$ is associated with $(g_j - g_i)(g_l - g_i)$, i.e., $\mathcal{D}_2 = \mathcal{D}_2^0 + (s.o.)$, where

$$\mathcal{D}_2^0 = \frac{1}{n^3} \sum_{i} \sum_{j \neq i} \sum_{l \neq i,j} G(x_i)(g_j - g_i)W_{h,ji}(g_l - g_i)W_{h,li}$$

$$= E\{G(x_i)(g_j - g_i)W_{h,ji}(g_l - g_i)W_{h,li}\} + (s.o.)$$

$$= E\{G(x_i)[E((g_j - g_i)W_{h,ji}|x_i)]^2\}(s.o.)$$

$$= h^4E[G(x_i)C(x_i)^2] + (s.o.),$$
where the second equality follows from the U-statistic H-decomposition, and $C(x_i) = \frac{1}{2} \mu_2 [g''(x_i) f(x_i) + 2g'(x_i) f'(x_i)]$, and $\mu_2 = \int v^2 \mathbb{W}(v) dv$.

\[ \] \[ \]

Lemma B.4. Let $D_{4,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i,j} e_{ji} e_{li} \tilde{m}_i^2 / f_i^2$ as defined in lemma A.2. Then $D_{4,2} = O_p((nh)^{-2}) = o_p((nh)^{-1})$.

Proof of Lemma B.4: Recall that $e_{ji} = P_{ji} - E(P_{ji} | x_i)$, $D_{4,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i,j} e_{ji} e_{li} \tilde{m}_i^2$. Using (A.14) we have

\[
D_{4,2} = \sum_i \sum_{j \neq i} \sum_{l \neq i,j} \sum_{k \neq i} \sum_{m \neq i} e_{ji} e_{li} W_{h,ki} W_{h,mi} [(g_k - g_l)(g_m - g_l) + u_k u_m + 2u_k (g_m - g_l)] / f_i^2
\]

\[
= D_3 + D_4 + 2D_5,
\]

(B.7)

where $D_3 = \sum_i \sum_{j \neq i} \sum_{l \neq i,j} \sum_{k \neq i} \sum_{m \neq i} e_{ji} e_{li} W_{h,ki} W_{h,mi} (g_k - g_l)(g_m - g_l) / f_i^2$, $D_4 = \sum_i \sum_{j \neq i} \sum_{l \neq i,j} \sum_{m \neq i} e_{ji} e_{li} W_{h,ki} W_{h,mi} u_k u_m / f_i^2$ and $D_5 = \sum_i \sum_{j \neq i} \sum_{l \neq i,j} \sum_{k \neq i} \sum_{m \neq i} e_{ji} e_{li} W_{h,ki} W_{h,mi} u_k (g_m - g_l) / f_i^2$.

We consider $D_3$. We first evaluate $E(D_3)$. Since $i \neq j \neq l$, and $E(e_{ji} | x_i) = 0$ and $E(e_{li} | x_i) = 0$, it is easy to see that if $j$ or $l$ differ from $k$ and $m$, then $E(D_3) = 0$. Hence, for $E(D_3) \neq 0$, we must have $k$ and $m$ to match $j$ and $l$, say $k = j$ and $m = l$. Therefore, the leading term of $E(D_3)$ corresponds to the case these five indices take at most three different values, say $i \neq j \neq l$ with $k = j$ and $m = l$. Since we only need to evaluate the order of $D_3$, it is sufficient to consider $e_{ji} \sim P_{ji} \sim h^{-2} W_{h,ji}(x_j - x_i)$. Then, we have

\[
E[e_{ji} W_{h,ji}(g_j - g_l) | x_i] \sim h^{-2} E[W_{h,ji}^2 (x_j - x_i)(g_j - g_l) | x_i]
\]

\[
= h^{-2} \int h^{-2} W^2(v) hv [g'_h h v + O(h^2)] [f_i + O(h)] h dv
\]

\[
= h^{-2} v_2 g'_h f_i + O(1) = O(h^{-2})
\]

uniformly in $x_i \in \mathcal{M}$.

Hence, $E[e_{ji} e_{li} W_{h,ji}(g_j - g_l) W_{h,li}(g_l - g_i)] = E\{E(e_{ji} W_{h,ji}(g_j - g_l) | x_i) E(e_{li} W_{h,li}(g_l - g_i) | x_i)\} = O(h^{-4})$. This leads to

\[
E(D_3) = n^{-5} n^3 O(h^{-4}) = O((nh)^{-2}).
\]

Similar arguments lead to $Var(D_3) = O((nh)^{-4})$. Hence,

\[ D_3 = O_p((nh)^{-1}). \]

Next, we consider $D_4$. It is easy to see that the leading term of $D_4$ corresponds to the case that the summation indices $i, j, l, k, m$ all different from each other. We use $D_{4,1}$ to denote this leading term of $D_4$. Obviously, $E(D_{4,1}) = 0$. By noting that $E(e_{ji}^2) = O(h^{-3})$ and $E(W_{h,ji}^2) = O(h^{-1})$, it is straightforward to show that $E(D_{4,1}^2) = n^{-8} n^4 O(h^{-8}) = O((nh)^{-4})$. Hence, $D_{4,1} = O_p((nh)^{-2})$.

Similarly, one can show that $D_5 = O_p((nh)^{-2})$.

\[ \]

Lemma B.5. Let $D_{6,2}$ and $D_{6,3}$ be as defined in lemma A.2. Then we have (i) $D_{6,2} = o_p(n^{-1}) = o_p((nh)^{-1})$; (ii) $D_{6,3} = o_p(h^4 + (nh)^{-1})$.\]
**Proof of Lemma (i):** Using $\tilde{\beta}_i - \beta_i = h^2 B_{2i} + J_{3i} + (s.o.)$, we have

\[
D_{6,2} = n^{-3}\sum_i \sum_{j \neq i} \sum_{l \neq i,j} \eta_{ji} \eta_{li} (\tilde{\beta}_i - \beta_i)^2 \\
= n^{-3}\sum_i \sum_{j \neq i} \sum_{l \neq i,j} \eta_{ji} \eta_{li} (h^4 B_{2i}^2 + J_{3i}^2 + 2h^2 B_{2i} J_{3i}) + (s.o.) \\
= D_6 + D_7 + 2D_8 + (s.o.),
\]

where $D_6 = (h^4/n^3) \sum_i \sum_{j \neq i} \sum_{l \neq i,j} \eta_{ji} \eta_{li} B_{2i}^2$. From $E(\eta_{ji}|x_i) = 0$ and $E(\eta_{ji}^2) \sim E(F_{ji}^2) \sim h^{-4}E[W_{h,ji}(x_j-x_i)^4] = h^{-4}O(h^3) = O(h^{-1})$, it is easy to show that

\[
E(D_6^2) = (h^8/n^6)O(n^4 h^{-2}) = O(h^6/n^2).
\]

Hence, $D_6 = O_p(h^3/n) = o_p(n^{-1})$.

\[
D_7 = n^{-5}\sum_i \sum_{j \neq i} \sum_{l \neq i,j} \sum_{k \neq i} \sum_{m \neq i} \eta_{ji} \eta_{li} u_k u_m W_{h,ki} W_{h,mi} / f_i^2.
\]

Using $E(\eta_{ji}|x_i) = 0$, $E(u_k|x_k) = 0$, $E(\eta_{ji}^2) \sim E(F_{ji}^2) = O(h^{-1})$ and $E((W_{h,ki})^2) = h^{-4}O(h) = O(h^{-3})$, we have

\[
E(D_7^2) = n^{-10}O(n^6 h^{-8}) = O((nh^2)^{-4}n^{-2}) = o(n^{-2}).
\]

Hence, $D_7 = o_p(n^{-1})$.

Similarly, one can show that $D_8 = o_p(n^{-1})$. Combining the above results we obtain

\[
D_{6,2} = D_6 + D_7 + 2D_8 + (s.o.) = o_p(n^{-1}).
\]

This completes the proof of (i). \[\square\]

**Proof of Lemma (ii):** Using $E(F_{ji}|x_i) = 1 + O(h^2)$ and $\tilde{\beta}_i - \beta_i = h^2 B_{2i} + J_{3i} + (s.o.)$, we have

\[
D_{6,3} \sim n^{-3}\sum_i \sum_{j \neq i} \sum_{l \neq i,j} \eta_{li} (\tilde{\beta}_i - \beta_i)^2 \\
\sim n^{-3}\sum_i \sum_{j \neq i} \sum_{l \neq i,j} \eta_{li} [h^4 B_{2i}^2 + J_{3i}^2 + 2h^2 B_{2i} J_{3i}] \\
= D_9 + D_{10} + D_{11} + (s.o.),
\]

where $D_9 = (h^4/n^3) \sum_i \sum_{j \neq i} \sum_{l \neq i,j} \eta_{li} B_{2i}^2$. Its second moment is

\[
E(D_9^2) \sim (h^8/n^6)O(E(\eta_{ji}^2)) = (h^8/n)O(h^{-1}) = O(h^7/n) = o(h^8).
\]

Hence, $D_9 = o_p(h^4)$.

\[
D_{10} = n^{-5}\sum_i \sum_{j \neq i} \sum_{l \neq i,j} \sum_{k \neq i} \sum_{m \neq i} \eta_{ji} u_k u_m W_{h,ki} W_{h,mi} / f_i^2.
\]

From $E(\eta_{ji}^2) = O(h^{-1})$ and $E((W_{h,ki})^2) = O(h^{-3})$, we get

\[
E(D_{10}^2) = n^{-10}O(h^{-7}) = O((nh^3)^{-2}(nh)^{-1}) = o((nh^3)^{-2}).
\]

Hence, $E(D_{10}) = o_p((nh^3)^{-1})$.

Similarly, one can show that $D_{11} = o_p(h^4 + (nh^3)^{-1})$. 

Summarizing the above we have shown that
\[ D_{0,3} = D_0 + D_{10} + D_{11} + (s.o.) = o_p(h^4 + (nh^3)^{-1}). \]

The next two lemmas give the asymptotic biases, variances and distributions of the local linear and local constant derivative estimators for the case when \( x \) is a \( d \times 1 \) vector.

**Lemma B.6.** Define a \( d \times d \) diagonal matrix \( D_h = \text{diag}(h_s) \). Then under the conditions of Theorem 2.2 we have

\[
\sqrt{n} h_{11} \ldots h_{dd} D_h (\hat{\beta}(x) - \beta(x) - B_{1,h}(x)) \xrightarrow{d} N(0, c_{ll}(x) I_d),
\]

where \( B_{1,h}(x) \) is a \( d \times 1 \) vector with its \( s^{th} \) component given by \( B_{1,s,h}(x) = \frac{h_s^2 \mu_4 g''(x)}{6\mu_2} + \frac{h_s^2 g''(x) f_s(x) \mu_4}{2\mu_2 f(x)} - \mu_2 f_s(x) \sum_{i=1}^d g_n(x) h_i^2 f(x) \), \( c_{ll}(x) = \nu_0 d^{-1} \nu_2 \sigma^2(x) / (\mu_2^2 f(x)) \) and \( I_d \) is \( n \times n \) identity matrix.

**Proof of Lemma B.6:** By (14) we know that \( \hat{\beta}(x) = \beta(x) + e_h A_{2x}^{-1} A_{1x} \), where \( A_{1x} \) and \( A_{2x} \) are obtained from \( A_{1i} \) and \( A_{2i} \) with \( x_t \) replaced by \( x \), and \( \sum_{j \neq i} \sum_{j'=i} \sum_{j''=i} (A_{1x}) \) replaced by \( \sum_j = \sum_{j=1}^d \). By the same derivation that leads to (A.8), we have \( e_h A_{2x}^{-1} A_{1x} = n^{-1} \sum_j W_{h,jx} [\frac{1}{\mu_2 f(x)} D_h^{-2}(x_j - x) - f'(x)/f(x)^2] (R_{jx} + u_j) + (s.o.) = A_{1n}(x) + A_{2n}(x) \), where \( D_h^{-2} = \text{diag}(h_s^{-2}) \), \( R_{jx} = g(x_j) - g(x) - (g'(x))' (x_j - x) \), \( A_{1n}(x) = n^{-1} \sum_j W_{h,jx} [\frac{1}{\mu_2 f(x)} D_h^{-2}(x_j - x) - f'(x)/f(x)^2] u_j \). Let \( A_{1n,s}(x) \) be the \( s^{th} \) component of \( A_{1n}(x) \). Note that the \( s^{th} \) component of \( D_h^{-2}(x_j - x) - f'(x)/f(x)^2 \) is \( h_s^{-2}(x_{js} - x_s) - f_s(x)/f(x)^2 \), where \( f_s(x) = \partial f(x)/\partial x_s \).

We have

\[
E(A_{1n,s}(x)) = \int W_{h,jx} \left[ \left( \frac{1}{\mu_2 f(x)} \right) h_s^{-2}(x_{js} - x_s) - \frac{f_s(x)}{(f(x))^2} \right] R_{jx} f(x_j) dx_j
\]

\[
= \int W(v) \left[ \left( \frac{1}{\mu_2 f(x)} \right) \frac{v_s}{h_s} - \frac{f_s(x)}{f(x)^2} \right] \left[ \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^d g_{1i,j}(x_i) h_i h_j v_i v_j \right]
\]

\[
+ \frac{1}{6} \sum_{i=1}^d \sum_{j=1}^d \sum_{l=1}^d g_{1i,j,l}(x_i) h_i h_j v_i v_j v_l + O(||h||^4)
\]

\[
\left[ f(x) + \sum_{i=1}^d f_i(x) h_i v_i + O(||h||^2) \right]
\]

\[
= \frac{h_s^2 \mu_4 g_{sss}(x)}{6\mu_2} + \frac{h_s^2 g_{ss}(x) f_s(x) \mu_4}{2\mu_2 f(x)} - \mu_2 f_s(x) \sum_{i=1}^d g_u(x) h_i^2 f(x)
\]

\[
+ \frac{1}{2} \frac{f_s(x) \sum_{t \neq s} h_s^2 g_{tt}(x)}{2f(x)} + \mu_2 \sum_{t \neq s} h_s^2 f_t(x) g_{ts}(x)/f(x) + (\mu_2/2) \sum_{t \neq s} h_s^2 g_{tls}(x) + O(||h||^3)
\]

\[
= B_{1,s,h}(x) + O(||h||^3).
\]
Proof of Lemma B.7: We use Lemma B.7. Under the conditions of Theorem 2.2 we have

\[ \mathcal{A}_n(x) = B_{1,h}(x) + O_p(||h||^3 + (nh_1...h_d||h||)^{-1/2}). \]

Next, \( E(\mathcal{A}_n(x)) = 0 \) and

\[
\operatorname{Var}(\mathcal{A}_n(x)) = \frac{1}{n\mu_2 f(x)^2} \int f(x)j \sigma^2(x) jW_{h,jx}^2 D_h^{-2}(x_j - x_j)(x_j - x)^T D_h^{-2} dx_j
\]

\[
= \frac{1}{nh_1...h_d \mu_2 f(x)^2} \int (f \sigma^2)(x + hv)W(h)v^2 D_h^{-2} D_h vv^T D_h D_h^{-2} dv
\]

\[
= \frac{\nu_0^{-d} \nu_2 \sigma^2(x)}{nh_1...h_d \mu_2 f(x)} D_h^{-2} + O((nh_1...h_d)^{-1}),
\]

where in the last equality we used \( \int W(v)^2 vv^T dv = \nu_0^{-d} \nu_2 I_d \).

Summarizing the above results and under the condition \( nh_1...h_d||h||^2 \to \infty, nh_1...h_d||h||^4 \to 0 \), and by applying Lyapunov’s central limit theorem, we have

\[ (nh_1...h_d)^{1/2} D_h (\bar{\beta}(x) - \beta(x) - B_{1,h}(x)) \xrightarrow{d} N(0, c_{0,l}(x) I_d), \]

where \( c_{0,l}(x) = \frac{\nu_0^{-d} \nu_2 \sigma^2(x)}{\mu_2^2 f(x)} \) and \( D_h = \text{diag}(h_s) \).

Lemma B.7. Under the conditions of Theorem 2.2 we have

\[ \sqrt{nh_1...h_d} D_h (\bar{\beta}(x) - \beta(x) - B_{2,h}(x)) \xrightarrow{d} N(0, c_{l}(x) I_d), \]

where \( B_{2,h}(x) \) is a \( d \times 1 \) vector with its \( s^{th} \) element given by

\[ B_{2,s,h}(x) = \frac{\mu_2}{f(x)} \left[ \frac{1}{2} g_{s,s,s}(x) f(x) h_s^2 + \sum_{t=1}^{d} \left[ f_t(x) g_{t,s}(x) + g_t(x) f_{t,s}(x) \right] h_t^2 - \frac{f_s(x)}{f(x)} \sum_{t=1}^{d} g_t(x) f_t(x) h_t^2 \right. \]

\[ + g_s(x) \sum_{t=1}^{d} f_{t,t}(x) h_t^2 \] \text{ with } \( m_t(x) = \frac{\partial m(x)}{\partial x_s}, m_{ts}(x) = \frac{\partial^2 m(x)}{\partial x_t \partial x_s}, m_{sss}(x) = \frac{\partial^3 m(x)}{\partial x_t^2 \partial x_s} \text{ for } m = g \text{ or } f, \]

\[ c_{l}(x) = \nu_0^{-d} \sum_{j=1}^{d} \beta^2(x) / f(x). \]

Proof of Lemma B.7: For the multivariate \( x \) case, \( \bar{\beta}(x) \) is still defined as in (B.5) with \( J_1(x) \) and \( J_2(x) \) defined below (B.5), except that now \( W_{h,tx} = \prod_{s=1}^{d} h_s^{-1} w_x(x_j-x_s) \) is a product kernel, and \( W_{h,t,jx} \) is a \( d \times 1 \) vector with its \( s^{th} \) component given by \( W_{h,t,jx,s} = -h_s^{-2} w_x(x_j-x_s) \prod_{t \neq s} h_t^{-1} w(x_t-x_s). \)

We use \( J_{1,s}(x) \) to denote the \( s^{th} \) component of \( J_1(x) \), then it is straightforward, though tedious, to show that \( (g_s(x) = \frac{\partial g(x)}{\partial x_s}). \)

\[
E(J_{1,s}(x)) = g_s(x) + \frac{\mu_2}{f(x)} \left[ \frac{1}{2} g_{s,s,s}(x) f(x) h_s^2 + \sum_{t=1}^{d} \left( f_t(x) g_{t,s}(x) + g_t(x) f_{t,s}(x) \right) h_t^2 \right. \]

\[ - \left. \frac{f_s(x)}{f(x)} \sum_{t=1}^{d} g_t(x) f_t(x) h_t^2 + g_s(x) \sum_{t=1}^{d} f_{t,t}(x) h_t^2 \right] + O(||h||^3)) \]

\[ = g_s(x) + B_{2,s,h}(x) + O(||h||^3). \]

Also, it is easy to show that \( \operatorname{Var}(J_1(x)) = O((nh_1...h_d||h||)^{-1}). \) Hence, by noting that \( \beta(x) = (g_1(x), ..., g_d(x))^T \), we have

\[ J_1(x) = \beta(x) + B_{2,h}(x) + O_p(||h||^3 + (nh_1...h_d||h||)^{1/2}). \]
It is easy to see that\[ W'_{h,jx} = -\frac{1}{h_1 \cdots h_d} D_h^{-1} W'(\frac{x_j - x}{h}) \] where\[ W'(\frac{x_j - x}{h}) \] is a \( d \times 1 \) vector with its \( s^{th} \) component given by\[ w'(\frac{x_j - x}{h}) \prod_{t \neq s} w'\left(\frac{x_t - x}{h}\right) \]. Similarly, define \( W'(v) \) as a \( d \times 1 \) vector with its \( s^{th} \) element given by\[ w'(v_s) \prod_{t \neq s} w(v_t) \]. Then it is easy to check that

\[
\int W'(v)(W'(v))^T dv = \nu_0^{d-1} \kappa_0 I_d, \tag{B.9}
\]

where \( \nu_0 = \int w(v_t)^2 dv_t \), \( \kappa_0 = \int (w'(v_s))^2 dv_s \), and \( I_d \) is a \( d \times d \) identity matrix.

Similar to the proof of lemma B.2, by H-decomposition one can show that \( J_2(x) = J_3(x) + (s.o.) \), where \( J_3(x) = \frac{1}{n f(x)} \sum_j u_j W'_{h,jx} \). Obviously, \( E(J_3(x)) = 0 \), and

\[
\text{Var}(J_3(x)) = \frac{1}{n f(x)^2} E[\sigma_j^2(W'_{h,jx})(W'_{h,jx})^T] \\
= \frac{1}{n f(x)^2} \int f(x_j) \sigma_j^2(x_j)(W'_{h,jx})(W'_{h,jx})^T dx_j \\
= \frac{1}{n(h_1 \cdots h_d)^2 f(x)^2} \int f(x_j) \sigma_j^2(x_j) D_h^{-1} W'(\frac{x_j - x}{h}) (W'(\frac{x_j - x}{h}))^T D_h^{-1} dx_j \\
= \frac{1}{n h_1 \cdots h_d f(x)^2} \int (f \sigma^2(x) + h \nu)(D_h^{-1} W'(v))(D_h^{-1} W'(v))^T dv \\
= \frac{\nu_0^{d-1} \kappa_0 \sigma^2(x)}{n h_1 \cdots h_d f(x)^2} D_h^{-2} + O_p((n h_1 \cdots h_d)^{-1}),
\]

where in the last equality we used (B.9).

Under the condition that \( nh_1 \cdots h_d \|h\|^2 \to \infty \), \( nh_1 \cdots h_d \|h\|^4 \to 0 \) as \( n \to \infty \), and by Lyapunov's central limit theorem, we have that

\[
(nh_1 \cdots h_d)^{1/2} D_h(\tilde{\beta}(x) - \beta(x) - B_{2,h}(x)) \stackrel{d}{\to} N(0, c_{lc}(x) I_d),
\]

where \( B_{2,h}(x) \) and \( c_{lc}(x) \) are defined in the beginning of lemma B.7. This completes the proof of lemma B.7.