

# GRADIENT BASED SMOOTHING PARAMETER SELECTION FOR NONPARAMETRIC REGRESSION ESTIMATION\*

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

ABSTRACT. Data-driven bandwidth selection based on the gradient of an unknown regression function is considered. Uncovering gradients nonparametrically is of crucial importance across a broad range of economic environments such as determining risk premium or recovering distributions of individual preferences. The procedure developed here is shown to deliver bandwidths which have the optimal rate of convergence for the estimation of gradients. We provide a detailed theoretical account of this new approach to smoothing parameter selection. An important additional advantage of our proposed method over the conventional cross-validation bandwidth selection method is that our approach overcomes the tendency of traditional data-driven approaches to engage in under smoothing. Both simulated and (several) empirical examples showcase the finite sample attraction of this new mechanism.

## 1. OVERVIEW

The success of nonparametric estimation methods 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, it is not entirely clear that the methods developed for recovering optimal smoothness levels are the proper surrogates when interest instead hinges on the derivative of the unknown function or density. Applications include the characterization of submicroscopic nanoparticles (Charnigo, Francoeur, Kenkel, Mengüç, Hall & Srinivasan 2007), inferring the chemical makeup of a sample of unknown composition (Charnigo, Hall & Srinivasan 2011), estimation of marginal willingness to pay within a two-stage hedonic regression (Bajari & Kahn 2005, Heckman, Matzkin & Nesheim 2010) analysis of human growth data (Ramsay & Silverman 2005) and investigating simple turbulent flows in fluid physics (Overholt & Pope 1996).

The importance of appropriate smoothness selections for higher order derivatives was illustrated by Wahba & Wang (1990) who showed in the smoothing spline setting that the ideal smoothing

---

STATE UNIVERSITY OF NEW YORK AT BINGHAMTON, TEXAS A&M UNIVERSITY, UNIVERSITY OF MIAMI

*Date:* October 12, 2011.

*Key words and phrases.* Gradient Estimation, Kernel Smoothing, Least Squares Cross Validation.

Daniel J. Henderson, Department of Economics, State University of New York at Binghamton; e-mail: djhender@binghamton.edu. Qi Li, Corresponding Author, Department of Economics, Texas A&M University; e-mail: qi@econmail.tamu.edu. Christopher F. Parmeter, Department of Economics, University of Miami; e-mail: cparmeter@bus.miami.edu.

\*The authors are grateful for comments made in seminars at Columbia University, London School of Economics, North Carolina State University, the State University of New York at Binghamton, the University of Florida and the University of Miami as well as by participants at New York Camp Econometrics (April, 2011).

parameter depends on the order of the derivative. Yet, standard selection methods are designed strictly around the fit ( $0^{\text{th}}$ ) of the density or function. A small strand of literature has developed focusing attention on smoothing parameter selection when interest hinges on a higher order derivative. Within this literature there exist several different approaches to construction of the bandwidth. The simplest 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 good bandwidth for estimation of the  $q^{\text{th}}$  order derivative. The analog to traditional data-driven methods (such as cross-validation) has been extended using empirically constructed noise-corrupted derivatives to replace the oracle inside the criterion function. A third approach that has emerged has been to create plug-in approaches.

Each of the existing methods leaves something to be desired in traditional multivariate, mixed data settings. The factor method requires bandwidth selection on the conditional mean followed by calculation of a scaling factor dependent upon the kernel function which for different kernels with mixed data can be tedious. The calculation of noise-corrupted derivatives also requires calculation of 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, the plug-in approaches, while having desirable theoretical properties require the calculation of numerous unknown quantities, which is further compounded 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, making the process streamlined.

To develop the intuition for existing approaches we explain in greater detail the setup for smoothing parameter selection for higher order derivatives. Consider a  $d$ -dimensional multivariate non-parametric regression model

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

Rice (1986) was perhaps the first to propose a method for selecting a smoothing parameter optimal for construction of the  $q^{\text{th}}$ -order derivative of  $g(x)$ . Rice's (1986) focus was univariate in nature. A differencing operator was suggested in Rice (1986) though it was not formally defined and the criterion proposed was a nearly unbiased estimator of mean integrated squared error (MISE) between the estimated  $q^{\text{th}}$ -order derivative and the oracle. Building on the insight of Rice (1986), Müller, Stadmüller & Schmitt (1987) used the 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  $q^{\text{th}}$ -order 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 et al. (2011) proposed a  $q^{\text{th}}$ -order differencing operator with more desirable variance properties as well as a generalized  $C_p$  criterion to be used for selecting the optimal smoothing parameter.

As an alternative to constructing 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 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  $g^{(q)}(x)$  depends on the kernel. Using this fact, Müller et al. (1987) were able to recover an optimal bandwidth for  $g^{(q)}(x)$  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  $q^{\text{th}}$  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 integrated mean square error. Fan, Gijbels, Hu & Huang (1996) show that this bandwidth selection mechanism has desirable properties both theoretically as well as in simulated settings. The approach can also be adapted to find local bandwidths instead of a single, global bandwidth.

As an alternative to the factor method, plug-in and refined bandwidths proposed by Fan & Gijbels (1995), Ruppert (1997) developed empirical-bias bandwidth selection (EBBS). A key difference from Ruppert's (1997) approach is that instead of fitting a higher-order local polynomial to obtain estimates for unknown components in the bias expansion for  $\hat{g}^{(q)}(x)$ , he instead estimates  $g^{(q)}(x)$  for several different bandwidths, then uses least squares to fit a Taylor expansion to estimated the unknown components of the bias. Ruppert (1997) uses the same estimate of the exact, small sample conditional variance as found in Fan & Gijbels (1995), thus the key difference is in the construction of the bias. To understand the differences in estimating the unknown bias components, if one were interested in the first derivative of a two covariate regression function (using local quadratic regression), the Fan & Gijbels's (1995) bandwidth selector would require estimation of at least a 15 variable local quartic whereas the method of Ruppert (1997) would only require estimation of (several) six variable local quadratic procedures. The differences become even more stark as the dimensionality and/or the order of the derivative required increases.

The approach developed in this paper does not rely on the construction of noise-corrupted observations of the appropriate gradients, an adjustment factor, or the calculation of finite sample bias and variance. Rather, we begin with the oracle least-squares cross validation setup for  $\hat{g}^{(1)}(x)$  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. Since we require only the order of the local polynomial to be one degree higher than the desired derivative, we also realize the reduction in computational burden noted in Ruppert (1997) over. 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 2.1 gives a cursory discussion of our cross-validation procedure through the lens of a simple univariate example. Section 2.2 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 GBCV 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 Two Illustrative examples.** In this section we provide an illustrative example to show the conventional LSCV method supplies bandwidths which often lead to estimated regression curves that are too wiggly. In other words, they lead to too much variation in the gradient estimates. Our proposed method delivers bandwidths which result in much smoother fitted curves. Consider a univariate nonparametric regression model

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

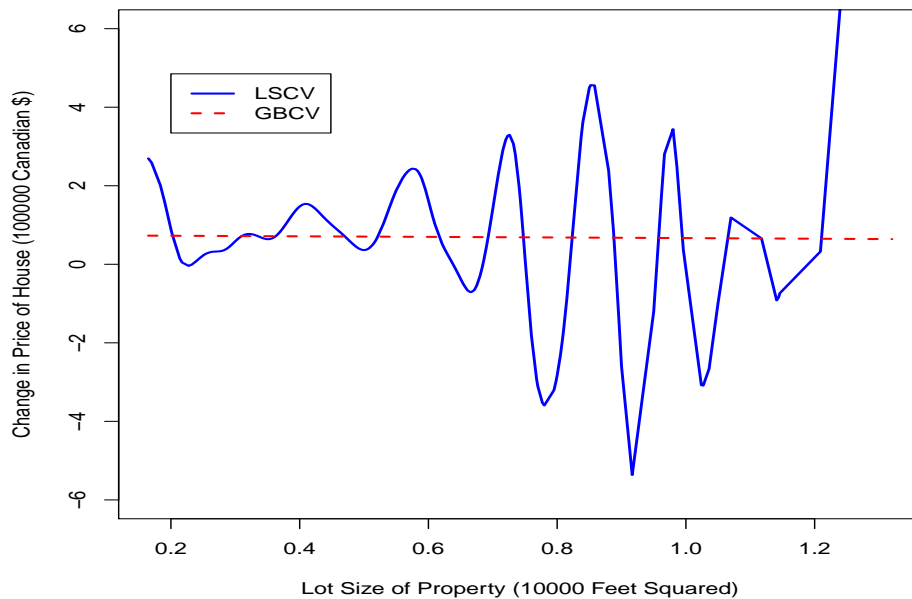
To operationalize a nonparametric kernel estimator for the conditional mean function  $g(\cdot)$ , a smoothing parameter must be selected. The LSCV approach aims to minimize the sample analogue of estimated mean square error, i.e., select  $h$  to minimize  $C_1(h) = n^{-1} \sum_{i=1}^n [g(x_i) - \hat{g}_h(x_i)]^2$ , where  $\hat{g}_h(x_i)$  is a kernel estimator of  $g(x_i)$  that depends on the smoothing parameter  $h$ . In practice, one replaces the unknown function  $g(x_i)$  by  $y_i$ , and to avoid over-fitting, one also needs to replace  $\hat{g}_h(x_i)$  by a leave-one-out version of it, say  $\hat{g}_{-i,h}(x_i)$ . Then one minimizes  $C_2(h) = n^{-1} \sum_{i=1}^n [y_i - \hat{g}_{-i,h}(x_i)]^2$ . It can be shown that the leading terms of  $C_1(h)$  and  $C_2(h)$  are the same, 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 functions. 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 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. Within this literature interest in the gradients stems from the fact that 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).<sup>1</sup> 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. Figure 1 provides the conditional first derivative plots using the two competing data-driven approaches. The LSCV bandwidth produces 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 fitted curve. In contrast, our GBCV method produces a smooth, almost constant, estimated gradient.<sup>2</sup> 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 optimality 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 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 Gaussian kernel. The solid line uses a smoothing parameter of  $h = 0.0272$  obtained via LSCV while the dashed line uses a smoothing parameter of  $h = 1.0139$  obtained via GBCV.



<sup>1</sup>We consider houses which have less than 15,000 square foot lots.

<sup>2</sup>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 .72 to 0.64.

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

$$y_j = g(x_j) + u_j = g(x_{j1}, \dots, x_{jd}) + u_j, \quad j = 1, \dots, n. \quad (3)$$

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

$$\beta(x) \stackrel{\text{def}}{=} \frac{\partial g(x)}{\partial x} = \begin{pmatrix} \frac{\partial g(x)}{\partial x_1} \\ \vdots \\ \frac{\partial g(x)}{\partial x_d} \end{pmatrix}$$

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

$$g(x_j) = g(x_i) + (x_j - x_i)^T \beta(x_i) + R_{ji}, \quad (4)$$

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  $(a, b^T)^T \in \mathcal{R}^{d+1}$  to minimize the following objective function

$$\min_{a, b} \sum_{j=1}^n [y_j - a - (x_j - x)^T b]^2 W_{h, jx}, \quad (5)$$

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

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

$$\sum_{j=1}^n [y_j - a - (x_j - x)^T b] \begin{pmatrix} 1 \\ x_j - x \end{pmatrix} W_{h, jx} = 0, \quad (6)$$

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

$$\hat{\delta}(x) = \begin{pmatrix} \hat{g}(x) \\ \hat{\beta}(x) \end{pmatrix} = \left[ \sum_{j=1}^n W_{h, jx} \begin{pmatrix} 1, & (x_j - x)^T \\ x_j - x, & (x_j - x)(x_j - x)^T \end{pmatrix} \right]^{-1} \sum_{j=1}^n W_{h, jx} \begin{pmatrix} 1 \\ x_j - x \end{pmatrix} y_j. \quad (7)$$

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  $\sum_{j \neq i}^n$ .

$$\hat{\delta}_{-i}(x_i) = \begin{pmatrix} \hat{g}_{-i}(x_i) \\ \hat{\beta}_{-i}(x_i) \end{pmatrix} = \left[ \sum_{j \neq i}^n W_{h,j,i} \begin{pmatrix} 1, & (x_j - x_i)^T \\ x_j - x_i, & (x_j - x_i)(x_j - x_i)^T \end{pmatrix} \right]^{-1} \sum_{j \neq i}^n W_{h,j,i} \begin{pmatrix} 1 \\ x_j - x_i \end{pmatrix} y_j, \quad (8)$$

where  $W_{h,j,i} = \prod_{s=1}^d h_s^{-1} w((x_{js} - x_{is})/h_s)$ .

Define two column vectors  $e_g = (1, \mathbf{0}_{1 \times d})^T$  and  $e_\beta = (0, \iota^T)^T$ , where  $\mathbf{0}_{1 \times d}$  is a  $1 \times d$  row vector of zeros, and  $\iota = (1, \dots, 1)^T$  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  $h = (h_1, \dots, h_d)$  to minimize a feasible version of the following (infeasible) cross-validation objective function:

$$C_{0,\beta}(h) = \frac{1}{n} \sum_{i=1}^n \|\hat{\beta}_{-i}(x_i) - \beta(x_i)\|^2 M(x_i), \quad (9)$$

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  $s^{\text{th}}$  diagonal equals  $h_s^{-2}$ ), and note that  $A_n^{-1} B_n = A_n^{-1} G_n^{-1} G_n B_n = (G_n A_n)^{-1} G_n B_n$ , we obtain

$$\begin{aligned} \hat{\delta}_{-i}(x_i) &= \left[ \sum_{j \neq i} W_{h,j,i} G_n \begin{pmatrix} 1 \\ x_j - x_i \end{pmatrix} (1, (x_j - x_i)^T) \right]^{-1} \sum_{j \neq i} W_{h,j,i} G_n \begin{pmatrix} 1 \\ x_j - x_i \end{pmatrix} y_j \\ &= \left[ \sum_{j \neq i} W_{h,j,i} \begin{pmatrix} 1 & (x_j - x_i)^T \\ D_h^{-2}(x_j - x_i) & D_h^{-2}(x_j - x_i)(x_j - x_i)^T \end{pmatrix} \right]^{-1} \sum_{j \neq i} W_{h,j,i} \begin{pmatrix} 1 \\ D_h^{-2}(x_j - x_i) \end{pmatrix} y_j. \end{aligned} \quad (10)$$

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

$$n^{-1} \sum_{j \neq i} W_{h,j,i} \begin{pmatrix} 1 & (x_j - x_i)^T \\ D_h^{-2}(x_j - x_i) & D_h^{-2}(x_j - x_i)(x_j - x_i)^T \end{pmatrix}$$

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

$$\begin{aligned} y_j &= g(x_j) + u_j = g(x_i) + (x_j - x_i)^T \beta(x_i) + R_{ji} + u_j \\ &= \left(1, (x_j - x_i)^T\right) \begin{pmatrix} g(x_i) \\ \beta(x_i) \end{pmatrix} + R_{ji} + u_j. \end{aligned} \quad (11)$$

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

$$\hat{\delta}_{-i}(x_i) = \delta(x_i) + A_{2i}^{-1} A_{1i},$$

where

$$A_{1i} = \frac{1}{n} \sum_{j \neq i} W_{h,ji} \begin{pmatrix} 1 \\ D_h^{-2}(x_j - x_i) \end{pmatrix} [R_{ji} + u_j], \quad (12)$$

and

$$A_{2i} = \begin{pmatrix} \hat{f}_i & \mathcal{B}_{1i}^T \\ D_h^{-2} \mathcal{B}_{1i} & D_h^{-2} \mathcal{B}_{2i} \end{pmatrix}, \quad (13)$$

where  $\hat{f}_i = n^{-1} \sum_{j \neq i} W_{h,ji}$ ,  $\mathcal{B}_{1i} = n^{-1} \sum_{j \neq i} W_{h,ji}(x_j - x_i)$ , and  $\mathcal{B}_{2i} = n^{-1} \sum_{j \neq i} W_{h,ji}(x_j - x_i)(x_j - x_i)^T$ .

Recall that  $e_\beta = (0, \iota^T)^T$ , is a  $(d+1)$  column vector whose first element is zero with all other elements are equal to one. Therefore we have  $\hat{\beta}_{-i}(x_i) = e_\beta^T \hat{\delta}_{-i}(x_i) = \beta(x_i) + e_\beta^T A_{2i}^{-1} A_{1i}$ . Rearranging terms we obtain

$$\hat{\beta}_{-i}(x_i) - \beta(x_i) = e_\beta^T A_{2i}^{-1} A_{1i}. \quad (14)$$

$A_{2i}$  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

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

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

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

One may be tempted 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 (16) are the normal equations for the local linear estimators (see 6). Hence, (16) becomes identically zero if one replaces  $\tilde{g}_{-i}(x_i)$  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.,

$$\tilde{g}_{-i}(x_i) = \frac{n^{-1} \sum_{j \neq i}^n y_j W_{h,ji}}{n^{-1} \sum_{j \neq i}^n W_{h,ji}}, \quad (17)$$

$$\begin{aligned} \tilde{\beta}_{-i}(x_i) &= \left[ \frac{n^{-1} \sum_{j \neq i}^n y_j W'_{h,ji}}{n^{-1} \sum_{l \neq i}^n W_{h,li}} - \frac{(n^{-1} \sum_{l \neq i}^n y_l W_{h,li})(n^{-1} \sum_{j \neq i}^n W'_{h,ji})}{(n^{-1} \sum_{l \neq i}^n W_{h,li})^2} \right] \\ &= \frac{n^{-2} \sum_{j \neq i}^n \sum_{l \neq i, j}^n (y_j - y_l) W'_{h,ji} W_{h,li}}{(n^{-1} \sum_{l \neq i}^n W_{h,li})^2}, \end{aligned} \quad (18)$$

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

$$W'_{h,ji,s} = \frac{\partial}{\partial x_{is}} W_{h,ji} = - \left[ \prod_{t \neq s} h_t^{-1} w \left( \frac{x_{jt} - x_{it}}{h_t} \right) \right] \frac{1}{h_s^2} \frac{\partial w(u)}{\partial u} \Big|_{u=(x_{js}-x_{is})/h_s}.$$

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

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

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  $\mathcal{R}^d$ ). (iii)  $f(x)$  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)  $w'(v) = dw(v)/dv$  is a continuous and bounded function; (iii)  $\int w(v)v^6$  and  $\int |w'(v)|v^6 dv$  are both finite.

**Assumption 2.4.**  $(h_1, \dots, h_d) \in H_n$ , where  $H_n = \{h \in \mathcal{R}_+^{d+2} : c_1 n^{-1/(d+\delta_1)} \leq h_s \leq c_2 n^{-1/(d+6+\delta_2)}, s = 1, \dots, d\}$ , for 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  $\max_{1 \leq s \leq d} h_s \rightarrow 0$  and  $nh_1 \dots h_d \sum_{s=1}^d h_s^2 \rightarrow \infty$  as  $n \rightarrow \infty$ , which are needed for the estimation bias and variance to converge to 0 as  $n$  gets large.

For expositional 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*

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

uniformly in  $h \in H_n$ ,

$$(ii) \quad \hat{h} = c_{opt}n^{-1/7} + o_p(n^{-1/7}),$$

where  $B(x) = B_1(x)^2 + B_2(x)^2$ ,  $B_1(x) = \left(\frac{\mu_4 - \mu_2^2}{2\mu_2}\right) \frac{g''(x)f'(x)}{f(x)} + \frac{\mu_4 g'''(x)}{6\mu_2}$ ,  $B_2(x) = \mu_2[g''(x)f'(x)f(x) + 2g'(x)f''(x)f(x) + (1/2)g'''(x)f(x)^2 - g'(x)(f'(x))^2]/f(x)^2$ ,  $\zeta_0 = \kappa_0 + \nu_2/\mu_2^2$ ,  $\kappa_0 = \int [W'(v)]^2 dv$ ,  $\mu_j = \int W(v)v^j dv$ ,  $\nu_j = \int W(v)^2 v^j dv$ ,  $\sigma^2(x) = E(u_i^2 | x_i = x)$ ,  $c_{opt} = \left[\frac{3\zeta_0 \int \sigma^2(x)M(x)dx}{4 \int B(x)M(x)f(x)dx}\right]^{1/7}$ ,  $m'(\cdot)$ ,  $m''(\cdot)$  and  $m'''(\cdot)$  denote the first, second and third derivative functions of  $m(\cdot)$  with  $m = g$  or  $m = f$ .

It is interesting to observe that the leading terms of  $C_\beta(h)$  have two parts, one part is related to a weighted version of the estimated MSE of  $\hat{\beta}(x)$ , the local linear estimator of  $\beta(x)$ , the second part 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

$$\sqrt{nh^3} \left[ \hat{\beta}(x) - \beta(x) - h^2 B_1(x) \right] \xrightarrow{d} N \left( 0, \frac{\nu_2}{\mu_2^2} \frac{\sigma^2(x)}{f(x)} \right), \quad (20)$$

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

$$\sqrt{nh^3} \left[ \tilde{\beta}(x) - \beta(x) - h^2 B_2(x) \right] \xrightarrow{d} N \left( 0, \frac{\kappa_0 \sigma^2(x)}{f(x)} \right). \quad (21)$$

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

$$C_\beta(h) = \int \left[ MSE(\hat{\beta}(x)) + MSE(\tilde{\beta}(x)) \right] f(x)M(x)dx + o_p(h^4 + (nh^3)^{-1}). \quad (22)$$

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  $C_\beta(h)$  objective function.

The next Theorem deals with the general multivariate  $x$  case. For the general  $d$ -dimensional  $x$ , we denote  $m_s(x) = \frac{\partial m(x)}{\partial x_s}$ ,  $m_{ts}(x) = \frac{\partial^2 m(x)}{\partial x_t \partial x_s}$  and  $m_{sss}(x) = \frac{\partial^3 m(x)}{\partial x_s^3}$ , 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*

$$(i) \quad C_\beta(h) = \int B_h(x)M(x)f(x)dx + \frac{1}{nh_1\dots h_d}\zeta_0 \sum_{s=1}^d \frac{1}{h_s^2} + o_p\left(\|h\|^4 + \frac{1}{nh_1\dots h_d\|h\|^2}\right)$$

uniformly in  $h \in H_n$ ,

$$(ii) \quad \hat{h}_s = c_{0,s}n^{-1/(d+6)} + o_p\left(n^{-1/(d+6)}\right),$$

where  $\zeta_0 = \int [c_{U}(x) + c_{Ic}(x)]M(x)dx$ ,  $c_{U}(x) = \nu_0^{d-1}\nu_2\sigma^2(x)/\mu_2^2$ ,  $c_{Ic}(x) = \nu_0^{d-1}\kappa_0\sigma^2(x)$ ,  $\|h\|^2 = \sum_{s=1}^d h_s^2$ ,  $B_h(x) = \sum_{s=1}^d (B_{1s,h}(x)^2 + B_{2s,h}(x)^2)$  with

$$\begin{aligned} B_{1s,h}(x) &= \frac{h_s^2\mu_4g_{sss}(x)}{6\mu_2} + \frac{h_s^2g_{ss}(x)f_s(x)\mu_4}{2\mu_2f(x)} - \frac{\mu_2f_s(x)\sum_{t=1}^d g_{tt}(x)h_t^2}{2f(x)} \\ &\quad + \frac{1}{2} \frac{f_s(x)\sum_{t \neq s} h_t^2 g_{tt}(x)}{2f(x)} + \mu_2 \sum_{t \neq s} h_t^2 f_t(x)g_{ts}(x)/f(x) + (\mu_2/2) \sum_{t \neq s} h_t^2 g_{tt}(x), \end{aligned}$$

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

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

We can see from Theorem 2.2 that the leading squared bias term is complicated as it involves partial derivatives up to the 3<sup>rd</sup> order. If one elected to use a plug-in method to select optimal smoothing parameters, initial 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 and we omit its proof here. Although we do not provide a proof for Theorem 2.2, 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

$$C_\beta(h) = \int \text{Tr} \left[ \text{MSE} \left( \hat{\beta}(x) \right) + \text{MSE} \left( \tilde{\beta}(x) \right) \right] f(x)M(x)dx + o_p \left( \|h\|^4 + \frac{1}{nh_1 \dots h_d \|h\|^2} \right),$$

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 the local linear and the local constant estimators of  $\beta(\cdot)$ .

### 3. SIMULATION STUDY

In this section we perform a small scale 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

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

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

**DGP1:**  $g(x) = 2 + \frac{e^{-3x}}{1+e^{-3x}};$

**DGP2:**  $g(x) = x + 2e^{-16x^2}.$

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 10<sup>th</sup> and 90<sup>th</sup> percentiles of these values over the 1000 replications. We use a second-order Gaussian kernel for all simulations. The results are presented in Table 1 for each cross-validation procedure. We find uniform improvement for our cross-validation method versus rate-adjusted LSCV in terms of estimated MSE. Moreover, we see drastic improvements at the upper percentile. For example, when  $n = 200$ , the 90<sup>th</sup> percentile of the estimated MSE of the estimated gradient of DGP1 using rate adjusted bandwidths selected via LSCV is over 11 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.

TABLE 1. Summary of simulation results at the median [10<sup>th</sup> percentile, 90<sup>th</sup> percentile] for the gradient

	DGP1		DGP2	
	LSCV	GBCV	LSCV	GBCV
$n = 50$				
MSE	0.1835 [0.0685, 12.3627]	0.1536 [0.0593, 0.4709]	5.7797 [3.6825, 20.0336]	5.0647 [3.3787, 7.3079]
$n = 100$				
MSE	0.1268 [0.0635, 7.4148]	0.1145 [0.0422, 0.3238]	5.4223 [3.5026, 13.5550]	4.8030 [3.3424, 6.2517]
$n = 200$				
MSE	0.1033 [0.0583, 2.8705]	0.0833 [0.0325, 0.2247]	4.8286 [3.0643, 9.0929]	4.5282 [2.9498, 5.6846]

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 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 some simple univariate regression models. 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

$$y_{jt} = \alpha + \beta_1 kg_{jt} + \beta_2 kp_{jt} + \beta_3 emp_{jt} + \beta_4 unem_{jt} + \varepsilon_{jt} \quad (23)$$

to study the public capital productivity puzzle. Here,  $y_{jt}$  denotes the gross state product of state  $j$  ( $j = 1, \dots, 48$ ) in period  $t$  ( $t = 1970, \dots, 1986$ ). Covariates include public capital ( $kg$ ) which aggregates highways and streets, water and sewer facilities, and other public buildings and structures,  $kp$  is the Bureau of Economic Analysis' private capital stock estimates, and labor ( $emp$ ) is employment in non-agricultural payrolls. Each of these variables are measured in logarithms. Following Baltagi & Pinnoi (1995), we also use the unemployment rate ( $unem$ ) 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.<sup>3</sup>

TABLE 2. Bandwidths and scale factors obtained via LSCV and GBCV for the public capital data. Column labelled ratio is the ratio of the GBCV bandwidth (scale factor) and the LSCV bandwidth (scale factor) for each regressor. Bandwidths obtained for both methods using Nelder-Mead non-gradient optimization.

	Bandwidths			Scale Factors		
	LSCV	GBCV	Ratio	LSCV	GBCV	Ratio
$\ln(kg)$	0.087	1.103	12.690	0.213	2.288	10.732
$\ln(kp)$	0.100	0.582	5.804	0.251	1.231	4.909
$\ln(emp)$	0.042	0.600	14.285	0.095	1.152	12.081
$unem$	3.923	1.407	0.359	4.061	1.232	0.303

We estimate nonparametric versions of Equation (23) using local linear least-squares with second-order Gaussian 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 from GBCV relative to LSCV, the bandwidth for unemployment actually decreases (the other three bandwidths increase as expected). The ratio of the estimated bandwidths reveals that indeed the two methods can provide substantially different bandwidths in applied settings and that the 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 90th and 10th percentile point estimates,  $D_{90}$  and  $D_{10}$ , 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.<sup>4</sup> The median  $t$ -ratio is greater than four, suggesting that public capital has a significant positive impact on gross state product.

<sup>3</sup>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.

<sup>4</sup>Negative 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 10<sup>th</sup> and 90<sup>th</sup> 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.

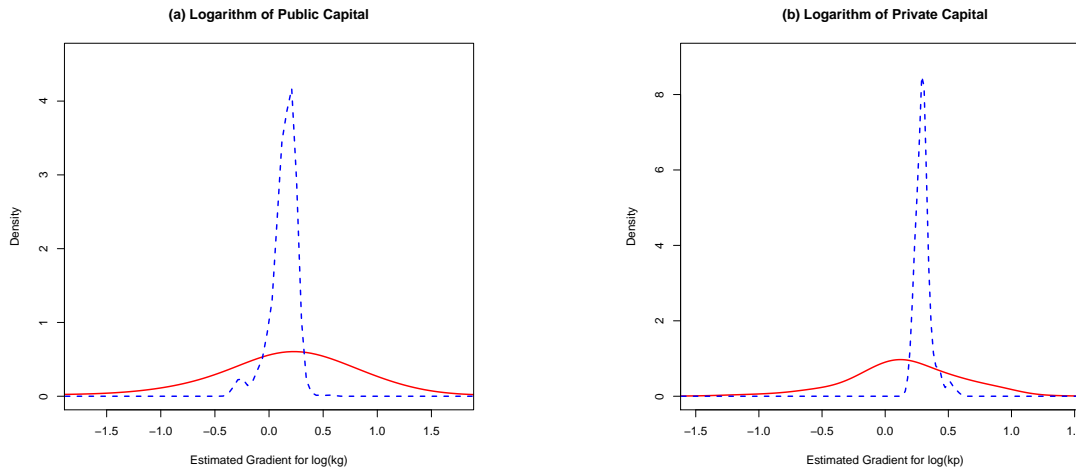
	$D_{10}$	<i>Median</i>	$D_{90}$	<i>SD</i>
$\widehat{\beta}_{kg}(LSCV)$	-0.6754	0.2010	0.9173	1.3902
	1.4462	0.1922	0.4698	
$\widehat{\beta}_{kg}(GBCV)$	-0.0094	0.1590	0.2504	0.1204
	0.0323	0.0373	0.0262	
$\widehat{\beta}_{kp}(LSCV)$	-0.4167	0.1525	0.7184	0.5703
	0.0854	0.0768	0.1446	
$\widehat{\beta}_{kp}(GBCV)$	0.2322	0.2946	0.3668	0.0634
	0.0278	0.0268	0.0281	
$\widehat{\beta}_{emp}(LSCV)$	-0.0477	0.8260	1.6496	0.7627
	0.2490	0.2212	0.7219	
$\widehat{\beta}_{emp}(GBCV)$	0.5291	0.6340	0.6974	0.0674
	0.0349	0.0380	0.0351	
$\widehat{\beta}_{unem}(LSCV)$	-0.0216	-0.0048	0.0103	0.0167
	0.0108	0.0644	0.0059	
$\widehat{\beta}_{unem}(GBCV)$	-0.0133	-0.0055	0.0019	0.0061
	0.0059	0.0032	0.0049	

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, very few of the estimates are statistically significant. It is difficult to place much faith on an elasticity estimated with such a low 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 Figures 3 and 4 which show three-dimensional plots of the estimated conditional gradients for public capital as public capital and employment vary (we hold private capital and unemployment fixed at their respective medians). Figure 3 uses the bandwidths from LSCV to construct this grid of estimates. It is difficult to provide an intuitive description of this figure. The shape of the curve does not reflect

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.



the assumed relationship between these variables and it is quite bumpy. On the other hand, Figure 4 shows that the estimated gradient is quite smooth and we do not see any of the large upward or downward spikes in the conditional mean estimates.

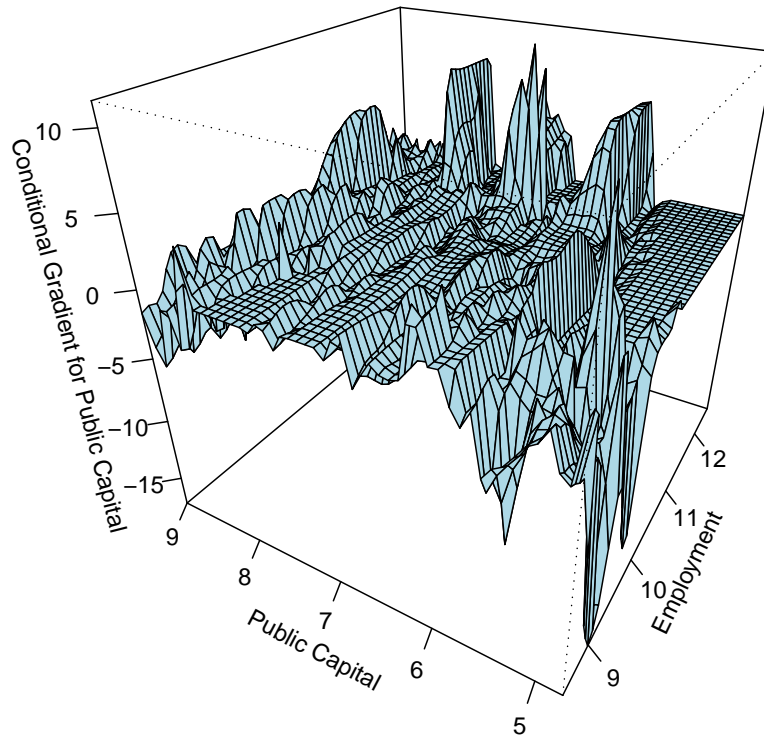
The result from the GBCV bandwidths are in line with conventional wisdom. We see that as the levels of private and public capital rise, we expect to have higher values of gross state product. Further, the nonlinear relationship shows that the return to public capital is lower than that of private capital. 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. It leads to a much smoother function estimation result than that obtained by using 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

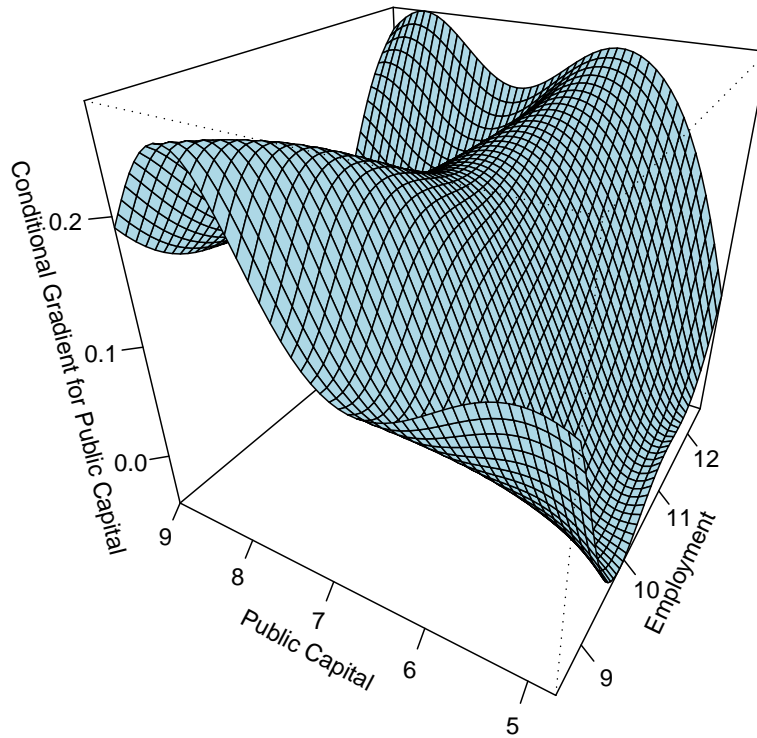
In this paper we propose a novel approach to select bandwidths in nonparametric kernel regression. In contrast to past measures, we are primarily concerned with estimation of the gradient function. Estimation of gradients is often of more interest as studying ‘marginal effects’ is a corner

FIGURE 3. Estimated conditional gradient for public capital using LSCV bandwidths, holding private capital and unemployment fixed at their medians. This curve is estimated over a grid of 2500 points based on the rectangle constructed from the limits of the logarithms of public capital and the employment rate.



stone 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.

FIGURE 4. Estimated conditional gradient for public capital using GBCV bandwidths, holding private capital and unemployment fixed at their medians. This curve is estimated over a grid of 2500 points based on the rectangle constructed from the limits of the logarithms of public capital and the employment rate.



Our simulations show large improvements in performance when estimating the gradient function. When we applied our method to several 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 significant variation.

There exist many possible extensions of our proposed method. For example, one can extend our method to the case of selecting smoothing parameters that are optimal for estimating higher-order derivation functions. 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

- Anglin, P. M. & Gençay, R. (1996), ‘Semiparametric estimation of a hedonic price function’, *Journal of Applied Econometrics* **11**, 633–648.
- Bajari, P. & Kahn, M. E. (2005), ‘Estimating housing demand with an application to explaining racial segregation in cities’, *Journal of Business and Economic Statistics* **23**(1), 20–33.
- Baltagi, B. H. & Pinnoi, N. (1995), ‘Public capital stock and state productivity growth: Further evidence from an error components model’, *Empirical Economics* **20**, 351–359.
- Bellver, C. (1987), ‘Influence of particulate pollution on the positions of neutral points in the sky at Seville (Spain)’, *Atmospheric Environment* **21**(3), 699–702.
- Cai, Z., Fan, J. & Yao, Q. (2000), ‘Functional coefficient regression models for nonlinear time series’, *Journal of the American Statistical Association* **95**(451), 941–956.
- Charnigo, R., Francoeur, M., Kenkel, P., Mengüç, M. P., Hall, B. & Srinivasan, C. (2007), ‘Derivatives of scattering profiles: Tools for nanoparticle characterization’, *Journal of the Optical Society of American A* **24**, 2578–2589.
- Charnigo, R., Hall, B. & Srinivasan, C. (2011), ‘A generalized  $c_p$  criterion for derivative estimation’, *Technometrics* **53**(3), 238–253.
- Cleveland, W. S. (1993), *Visualizing Data*, Hobart Press, Summit, NJ.
- Fan, J. & Gijbels, I. (1995), ‘Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation’, *Journal of the Royal Statistical Society. Series B* **57**(2), 371–394.
- Fan, J., Gijbels, I., Hu, T.-C. & Huang, L. (1996), ‘A study of variable bandwidth selection for local polynomial regression’, *Statistica Sinica* **6**, 113–127.
- Hall, P., Li, Q. & Racine, J. S. (2007), ‘Nonparametric estimation of regression functions in the presence of irrelevant regressors’, *Review of Economics and Statistics* **89**(4), 784–789.
- Heckman, J. J., Matzkin, R. L. & Nesheim, L. (2010), ‘Nonparametric identification and estimation of nonadditive hedonic models’, *Econometrica* **78**(5), 1569–1591.
- Lee (1990), *U-Statistics*, Marcel Dekker, New York, New York.
- Müller, H.-G., Stadtmüller, U. & Schmitt, T. (1987), ‘Bandwidth choice and confidence intervals for derivatives of noisy data’, *Biometrika* **74**(4), 743–749.
- Munnell, A. H. (1990), ‘How does public infrastructure affect regional economic performance?’, *New England Economic Review* **September**, 11–32.
- Overholt, M. R. & Pope, S. B. (1996), ‘Direct numerical simulation of a passive scalar with imposed mean gradient in isotropic turbulence’, *Physics of Fluids* **8**(11), 3128–3148.
- Parmeter, C. F., Henderson, D. J. & Kumbhakar, S. C. (2007), ‘Nonparametric estimation of a hedonic price function’, *Journal of Applied Econometrics* **22**, 695–699.
- Ramsay, J. O. & Silverman, B. W. (2005), *Applied Functional Data Analysis*, Springer-Verlag, New York.
- Rice, J. A. (1986), ‘Bandwidth choice for differentiation’, *Journal of Multivariate Analysis* **19**, 251–264.
- Ruppert, D. (1997), ‘Empirical-bias bandwidth for local polynomial nonparametric regression and density estimation’, *Journal of the American Statistical Association* **92**(439), 1049–1062.
- Wahba, G. & Wang, Y. (1990), ‘When is the optimal regularization parameter insensitive to the choice of loss function?’, *Communications in Statistics* **19**, 1685–1700.

## APPENDIX A. PROOF OF THEOREM 2.1

The proof of Theorem 2.1 is quite tedious. Therefore, it is necessary to introduce some short-hand 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:  $g_i = g(x_i)$ ,  $\beta_i = \beta(x_i)$ ,  $f_i = f(x_i)$ ,  $\hat{f}_i = \hat{f}(x_i)$ ,  $\tilde{g}_i = \tilde{g}_{-i}(x_i)$ ,  $\tilde{\beta}_i = \tilde{\beta}_{-i}(x_i)$ , etc. Also, we will often omit the weight function  $M(x_i)$  to save space.
- (2) We define  $\sum_i = \sum_{i=1}^n$ ,  $\sum \sum_{j \neq i} = \sum_{i=1}^n \sum_{j=1, j \neq i}^n$ ,  $\sum \sum \sum_{l \neq j \neq i} = \sum_{i=1}^n \sum_{j=1, j \neq i}^n \sum_{l=1, l \neq i, l \neq j}^n$ . Similarly,  $\sum \sum \sum \sum \sum_{i \neq j \neq l \neq k \neq m}$  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

$$\begin{aligned}
\hat{A}_{1i} &= \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( h^{-2}(x_j - x_i) \right) \left\{ g_j + u_j - g_i - (x_j - x_i)\beta(x_i) - [\tilde{g}_i - g_i] - (x_j - x_i)[\tilde{\beta}_i - \beta_i] \right\} \\
&= \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( h^{-2}(x_j - x_i) \right) \left\{ R_{ji} + u_j - [\tilde{g}_i - g_i] - (x_j - x_i)[\tilde{\beta}_i - \beta_i] \right\} \\
&= A_{1i} - \Delta_i
\end{aligned} \tag{A.1}$$

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

$$\Delta_i = \frac{1}{n} \sum_{j \neq i} W_{h,ji} \left( h^{-2}(x_j - x_i) \right) \left\{ [\tilde{g}_i - g_i] + (x_j - x_i)[\tilde{\beta}_i - \beta_i] \right\}. \tag{A.2}$$

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

$$\begin{aligned}
C_\beta(h) &= \frac{1}{n} \sum_{i=1}^n [e_\beta^T A_{2i}^{-1} A_{1i} - e_\beta^T A_{2i}^{-1} \Delta_i]^2 \\
&= \frac{1}{n} \sum_{i=1}^n [(e_\beta^T A_{2i}^{-1} A_{1i})^2 + (e_\beta^T A_{2i}^{-1} \Delta_i)^2 - 2(e_\beta^T A_{2i}^{-1} A_{1i})(e_\beta^T A_{2i}^{-1} \Delta_i)] \\
&= C_1 + C_2 - 2C_3
\end{aligned} \tag{A.3}$$

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

$$\begin{aligned} C_1 &= h^4 \int B_1(x)^2 M(x) f(x) dx + \frac{1}{nh^3} \frac{\nu_2}{\mu_2^2} \int \sigma(x)^2 M(x) dx + (s.o.), \\ C_2 &= h^4 \int B_2(x)^2 M(x) f(x) dx + \frac{1}{nh^3} \kappa_0 \int \sigma(x)^2 M(x) dx + (s.o.), \\ C_3 &= o_p(h^4 + (nh^3)^{-1}) \end{aligned}$$

uniformly in  $h \in H_n$ .

Note that  $B(x) = B_1(x)^2 + B_2(x)^2$  and  $\zeta_0 = \kappa_0 + \nu_2/\mu_2^2$ . This proves Theorem 2.1 (i). Theorem (ii) follows from Theorem (i).  $\square$

**Lemma A.1.**  $C_1 = h^4 \int B_1(x)^2 M(x) f(x) dx + \frac{\nu_0}{nh^3 \mu_2^2} \int \sigma(x)^2 M(x) dx + o_p(h^4 + (nh^3)^{-1})$  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  $x_i \in \mathcal{M}$ , where  $\mathcal{M}$  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

$$e_\beta^T A_{2i}^{-1} = (-C_{1i}, C_{2i}) + O_p \left( h^2 + \frac{(\ln(n))^{1/2}}{\sqrt{nh}} \right). \quad (\text{A.4})$$

Combining (A.4) and (12) lead to

$$\begin{aligned} e_\beta^T A_{2i}^{-1} A_{1i} &= n^{-1} \sum_{j \neq i} W_{h,ji} (R_{ji} + u_j) [C_{2i} h^{-2} (x_j - x_i) - C_{1i}] + (s.o.) \\ &= n^{-1} \sum_{j \neq i} P_{ji} (R_{ji} + u_j) + (s.o.), \end{aligned} \quad (\text{A.5})$$

where

$$P_{ji} = W_{h,ji} [C_{2i} h^{-2} (x_j - x_i) - C_{1i}]. \quad (\text{A.6})$$

Substituting the above result into  $C_1$  gives,

$$\begin{aligned}
C_1 &= n^{-1} \sum_{i=1}^n (e_{\beta}^T A_{2i}^{-1} A_{1i})^2 \\
&= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji} P_{li} (R_{ji} + u_j)(R_{li} + u_l) + (s.o.) \\
&= n^{-3} \sum_i \sum_{j \neq i} P_{ji}^2 (R_{ji} + u_j)^2 + n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq j \neq i} P_{ji} P_{li} (R_{ji} + u_j)(R_{li} + u_l) + (s.o.) \\
&= C_{1,1} + C_{1,2} + (s.o.)
\end{aligned}$$

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

From  $(R_{ji} + u_j)^2 = R_{ji}^2 + 2u_j R_{ji} + u_j^2$ , it is easy to see that the leading term of  $C_{1,1}$  is associated with the term  $u_j^2$ , we use  $C_{1,1}^0$  to denote this leading term.

$$\begin{aligned}
C_{1,1}^0 &= n^{-3} \sum_i \sum_{j \neq i} P_{ji}^2 u_j^2 = n^{-1} E[\sigma_j^2 P_{ji}^2] + (s.o.) = n^{-1} E[E(\sigma_j^2 P_{ji}^2 | x_i)] + (s.o.) \\
&= \frac{\nu_2}{nh^3 \mu_2^2} \int \sigma(x)^2 dx + (s.o.)
\end{aligned}$$

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

$$\begin{aligned}
E\{E[\sigma_j^2 P_{ji}^2 | x_i]\} &= E\left\{E\left[W_{h,ji}^2 u_j^2 (C_{1i} - C_{2i} h^{-2} (x_j - x_i))^2 \mid x_i\right]\right\} \\
&= E\left\{E\left[W_{h,ji}^2 \sigma(x_j)^2 (C_{2i}^2 h^{-4} (x_j - x_i)^2 + C_{1i}^2 - 2C_{2i} h^{-2} (x_j - x_i) C_{1i}) \mid x_i\right]\right\} \\
&= E\left[\frac{\nu_2}{h^3} \sigma(x_i)^2 C_{2i}^2 f(x_i) + \frac{\mu_0}{nh} \sigma(x_i)^2 C_{1i}^2 f(x_i) - \frac{2\mu_2}{nh} \sigma(x_i)^2 C_{1i} C_{2i} f'(x_i)\right] \\
&= \frac{\nu_2}{h^3 \mu_2^2} \int \sigma(x)^2 dx + O(h^{-1}),
\end{aligned}$$

where  $\sigma^2(x_i) = E(u_i^2 | x_i)$  and we used  $C_{2i} = 1/(\mu_2 f(x_i))$  and  $\nu_2 = \int W(v) v^2 dv$ .

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

$$C_{1,1} = C_{1,1}^0 + (s.o.) = \frac{\nu_2}{nh^3 \mu_2^2} \int \sigma(x)^2 M(x) dx + (s.o.). \quad (\text{A.7})$$

Next, we consider  $C_{1,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq j \neq i} P_{ji} P_{li} (R_{ji} R_{li} + u_j u_l + 2R_{ji} u_l) = C_{1,2,a} + C_{1,2,b} + 2C_{1,2,c}$ .  $C_{1,2,a} = \sum_{l \neq j \neq i} P_{ji} P_{li} R_{ji} R_{li}$ , which can be written as a third-order U-statistic, whose

leading term is its expectation:

$$\begin{aligned}
E(C_{1,2,a}) &= \left\{ E \left\{ E \left[ P_{ji} R_{ji} \middle| x_i \right] \right\} \right\}^2 = \left\{ E \left\{ E \left[ W_{h,ji} (h^{-2} C_{2i} (x_j - x_i) - C_{1i}) R_{ji} \middle| x_i \right] \right\} \right\}^2 \\
&= \left\{ h^2 E \left[ \frac{\mu_4}{2} g''(x_i) C_{2i} f'(x_i) + \frac{\mu_4}{6} f(x_i) g'''(x_i) - \frac{\mu_2}{2} f(x_i) g''(x_i) C_{1i} \right] \right\}^2 + O(h^6) \\
&= h^4 \int B_1(x)^2 f(x) dx + O(h^6),
\end{aligned}$$

where  $B_1(x) = \left( \frac{\mu_4 - \mu_2^2}{2\mu_2} \right) \frac{g''(x)f'(x)}{f(x)} + \frac{\mu_4 g'''(x)}{6\mu_2}$ ,  $\mu_l = \int W(v) v^l dv$  ( $l = 2, 4$ ).

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  $E(C_{1,2,c}^2) = n^{-6} [O(n^5 h^6) + O(n^4 h^4)] = O(h^6/n) = o(h^8)$ . Hence,  $C_{1,2,c} = o_p(h^4)$ .

Summarizing the above we have shown that (adding back the trimming function  $M(\cdot)$ )

$$C_1 = h^4 \int B_1(x)^2 f(x) M(x) dx + \frac{\nu_2}{nh^3 \mu_2^2} \int \sigma(x)^2 M(x) dx + o_p(h^4 + (nh^3)^{-1}). \quad (\text{A.8})$$

Moreover, by using Rosenthal's and Markov's inequalities, one can show that (A.8) holds true uniformly in  $h \in H_n$ . This completes the proof of lemma A.1.  $\square$

**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  $P_{ji} = W_{h,ji} [C_{2i} h^{-2} (x_j - x_i) - C_{1i}]$ . Then By (A.2) and (A.4), we have

$$\begin{aligned}
e_\beta^T A_{2i}^{-1} \Delta_i &= n^{-1} \sum_{j \neq i} P_{ji} \left\{ [\tilde{g}_i - g_i] + (x_j - x_i) [\tilde{\beta}_i - \beta_i] \right\} + (s.o.) \\
&= D_{1i} + D_{2i} + (s.o.),
\end{aligned} \quad (\text{A.9})$$

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.9) into  $C_2$  we get

$$\begin{aligned}
C_2 &= n^{-1} \sum_i D_{1i}^2 + n^{-1} \sum_i D_{2i}^2 + 2n^{-1} \sum_i D_{1i} D_{2i} + (s.o.) \\
&= C_{2,1} + C_{2,2} + 2C_{2,3} + (s.o.),
\end{aligned}$$

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

$$\tilde{g}_i - g_i = (\tilde{g}_i - g_i) \hat{f}_i / \hat{f}_i = (\tilde{g}_i - g_i) \hat{f}_i / f_i + (s.o.) = \tilde{m}_i + (s.o.), \quad (\text{A.10})$$

where

$$\tilde{m}_i = (\tilde{g}_i - g_i) \hat{f}_i / f_i = \frac{1}{n} \sum_{k \neq i}^n [g_k + u_k - g_i] W_{h,ki} / f_i. \quad (\text{A.11})$$

Hence, we can replace  $(\tilde{g}_i - g_i)^2$  by  $\tilde{m}_i^2$  to obtain the leading term of  $C_{2,1}$ , i.e., Using (A.11) we get

$$\begin{aligned} C_{2,1} &= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji} P_{li} \tilde{m}_i^2 / f_i^2 + (s.o.) \\ &= n^{-3} \sum_i \sum_{j \neq i} P_{ji}^2 \tilde{m}_i^2 / f_i^2 + n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} P_{ji} P_{li} \tilde{m}_i^2 / f_i^2 + (s.o.) \\ &= D_3 + D_4 + (s.o.), \end{aligned}$$

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

$$D_3 \leq \sup_{1 \leq i \leq n} (\tilde{m}_i^2 / f_i^2) n^{-3} \sum_i \sum_{j \neq i} P_{ji}^2 = O_p \left( h^4 + \frac{\ln(n)}{nh} \right) O_p((nh^3)^{-1})$$

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

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  $P_{ji} P_{li} = [h^2 G_1(x_i) + e_{ji}][h^2 G_1(x_i) + e_{li}] + (s.o.)$ , 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

$$\begin{aligned} D_{4,1} &= n^{-3} h^4 \sum_i \sum_{j \neq i} \sum_{l \neq i, j} G_1(x_i)^2 \tilde{m}_i^2 / f_i^2 = h^4 \left[ n^{-1} \sum_i G_1(x_i)^2 \tilde{m}_i^2 / f_i^2 \right] \\ &= h^4 O_p(h^4 + (nh)^{-1}) \end{aligned} \tag{A.12}$$

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

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

by lemma B.4.

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

Hence, we have shown that

$$C_{2,1} = D_4 + (s.o.) = o_p((nh)^{-1}) = o_p((nh^3)^{-1}). \tag{A.13}$$

Next, we consider  $C_{2,2}$ . Define  $F_{ji} = (x_j - x_i) P_{ji}$ . We have

$$\begin{aligned} C_{2,2} &= n^{-3} \sum_i \sum_{j \neq i} F_{ji}^2 (\tilde{\beta}_i - \beta_i)^2 + n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} F_{ji} F_{li} (\tilde{\beta}_i - \beta_i)^2 \\ &= D_5 + D_6, \end{aligned}$$

where the definitions of  $D_5$  and  $D_6$  should be obvious.

$$\begin{aligned} D_5 &\leq n^{-1} \left[ \sup_{1 \leq i \leq n} (\tilde{\beta}_i - \beta_i)^2 / f_i^2 \right] \left[ n^{-3} \sum_i \sum_{j \neq i} F_{ji}^2 \right] \\ &= n^{-1} O_p \left( h^4 + \frac{\ln(n)}{nh^3} \right) O_p(h^{-1}) = O_p \left( (nh)^{-1} \left( h^4 + \frac{\ln(n)}{nh^3} \right) \right) \end{aligned}$$

by lemma B.1 (iv) and the fact that  $\sup_{1 \leq i \leq n} (\tilde{\beta}_i - \beta_i)^2 / f_i^2 = O_p \left( h^4 + \frac{\ln(n)}{nh^3} \right)$ .

Define  $\eta_{ji} = E(F_{ji}|x_i)$ , 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  $D_{6,1} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i,j} E(F_{ji}|x_i) E(F_{li}|x_i) (\tilde{\beta}_i - \beta_i)^2$ ,  $D_{6,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i,j} \eta_{ji} \eta_{li} (\tilde{\beta}_i - \beta_i)^2$  and  $D_{6,3} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i,j} E(F_{ji}|x_i) \eta_{li} (\tilde{\beta}_i - \beta_i)^2$ .

We first consider  $D_{6,1}$ . By lemma B.1 (ii) 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

$$\begin{aligned} D_{6,1} &= n^{-1} \sum_i (h^2 B_{2i} + J_{3i})^2 + (s.o.) = n^{-1} \sum_i (h^4 B_{2i}^2 + J_{3i}^2 + 2h^2 B_{2i} J_{3i}) + (s.o.) \\ &= D_7 + D_8 + 2D_9 + (s.o.), \end{aligned}$$

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

$$D_7 = n^{-1} h^4 \sum_i B_2(x_i)^2 + (s.o.) = h^4 E[B_2(x_i)^2] + (s.o.).$$

Next, we consider  $D_8$ . Using  $J_{3i} = n^{-1} \sum_{j \neq i} u_j W'_{h,ji} / f_i$ .

$$\begin{aligned} D_8 &= \frac{1}{n} \sum_i J_{3i}^2 = \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} u_j W'_{h,ji} u_l W'_{h,li} / f_i^2 \\ &= \frac{1}{n^3} \sum_i \sum_{j \neq i} u_j^2 (W'_{h,ji})^2 / f_i^2 + \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i,j} u_j W'_{h,ji} u_l W'_{h,li} / f_i^2 \\ &= V_1 + V_2, \end{aligned}$$

where the definitions of  $V_2$  and  $V_3$  should be obvious.

By the H-decomposition we get

$$\begin{aligned} V_1 &= E(V_1) + (s.o.) = \frac{1}{n} E[u_j^2 (W'_{h,ji})^2 / f_i^2] \\ &= \frac{1}{nh^4} \int \sigma^2(x_j) \left( W' \left( \frac{x_j - x_i}{h} \right) \right)^2 f_i^{-1} f_j dx_i dx_j + (s.o.) \\ &= \frac{\kappa_0}{nh^3} \int \sigma^2(x) dx + (s.o.), \end{aligned}$$

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

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

Hence, we have

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

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

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

Finally,  $D_9 = (h^2/n^2) \sum_i \sum_{j \neq i} u_j W'_{h,ji} / f_i$ . 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

$$D_{6,1} = D_7 + V_1 + (s.o.) = h^4 \int B_2^2(x) M(x) dx + \frac{1}{nh^3} \kappa_0 \int \sigma^2(x) M(x) dx + (s.o.) \quad (\text{A.14})$$

$D_{6,2} = o_p(h^4 + (nh^3)^{-1})$  and  $D_{6,3} = o_p(h^4 + (nh^3)^{-1})$  by lemma B.5. Hence, we have shown that (adding back the weight function  $M(\cdot)$ )

$$\begin{aligned} C_{2,2} &= D_5 + D_6 = D_{6,1} + (s.o.) \\ &= h^4 \int B_2^2(x) M(x) dx + \frac{1}{nh^3} \kappa_0 \int \sigma^2(x) M(x) dx + (s.o.). \end{aligned} \quad (\text{A.15})$$

Finally, by (A.13), (A.15) 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

$$C_2 = C_{2,2} + (s.o.) = h^4 \int B_2(x)^2 f(x) M(x) dx + \frac{1}{nh^3} \kappa_0 \int \sigma^2(x) M(x) dx + (s.o.).$$

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.  $\square$

**Lemma A.3.**  $C_3 = o_p(h^4 + (nh^3)^{-1})$  uniformly in  $h \in H_n$ .

*Proof of Lemma A.3:* Recall that  $F_{ji} = P_{ji}(x_j - x_i)$ . Then using similar arguments as we did in the proofs of lemmas A.1 and A.2, we have

$$\begin{aligned}
C_3 &= n^{-1} \sum_i e_{\beta}^T (A_{2i}^{-1} A_{1i}) (e_{\beta}^T A_{2i}^{-1} \Delta_i) \\
&= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji} (R_{ji} + u_j) F_{li} (\tilde{\beta}_l - \beta_l) \\
&\sim n^{-3} h^2 \sum_i \sum_{j \neq i} \sum_{l \neq i} P_{ji} (R_{ji} + u_j) (\tilde{\beta}_l - \beta_l) \\
&= n^{-3} h^2 \sum_i \sum_{j \neq i} \sum_{l \neq i} W_{h,ji} [h^{-2} C_{2i}(x_j - x_i) - C_{1i}] (R_{ji} + u_j) (\tilde{\beta}_l - \beta_l) \\
&\sim n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} W_{h,ji} (x_j - x_i) (R_{ji} + u_j) (\tilde{\beta}_l - \beta_l) \\
&= C_{3,1} + C_{3,2},
\end{aligned}$$

where  $C_{3,1} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} W_{h,ji} (x_j - x_i) R_{ji} (\tilde{\beta}_l - \beta_l)$  and  $C_{3,2} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} W_{h,ji} (x_j - x_i) u_j (\tilde{\beta}_l - \beta_l)$ , the third line (with a ‘ $\sim$ ’ sign) above follows from  $E(F_{ji}|x_i) = 1 + O(h^2)$ , and the fifth line (with a ‘ $\sim$ ’ sign) holds because  $C_{1i}$  has an order no larger than that of  $h^{-2} C_{2i}(x_j - x_i)$ , and then we replaced  $C_{2i}$  by 1.

Using the U-statistic H-decomposition, it is easy to show that the leading term of  $C_{3,1}$  can be obtained by replacing  $W_{h,ji}(x_j - x_i)R_{ji}$  by  $E[W_{h,ji}(x_j - x_i)R_{ji}|x_i] = \mu_4 h^4 H(x_i) + O_p(h^6)$ , where  $H(x_i) = (1/6)f(x_i)g'''(x_i) + (1/2)f'(x_i)g''(x_i)$ . Then it is easy to show that  $C_{3,1} = O_p(h^4(h^2 + (\ln(n))^{1/2}(nh^3)^{-1/2})) = o_p(h^4)$ . Similarly, one can show that  $C_{3,2} = o_p(h^4 + (nh^3)^{-1})$ .  $\square$

## APPENDIX B. SOME USEFUL LEMMAS (NOT FOR PUBLICATION)

Appendix B is for referees' convenience, not for publication. This appendix is also available to readers from the authors upon request.

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  $k^{\text{th}}$  order U-statistic  $U_{(k)}$  is defined by*

$$U_{(k)} = \frac{1}{C_n^k} \sum_{1 \leq i_1 < \dots < i_k} H_n(x_{i_1}, \dots, x_{i_k}), \quad (\text{B.1})$$

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

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

$$U_{(k)} = \sum_{j=0}^k C_n^k H_n^{(j)} = C_k^0 H_n^{(0)} + C_k^1 H_n^{(1)} + C_k^2 H_n^{(2)} + \dots + C_k^k H_n^{(k)}, \quad (\text{B.2})$$

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

When  $\theta \neq 0$ ,  $H_n^{(0)} = \theta$  is usually the leading term of the U-statistic. If  $\theta = 0$  and  $H_{ni} \neq 0$ , usually the partial sum  $H_n^{(1)}$  becomes the leading term of  $H_{(k)}$ . If both  $\theta = 0$  and  $H_{ni} = 0$ , 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  $P_{ji} = W_{h,ji}[C_{2i}h^{-2}(x_j - x_i) - C_{1i}]$  and  $F_{ji} = (x_j - x_i)P_{ji}$ .*

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

$$(ii) E(P_{ji}|x_i) = G_2(x_i)h^2 + O(h^3) \text{ uniformly in } x_i \in \mathcal{M}, \text{ where } G_2(x_i) = \frac{\mu_4 f'''(x_i)}{6\mu_2 f(x_i)} - \frac{\mu_2 f''(x_i)f'(x_i)}{2 f^2(x_i)}.$$

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

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

**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

$$\begin{aligned} E(F_{ji}|x_i) &= E\{W_{h,ji}(x_j - x_i)[C_{2i}h^{-2}(x_j - x_i) - C_{1i}] | x_i\} \\ &= C_{2i}h^{-2}E[W_{h,ji}(x_j - x_i)^2 | x_i] - C_{1i}E[(x_j - x_i)W_{h,ji} | x_i] \\ &= 1 - h^2 \mu_2 f'(x_i)^2 / f(x_i)^2 + O(h^3) \end{aligned}$$

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

$$\begin{aligned} E(W_{h,ji}|x_i) &= f(x_i) + \frac{1}{2}h^2\mu_2 f''(x_i) + O_p(h^4), \\ h^{-2}E[(x_j - x_i)W_{h,ji}|x_i] &= h^{-2}[0 + \frac{1}{2}h^2 f'(x_i)\mu_2 + 0 + \frac{1}{6}h^4\mu_4 f'''(x_i) + O(h^6)] \\ &= f'(x_i)\mu_2 + \frac{1}{6}h^2\mu_4 f'''(x_i) + O(h^4). \end{aligned}$$

Hence, we have

$$\begin{aligned} E(P_{ji}|x_i) &= E[W_{h,ji}(C_{2i}h^{-2}(x_j - x_i) - C_{1i})|x_i] \\ &= C_{2i}h^{-2}E[W_{h,ji}(x_j - x_i)|x_i] - C_{1i}E[W_{h,ji}|x_i] \\ &= \frac{f'(x_i)}{f(x_i)} + \frac{h^2\mu_4 f'''(x_i)}{6\mu_2 f(x_i)} - \frac{f'(x_i)}{f(x_i)} - \frac{h^2\mu_2 f''(x_i)f'(x_i)}{2 f^2(x_i)} + O_p(h^3) \\ &= G_2(x_i)h^2 + O(h^3) \end{aligned} \tag{B.3}$$

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.,  $L_{1n,1} = n^{-2} \sum_i \sum_{j \neq i} W_{h,ji}^2 C_{2i} h^{-4} (x_j - x_i)^2$ . 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 (iv):** 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  $C_{1i}$  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  $J_{4n,2} = n^{-2} h^{-4} \sum_i \sum_{j \neq i} W_{h,ji}^2 (x_j - x_i)^4$ . 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  $\tilde{\beta}(x)$ , the local constant of  $\beta(x)$ . The local constant estimate of  $g(x)$  is given by

$$\tilde{g}(x) = \frac{n^{-1} \sum_j y_j W_{h,jx}}{n^{-1} \sum_j W_{h,jx}},$$

where  $W_{h,jx} = h^{-1} W((x_j - x)/h)$ ,  $y_j = g_j + u_j$ ,  $g_j = g(x_j)$ .

For  $W(\frac{x_j - x}{h})$ , let  $u = (x_j - x)/h$ , we have

$$W'_{h,jx} = \frac{\partial W_h((x_j - x)/h)}{\partial x} = \frac{1}{h} \frac{\partial W(u)}{\partial u} \frac{\partial u}{\partial x} = -\frac{1}{h^2} W' \left( \frac{x_j - x}{h} \right), \tag{B.4}$$

where  $W' \left( \frac{x_j - x}{h} \right) = W'(u)|_{u=(x_j - x)/h}$  and  $W'(u) = \partial W(u)/\partial u$ .

Using (B.4) we obtain the local constant first derivative estimator of  $\beta(x)$  given by

$$\begin{aligned}
\tilde{\beta}(x) &= \frac{n^{-1} \sum_j y_j W'_{h,jx}}{n^{-1} \sum_l W_{h,lx}} - \frac{(n^{-1} \sum_l y_l W_{h,lx})(n^{-1} \sum_j W'_{h,jx})}{(n^{-1} \sum_l W_{h,lx})^2} \\
&= \frac{\frac{1}{n^2} \sum_j \sum_{l \neq j} (y_j - y_l) W'_{h,jx} K_{h,lx}}{(\frac{1}{n} \sum_l K_{h,lx})^2} \\
&= \frac{\frac{1}{n^2} \sum_j \sum_{l \neq j} (g_j - g_l + u_j - u_l) W'_{h,jx} W_{h,lx}}{f(x)^2} + (s.o.) \\
&= J_1(x) + J_2(x) + (s.o.), \tag{B.5}
\end{aligned}$$

where

$$\begin{aligned}
J_1(x) &= \frac{1}{n^2} \sum_j \sum_{l \neq j} (g_j - g_l) W'_{h,jx} W_{h,lx} / f(x)^2, \\
J_2(x) &= \frac{1}{n^2} \sum_j \sum_{l \neq j} (u_j - u_l) W'_{h,jx} W_{h,lx} / f(x)^2.
\end{aligned}$$

**Lemma B.2.** (i)  $\tilde{\beta}(x) = \beta(x) + h^2 B_2(x) + J_3(x) + O_p(h^3 + (nh^2)^{-1/2})$ , 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  $J_3(x) = nh^{-1} \sum_j u_j W'_{h,jx} / f(x)$ .

(ii)  $\tilde{\beta}_i(x_i) = \beta(x_i) + h^2 B_{2i} + J_{3i} + O_p(h^3 + (nh^2)^{-1/2})$ ,  $B_{2i} = B_2(x_i)$  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  $W'_{h,ji} = -h^{-2} W' \left( \frac{x_j - x}{h} \right)$ , we get

$$\begin{aligned}
E[J_1(x)] &= \frac{1}{h^3 f(x)^2} \int \int [g(x_l) - g(x_j)] W' \left( \frac{x_j - x}{h} \right) W \left( \frac{x_l - x}{h} \right) f(x_j) f(x_l) dx_j dx_l \\
&= \frac{1}{h f(x)^2} \int \int [g(x + hv) - g(x + hu)] W'(u) W(v) f(x + hu) f(x + hv) dudv \\
&= \frac{1}{h f(x)^2} \int \int [(g'(x)hv + \frac{1}{2} g''(x)h^2 v^2 - g'(x)hu - \frac{1}{2} g''(x)h^2 u^2 - \frac{1}{6} g'''(x)h^3 u^3) \\
&\quad [f(x) + f'(x)hu + \frac{1}{2} f''(x)h^2 u^2] [f(x) + f'(x)hv + \frac{1}{2} f''(x)h^2 v^2] W'(u) W(v) dudv + O(h^3) \\
&= g'(x) + B_2(x)h^2 + O(h^3),
\end{aligned}$$

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,  $\int u W'(u) du = u W(u) |_{-\infty}^{\infty} - \int W(u) du = -1$  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

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

Next, we consider  $J_2(x)$ . Define  $J_{2,1} = J_2(x)f(x)^2$ . Then  $J_{2,1}(x) = (2/n^2) \sum_j \sum_{l>j} (1/2)[(u_j - u_l)W'_{h,jx}W_{h,lx} + (u_l - u_j)W'_{h,lx}W_{h,jx}] \equiv (2/n^2) \sum_j \sum_{l \neq j} H_n(z_j, z_l)$ , which is a second-order U-statistic, where  $H_n(z_j, z_l) = (1/2)[(u_j - u_l)W'_{h,jx}W_{h,lx} + (u_l - u_j)W'_{h,lx}W_{h,jx}]$  with  $z_j = (u_j, x_j)$ .

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

$$\begin{aligned} H_{nj} &= E[H_n(z_j, z_l)|z_j] \\ &= (1/2)u_j[W'_{h,jx}E(W_{h,lx}) - W_{h,jx}E(W'_{h,lx})] \\ &= (1/2)u_j \left\{ W'_{h,jx}[f(x) + O_p(h^2)] - [W_{h,jx}\mu_2 f'(x) + O_p(h^2)] \right\} \\ &= (1/2)u_j W'_{h,jx}f(x) + (s.o.) \end{aligned}$$

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  $W_{h,jx} = h^{-1}W(\frac{x_j-x}{h})$ .

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

$$J_2(x) = \frac{2}{nf(x)} \sum_j H_{nj} + (s.o.) = \frac{1}{n} \sum_j u_j W'_{h,jx} + (s.o.) = J_3(x) + (s.o.),$$

where  $J_3(x) = n^{-1} \sum_j u_j W'_{h,jx}$ .

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

$$\begin{aligned} E[J_3(x)^2] &= \frac{1}{nf(x)^2} E[\sigma_j^2 W_{h,jx}^{\prime 2}] \\ &= \frac{1}{nh^4 f(x)^2} \int \sigma^2(x_j) \left[ W' \left( \frac{x_j - x}{h} \right) \right]^2 f(x_j) dx_j \\ &= \frac{1}{nh^4 f(x)^2} h \int \sigma^2(x + hv) [W'(v)]^2 f(x + hv) dv \\ &= \frac{\kappa_0}{nh^3 f(x)} \sigma^2(x) + O((nh)^{-1}), \end{aligned}$$

where  $\kappa_0 = \int [W'(v)]^2 dv$  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  $\tilde{\beta}(x)$  (with  $h \rightarrow 0$ ,  $nh^3 \rightarrow \infty$  and  $nh^9 \rightarrow 0$  as  $n \rightarrow \infty$ ):

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

□

**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  $\tilde{m}_i = (\tilde{g}_i - g_i) \hat{f}_i$ . Then*

$$\mathcal{D} = h^4 E[G(x_i)C(x_i)^2] + \frac{1}{nh} \nu_0 E[G(x_i)\sigma^2(x_i)f(x_i)] + (s.o.),$$

where  $C(x_i) = \frac{1}{2}\mu_2[g''(x_i)f(x_i) + 2g'(x_i)f'(x_i)]$ ,  $\mu_2 = \int v^2 W(v)dv$  and  $\nu_0 = \int W^2(v)dv$ .

**Proof of Lemma B.3:** Using (A.11), we have

$$\begin{aligned}
\mathcal{D} &= \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i} G(x_i)(g_j - g_i + u_j)W_{h,ji}(g_l - g_i + u_j)W_{h,li} \\
&= \frac{1}{n^3} \sum_i \sum_{j \neq i} G(x_i)(g_j - g_i + u_j)^2 W_{h,ji}^2 \\
&\quad + \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} G(x_i)(g_j - g_i + u_j)W_{h,ji}(g_l - g_i + u_j)W_{h,li} \\
&= \mathcal{D}_1 + \mathcal{D}_2,
\end{aligned}$$

where the definitions of  $\mathcal{D}_l$  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

$$\begin{aligned}
\mathcal{D}_1 &= \frac{1}{n^3} \sum_i \sum_{j \neq i} G(x_i)u_j^2 W_{h,ji}^2 + (s.o.) \\
&= n^{-1}E[G(x_i)\sigma^2(x_j)W_{h,ji}^2] + (s.o.) \\
&= n^{-1}h^{-1}\nu_0 \left[ \int G(x_i)\sigma^2(x_i)f(x_i)^2 dx_i \right] + (s.o.) \\
&= \frac{1}{nh}\nu_0 E[G(x_i)\sigma^2(x_i)f(x_i)] + (s.o.),
\end{aligned}$$

where the second equality follows from the U-statistic H-decomposition, and  $\nu_0 = \int W^2(v)dv$ .

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

$$\begin{aligned}
\mathcal{D}_2^0 &= \frac{1}{n^3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} G(x_i)(g_j - g_i)W_{h,ji}(g_l - g_i)W_{h,li} \\
&= E\{G(x_i)(g_j - g_i)W_{h,ji}(g_l - g_i)W_{h,li}\} + (s.o.) \\
&= E\{G(x_i)[E((g_j - g_i)W_{h,ji}|x_i)]^2\} + (s.o.) \\
&= h^4 E[G(x_i)C(x_i)^2] + (s.o.),
\end{aligned}$$

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  $\mu_2 = \int v^2 W(v)dv$ .  $\square$

**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

$$D_{4,2} = O_p((nh^2)^{-2}) = o_p((nh)^{-1}).$$

**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.11) we have

$$\begin{aligned}
D_{4,2} &= \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \sum_{k \neq i} \sum_{m \neq i} e_{ji}e_{li}W_{h,ki}W_{h,mi} [(g_k - g_i)(g_m - g_i) + u_k u_m + 2u_k(g_m - g_i)] / f_i^2 \\
&= \mathcal{D}_3 + \mathcal{D}_4 + 2\mathcal{D}_5,
\end{aligned} \tag{B.7}$$

where  $\mathcal{D}_3 = \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \sum_{k \neq i} \sum_{m \neq i} e_{ji} e_{li} W_{h,ki} W_{h,mi} (g_k - g_i)(g_m - g_i) / f_i^2$ ,  $\mathcal{D}_4 = \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \sum_{k \neq i} \sum_{m \neq i} e_{ji} e_{li} W_{h,ki} W_{h,mi} u_k u_m / f_i^2$  and  $\mathcal{D}_5 = \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \sum_{k \neq i} \sum_{m \neq i} e_{ji} e_{li} W_{h,ki} W_{h,mi} u_k (g_m - g_i) / f_i^2$ .

We consider  $\mathcal{D}_3$ . We first evaluate  $E(\mathcal{D}_3)$ . 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  $E(\mathcal{D}_3) = 0$ . Hence, for  $E(\mathcal{D}_3) \neq 0$ , we must have  $k$  and  $m$  to match  $j$  and  $l$ , say  $k = j$  and  $m = l$ . Therefore, the leading term of  $E(\mathcal{D}_3)$  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  $\mathcal{D}_3$ , it is sufficient to consider  $e_{ji} \sim P_{ji} \sim h^{-2} W_{h,ji}(x_j - x_i)$ . Then, we have

$$\begin{aligned} E[e_{ji} W_{h,ji}(g_j - g_i)|x_i] &\sim h^{-2} E[W_{h,ji}^2(x_j - x_i)(g_j - g_i)|x_i] \\ &= h^{-2} \int h^{-2} W^2(v) h v [g'_i h v + O(h^2)] [f_i + O(h)] h d v \\ &= h^{-2} \nu_2 g'_i f_i + O(1) = O(h^{-2}) \end{aligned}$$

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

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

$$E(\mathcal{D}_3) = n^{-5} n^3 O(h^{-4}) = O((nh^2)^{-2}).$$

Similar arguments lead to  $Var(\mathcal{D}_3) = O((nh^2)^{-4})$ . Hence,

$$\mathcal{D}_3 = O_p((nh^2)^{-1}).$$

Next, we consider  $\mathcal{D}_4$ . It is easy to see that the leading term of  $\mathcal{D}_4$  corresponds to the case that the summation indices  $i, j, l, k, m$  all different from each other. We use  $\mathcal{D}_{4,1}$  to denote this leading term of  $\mathcal{D}_4$ . Obviously,  $E(\mathcal{D}_{4,1}) = 0$ . 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  $E(\mathcal{D}_{4,1}^2) = n^{-8} n^4 O(h^{-8}) = O((nh^2)^{-4})$ . Hence,  $\mathcal{D}_{4,1} = O_p((nh^2)^{-2})$ .

Similarly, one can show that  $\mathcal{D}_5 = O_p((nh^2)^{-2})$ .  $\square$

**Lemma B.5.** *Let  $D_{6,2}$  and  $D_{6,3}$  be as defined in lemma A.2. Then we have (i)  $D_{6,2} = o_p(n^{-1}) = o_p((nh^3)^{-1})$ ; (ii)  $D_{6,3} = o_p(h^4 + (nh^3)^{-1})$ ,*

**Proof of Lemma (i):** Using  $\tilde{\beta}_i - \beta_i = h^2 B_{2i} + J_{3i} + (s.o.)$ , we have

$$\begin{aligned} D_{6,2} &= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \eta_{ji} \eta_{li} (\tilde{\beta}_i - \beta_i)^2 \\ &= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \eta_{ji} \eta_{li} (h^4 B_{2i}^2 + J_{3i}^2 + 2h^2 B_{2i} J_{3i}) + (s.o.) \\ &= \mathcal{D}_6 + \mathcal{D}_7 + 2\mathcal{D}_8 + (s.o.), \end{aligned}$$

where  $\mathcal{D}_6 = (h^4/n^3) \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \eta_{ji} \eta_{li} B_{2i}^2$ . From  $E(\eta_{ji}|x_i) = 0$  and  $E(\eta_{ji}^2) \sim E(F_{ji}^2) \sim h^{-4} E[W_{h,ji}^2(x_j - x_i)^4] = h^{-4} O(h^3) = O(h^{-1})$ , it is easy to show that

$$E(\mathcal{D}_6^2) = (h^8/n^6) O(n^4 h^{-2}) = O(h^6/n^2).$$

Hence,  $\mathcal{D}_6 = O_p(h^3/n) = o_p(n^{-1})$ .

$$\mathcal{D}_7 = n^{-5} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \sum_{k \neq i} \sum_{m \neq i} \eta_{ji} \eta_{li} u_k u_m W'_{h,ki} W'_{h,mi} / f_i^2.$$

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

$$E(\mathcal{D}_7^2) = n^{-10} O(n^6 h^{-8}) = O((nh^2)^{-4} n^{-2}) = o(n^{-2}).$$

Hence,  $\mathcal{D}_7 = o_p(n^{-1})$ .

Similarly, one can show that  $\mathcal{D}_8 = o_p(n^{-1})$ . Combining the above results we obtain

$$\mathcal{D}_{6,2} = \mathcal{D}_6 + \mathcal{D}_7 + 2\mathcal{D}_8 + (s.o.) = o_p(n^{-1}).$$

This completes the proof of (i). □

**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

$$\begin{aligned} \mathcal{D}_{6,3} &\sim n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \eta_{li} (\tilde{\beta}_i - \beta_i)^2 \\ &\sim n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \eta_{li} [h^4 B_{2i}^2 + J_{3i}^2 + 2h^2 B_{2i} J_{3i}] \\ &= \mathcal{D}_9 + \mathcal{D}_{10} + \mathcal{D}_{11} + (s.o.), \end{aligned}$$

where  $\mathcal{D}_9 = (h^4/n^3) \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \eta_{li} B_{2i}^2$ . Its second moment is

$$E(\mathcal{D}_9^2) \sim (h^8/n^6) n^5 O(E(\eta_{ji}^2)) = (h^8/n) O(h^{-1}) = O(h^7/n) = o(h^8).$$

Hence,  $\mathcal{D}_9 = o_p(h^4)$ .

$\mathcal{D}_{10} = n^{-5} \sum_i \sum_{j \neq i} \sum_{l \neq i, j} \sum_{k \neq i} \sum_{m \neq i} \eta_{li} u_k u_m W'_{h,ki} W'_{h,mi} / f_i^2$ . From  $E(\eta_{ji}^2) = O(h^{-1})$  and  $E((W'_{h,ki})^2) = O(h^{-3})$ , we get

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

Hence,  $E(\mathcal{D}_{10}) = o_p((nh^3)^{-1})$ .

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

Summarizing the above we have shown that

$$\mathcal{D}_{6,3} = \mathcal{D}_9 + \mathcal{D}_{10} + \mathcal{D}_{11} + (s.o.) = o_p(h^4 + (nh^3)^{-1}).$$

□

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

$$\sqrt{nh_1 \dots h_d} D_h (\hat{\beta}(x) - \beta(x) - B_{1,h}(x)) \xrightarrow{d} N(0, c_U(x) I_d),$$

where  $B_{1,h}(x)$  is a  $d \times 1$  vector with its  $s^{\text{th}}$  component given by  $B_{1s,h}(x) = \frac{h_s^2 \mu_4 g_s'''(x)}{6\mu_2} + \frac{h_s^2 g_s''(x) f_s'(x) \mu_4}{2\mu_2 f(x)} - \frac{\mu_2 f_s(x) \sum_{t=1}^d g_{tt}(x) h_t^2}{2f(x)^2}$ ,  $c_U(x) = \nu_0^{d-1} \nu_2 \sigma^2(x) / (\mu_2^2 f(x))$  and  $I_d$  is  $n \times n$  identity matrix.

**Proof of Lemma B.6:** By (14) we know that  $\hat{\beta}(x) = \beta(x) + e_\beta^T A_{2x}^{-1} A_{1x}$ , where  $A_{1x}$  and  $A_{2x}$  are obtained from  $A_{1i}$  and  $A_{2i}$  with  $x_i$  replaced by  $x$ , and  $\sum_{j \neq i}^n$  replaced by  $\sum_j = \sum_{j=1}^n$ . Ny the same derivation that leads to (A.5), we have  $e_\beta^T A_{2x}^{-1} A_{1x} = n^{-1} \sum_j W_{h,jx} [\frac{1}{\mu_2 f(x)} D_h^{-2} (x_j - x) - f'(x)/f(x)^2] (R_{jx} + u_j) + (s.o.) \equiv \mathcal{A}_{1n}(x) + \mathcal{A}_{2n}(x)$ , where  $D_h^{-2} = \text{diag}(h_s^{-2})$ ,  $R_{jx} = g(x_j) - g(x) - (g'(x))^T (x_j - x)$ ,  $\mathcal{A}_{1n}(x) = n^{-1} \sum_j W_{h,jx} [\frac{1}{\mu_2 f(x)} D_h^{-2} (x_j - x) - f'(x)/f(x)^2] R_{jx}$  and  $\mathcal{A}_{2n}(x) = n^{-1} \sum_j W_{h,jx} [\frac{1}{\mu_2 f(x)} D_h^{-2} (x_j - x) - f'(x)/f(x)^2] u_j$ . Let  $\mathcal{A}_{1n,s}(x)$  be the  $s^{\text{th}}$  component of  $\mathcal{A}_{1n}(x)$ . Note that the  $s^{\text{th}}$  component of  $D_h^{-2} (x_j - x) - \frac{f'(x)}{f(x)^2}$  is  $h_s^{-2} (x_{js} - x_s) - \frac{f_s'(x)}{f(x)^2}$ , where  $f_s(x) = \frac{\partial f(x)}{\partial x_s}$ . We have

$$\begin{aligned} E(\mathcal{A}_{1n,s}(x)) &= \int W_{h,jx} \left[ \left( \frac{1}{\mu_2 f(x)} \right) h_s^{-2} (x_{js} - x_s) - \frac{f_s'(x)}{f(x)^2} \right] R_{jx} f(x_j) dx_j \\ &= \int W(v) \left[ \left( \frac{1}{\mu_2 f(x)} \right) \frac{v_s}{h_s} - \frac{f_s'(x)}{f(x)^2} \right] \left[ \frac{1}{2} \sum_{l_1=1}^d \sum_{l_2=1}^d g_{l_1, l_2}(x_i) h_{l_1} h_{l_2} v_{l_1} v_{l_2} \right. \\ &\quad \left. + \frac{1}{6} \sum_{l_1=1}^d \sum_{l_2=1}^d \sum_{l_3=1}^d g_{l_1, l_2, l_3}(x_i) h_{l_1} h_{l_2} h_{l_3} v_{l_1} v_{l_2} v_{l_3} + O(\|h\|^4) \right] \\ &\quad \left[ f(x) + \sum_{t=1}^d f_t(x) h_t v_t + O(\|h\|^2) \right] \\ &= \frac{h_s^2 \mu_4 g_{sss}(x)}{6\mu_2} + \frac{h_s^2 g_{ss}(x) f_s'(x) \mu_4}{2\mu_2 f(x)} - \frac{\mu_2 f_s(x) \sum_{t=1}^d g_{tt}(x) h_t^2}{f(x)^2} \\ &\quad + \frac{1}{2} \frac{f_s(x) \sum_{t \neq s} h_t^2 g_{tt}(x)}{2f(x)} + \mu_2 \sum_{t \neq s} h_t^2 f_t(x) g_{ts}(x) / f(x) + (\mu_2/2) \sum_{t \neq s} h_t^2 g_{tts}(x) + O(\|h\|^3) \\ &= B_{1s,h}(x) + O(\|h\|^3). \end{aligned}$$

It is easy to show that  $\text{Var}(\mathcal{A}_{1n}(x)) = O((nh_1 \dots h_d \|h\|)^{-1})$ . Hence, we have that

$$\mathcal{A}_{1n}(x) = B_{1,h}(x) + O_p(\|h\|^3 + (nh_1 \dots h_d \|h\|)^{-1/2}).$$

Next,  $E(\mathcal{A}_{2n}(x)) = 0$  and

$$\begin{aligned} \text{Var}(\mathcal{A}_{2n}(x)) &= \frac{1}{n\mu_2^2 f(x)^2} \int f(x_j) \sigma^2(x_j) W_{h,jx}^2 D_h^{-2} (x_j - x)(x_j - x)^T D_h^{-2} dx_j \\ &= \frac{1}{nh_1 \dots h_d \mu_2^2 f(x)^2} \int (f\sigma^2)(x + hv) W(v)^2 D_h^{-2} D_h v v^T D_h D_h^{-2} dv \\ &= \frac{\nu_0^{d-1} \nu_2 \sigma^2(x)}{nh_1 \dots h_d \mu_2^2 f(x)} D_h^{-2} + O((nh_1 \dots h_d)^{-1}), \end{aligned}$$

where in the last equality we used  $\int W(v)^2 v v^T dv = \nu_0^{d-1} \nu_2 I_d$ .

Summarizing the above results and under the condition  $nh_1 \dots h_d \|h\|^2 \rightarrow \infty$ ,  $nh_1 \dots h_d \|h\|^4 \rightarrow 0$ , and by applying Lyapunov's central limit theorem, we have

$$(nh_1 \dots h_d)^{1/2} D_h (\hat{\beta}(x) - \beta(x) - B_{1,h}(x)) \xrightarrow{d} N(0, c_{0,l}(x) I_d),$$

where  $c_{0,l}(x) = \frac{\nu_0^{d-1} \nu_2 \sigma^2(x)}{\mu_2^2 f(x)}$  and  $D_h = \text{diag}(h_s)$ . □

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

$$\sqrt{nh_1 \dots h_d} D_h (\tilde{\beta}(x) - \beta(x) - B_{2,h}(x)) \xrightarrow{d} N(0, c_{lc}(x) I_d),$$

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

$$\begin{aligned} B_{2s,h}(x) &= \frac{\mu_2}{f(x)} \left[ \frac{1}{2} g_{sss}(x) f(x) h_s^2 + \sum_{t=1}^d [f_t(x) g_{ts}(x) + g_t(x) f_{ts}(x)] h_t^2 - \frac{f_s(x)}{f(x)} \sum_{t=1}^d g_t(x) f_t(x) h_t^2 \right. \\ &\quad \left. + g_s(x) \sum_{t=1}^d f_{tt}(x) h_t^2 \right] \text{ with } m_t(x) = \frac{\partial m(x)}{\partial x_s}, m_{ts}(x) = \frac{\partial^2 m(x)}{\partial x_t \partial x_s}, m_{sss}(x) = \frac{\partial^3 m(x)}{\partial x_s^3} \text{ for } m = g \text{ or } f, \\ c_{lc}(x) &= \nu_0^{d-1} \kappa_0 \sigma^2(x) / f(x). \end{aligned}$$

**Proof of Lemma B.7:** For the multivariate  $x$  case,  $\tilde{\beta}(x)$  is still defined as in (B.5) with  $J_1(x)$  and  $J_2(x)$  defined below (B.5), except that now  $W_{h,lx} = \prod_{s=1}^d h_s^{-1} w(\frac{x_{ls} - x_s}{h_s})$  is a product kernel, and  $W'_{h,jx}$  is a  $d \times 1$  vector with its  $s^{\text{th}}$  component given by  $W'_{h,jx,s} = -h_s^{-2} w'(\frac{x_{js} - x_s}{h_s}) \prod_{t \neq s} h_t^{-1} w(\frac{x_{jt} - x_t}{h_t})$ . We use  $J_{1,s}(x)$  to denote the  $s^{\text{th}}$  component of  $J_1(x)$ , then it is straightforward, though tedious, to show that  $(g_s(x) = \frac{\partial g(x)}{\partial x_s})$ .

$$\begin{aligned} E(J_{1,s}(x)) &= g_s(x) + \frac{\mu_2}{f(x)} \left[ \frac{1}{2} g_{sss}(x) f(x) h_s^2 + \sum_{t=1}^d (f_t(x) g_{ts}(x) + f_t(x) f_{ts}(x)) h_t^2 \right. \\ &\quad \left. - \frac{f_s(x)}{f(x)} \sum_{t=1}^d g_t(x) f_t(x) h_t^2 + g_s(x) \sum_{t=1}^d f_{tt}(x) h_t^2 \right] + O(\|h\|^3) \\ &= g_s(x) + B_{2s,h}(x) + O(\|h\|^3). \end{aligned} \tag{B.8}$$

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

$$J_1(x) = \beta(x) + B_{2,h}(x) + O_p(\|h\|^3 + (nh_1 \dots h_d \|h\|)^{1/2}).$$

It is easy to see that  $W'_{h,jx} = -\frac{1}{h_1 \dots h_d} D_h^{-1} W'(\frac{x_j - x}{h})$ , where  $W'(\frac{x_j - x}{h})$  is a  $d \times 1$  vector with its  $s^{\text{th}}$  component given by  $w'(\frac{x_{js} - x_s}{h_s}) \prod_{t \neq s} w(\frac{x_{jt} - x_t}{h_t})$ . Similarly, define  $W'(v)$  as a  $d \times 1$  vector with

its  $s^{\text{th}}$  element given by  $w'(v_s) \prod_{t \neq s} w(v_t)$ . Then it is easy to check that

$$\int W'(v)(W'(v))^T dv = \nu_0^{d-1} \kappa_0 I_d, \quad (\text{B.9})$$

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  $J_3(x) = \frac{1}{nf(x)} \sum_j u_j W'_{h,jx}$ . Obviously,  $E(J_3(x)) = 0$ , and

$$\begin{aligned} \text{Var}(J_3(x)) &= \frac{1}{nf(x)^2} E[\sigma_j^2(W'_{h,jx})(W'_{h,jx})^T] \\ &= \frac{1}{nf(x)^2} \int f(x_j) \sigma^2(x_j) (W'_{h,jx})(W'_{h,jx})^T dx_j \\ &= \frac{1}{n(h_1 \dots h_d)^2 f(x)^2} \int f(x_j) \sigma^2(x_j) D_h^{-1} W'(\frac{x_j - x}{h}) (W'(\frac{x_j - x}{h}))^T D_h^{-1} dx_j \\ &= \frac{1}{nh_1 \dots h_d f(x)^2} \int (f\sigma^2)(x + hv) (D_h^{-1} W'(v)) (D_h^{-1} W'(v))^T dv \\ &= \frac{\nu_0^{d-1} \kappa_0 \sigma^2(x)}{nh_1 \dots h_d f(x)} D_h^{-2} + O_p((nh_1 \dots h_d)^{-1}), \end{aligned}$$

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

Under the condition that  $nh_1 \dots h_d \|h\|^2 \rightarrow \infty$ ,  $nh_1 \dots h_d \|h\|^4 \rightarrow 0$  as  $n \rightarrow \infty$ , and by Lyapunov's central limit theorem, we have that

$$(nh_1 \dots h_d)^{1/2} D_h(\tilde{\beta}(x) - \beta(x) - B_{2,h}(x)) \xrightarrow{d} N(0, c_{lc}(x) I_d),$$

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.  $\square$