NONPARAMETRIC ESTIMATION OF DERIVATIVE FUNCTIONS WITH DATA-DRIVEN OPTIMALLY SELECTED SMOOTHING PARAMETERS

A Dissertation

by

SHUANG YAO

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Qi Li
Co-Chair of Committee,	Ke-Li Xu
Committee Members,	Steve Puller
	Ximing Wu
Head of Department,	Timothy Gronberg

August 2014

Major Subject: Economics

ABSTRACT

Estimating gradients is of crucial importance across a broad range of applied economic domains. Here we consider data-driven bandwidth selection based on the gradient of an unknown regression function. This is a difficult problem empirically given that direct observation of the value of the gradient is typically not observed. The procedure developed here delivers bandwidths which behave asymptotically as though they were selected knowing the true gradient. This procedure is shown valid for semiparametric single index models. Simulated examples showcase the finite sample attraction of this new mechanism and confirm the theoretical predictions.

TABLE OF CONTENTS

LIST OF TABLES

1. INTRODUCTION AND LITERATURE REVIEW

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

The importance of appropriate smoothness selection for derivatives was illustrated by Wahba & Wang (1990) who showed in the smoothing spline setting that the ideal smoothing parameter depends on the derivative of the unknown function. A small strand of literature has developed focusing attention on smoothing parameter selection when interest hinges on the derivative. Within this literature there exist several different approaches for construction of the optimal bandwidth. To develop the intuition for existing approaches consider a univariate nonparametric regression model

$$y_j = g(x_j) + u_j \qquad\qquad j = 1, \ldots, n. \tag{1.1}$$

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

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

Beyond the factor method, Fan & Gijbels (1995) also proposed a two-step bandwidth selector which consists of constructing 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 found using the factor-method bandwidth. Once these measures are constructed, the final bandwidth, termed the refined bandwidth, is found by minimizing MISE. Fan, Gijbels, Hu & Huang (1996) showed that this bandwidth selection mechanism has desirable properties both theoretically as well as in simulated settings.

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

Each of the existing methods leaves something to be desired in a multivariate setting. The factor method requires bandwidth selection on the conditional mean followed by calculation of a scaling factor dependent upon the kernel function (in the univariate setting) which can be tedious. The calculation of noise-corrupted derivatives also requires computing the number of neighboring observations to construct the estimates prior to minimizing the criterion function. In high dimensional settings this may not be feasible. Lastly, plug-in approaches, while having desirable theoretical properties, require the calculation of numerous unknown quantities, neutering the ability of having a completely automatic procedure. All plug-in approaches require estimation of unknown functions and their derivatives prior to the formal selection of the bandwidth. Moreover, the plug-in formula for the optimal bandwidths can become quite complicated in high dimensional settings. The framework laid out here does not require adjustment, calculation of noise-corrupted derivatives or unknown quantities related to the underlying data generating process. The method also does not hinge on a pilot bandwidth nor a set of estimates being supplied to the criterion function, streamlining the process.

Our approach begins with the oracle LSCV setup for the gradient as in Müller et al. (1987), with a local-linear estimator. We then show that replacing the oracle gradient with a local-cubic estimator produces bandwidths which behave asymptotically as though the oracle was used. The intuition for this result is that the bias of the local-cubic estimator is of sufficiently smaller order relative to the local-linear estimator that the only aspect of the local-cubic estimator which appears in our asymptotic expansion of the LSCV criterion is the variance of the difference between these estimators (local-linear and local-cubic). In the limit, the variance of this difference behaves (up to a constant depending on the kernel) exactly as the case with the oracle gradient. Thus, bandwidths selected replacing the oracle gradient with the local-cubic estimator are asymptotically equivalent to those selected with the unknown oracle gradient.

The gradient based cross-validation (GBCV) approach studied here has several appealing features. First, the computational burden is dramatically decreased given that pilot bandwidths and first differences are not necessary to make the procedure operational. Further, the approach readily scales to the multivariate setting and is firmly entrenched within the data-driven bandwidth selection arena. Lastly, the method is intuitively appealing as it represents an easily explained procedure which mimics the traditional LSCV approach to bandwidth selection, albeit for gradients.

The remainder of the paper is as follows. Section 2 provides the formal details of our new cross-validation procedure and the asymptotic justification for our proposed method. Section 3 extends the GBCV approach to semiparametric single index models. Section 4 contains a set of simulations to show the performance of our bandwidth selection method for estimation of derivative functions compared with the oracle selection method. Concluding remarks appear in Section 5.

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

We consider the problem of using a data-driven method to select the smoothing parameters for estimation of the derivative of a function. Here we describe our gradient based cross-validation method first in the univariate setting and then for the general multivariate case.

2.1 The Univariate Case

To motivate the idea and keep the notational burden to a minimum, in this section we focus on the univariate nonparametric regression model in (B.8):

$$y_j = g(x_j) + u_j, \qquad\qquad j = 1, \ldots, n, \tag{2.1}$$

where the functional form of $g(\cdot)$ is not specified and the error term u_j satisfies $E(u_j|x_j) = 0$. Let $\beta(x) = dg(x)/dx$ denote the first order derivative function of $g(\cdot)$ with respect to x. Let $\hat{\beta}_{LL}(x)$ be the local-linear estimator of $\beta(x)$. Ideally, we would like to choose the smoothing parameter h to minimize the estimation mean squared error $E\{[\hat{\beta}_{LL}(x) - \beta(x)]^2\}$, or the sample analogue of it:

$$CV(h) \overset{def}{=} \frac{1}{n}\sum_{j=1}^{n}[\hat{\beta}_{LL}(x_j) - \beta(x_j)]^2 M(x_j), \tag{2.2}$$

where $M(\cdot)$ is a weight function with bounded support that trims out data near the boundary of the support of x.

Following the same arguments as in Racine & Li (2004) and Hall, Li & Racine

(2007), one can show that

$$CV(h) = \int E[\hat{\beta}_{LL}(x) - \beta(x)]^2 M(x)f(x)dx + (s.o.),$$

where $f(x)$ denotes the density function of x and $(s.o.)$ captures terms having probability orders smaller than the leading term $\int E\left[\hat{\beta}_{LL}(x) - \beta(x)\right]^2 M(x)f(x)dx$. Let $Bias^0\left(\hat{\beta}_{LL}(x)\right)$ and $Var^0\left(\hat{\beta}_{LL}(x)\right)$ denote the leading bias and leading variance terms of $\hat{\beta}_{LL}(x)$. Then the leading term of $CV(h)$ is given by

$$CV^0(h) \overset{def}{=} \int \left\{\left[Bias^0\left(\hat{\beta}_{LL}(x)\right)\right]^2 + Var^0\left(\hat{\beta}_{LL}(x)\right)\right\} M(x)f(x)dx. \tag{2.3}$$

Here, we explain the definition of leading bias and leading variance of $\hat{\beta}_{LL}(x)$. It can be shown that (e.g., Henderson et al (2012))

$$\hat{\beta}_{LL}(x) - \beta(x) = h^2 B(x) + \sqrt{\frac{V(x)}{nh^3}} Z_n + o_p(h^2 + (nh^3)^{-1/2}), \tag{2.4}$$

where $B(x) = \left(\frac{\mu_4 - \mu_2^2}{2\mu_2}\right)\frac{g''(x)f'(x)}{f(x)} + \frac{\mu_4 g'''(x)}{6\mu_2}$, $V(x) = \nu_2\sigma^2(x)/[\mu_2^2 f(x)]$, Z_n is a mean zero, unit variance random variable ($Z_n \xrightarrow{d} N(0,1)$ under some standard regularity conditions), $\mu_l = \int w(v)v^l dv$, $\nu_l = \int w(v)^2 v^l dv$, $m'(x)$, $m''(x)$ and $m'''(x)$ are the first, second and third derivative functions of m ($m = g$ or $m = f$). Note that $B(x)$ is nonrandom, therefore, we say that $h^2B(x) = Bias^0(\hat{\beta}_{LL}(x))$ is the leading bias of $\hat{\beta}_{LL}(x)$, and we say that $V(x)/(nh^3) = Var^0(\hat{\beta}_{LL}(x))$ is the leading variance of $\hat{\beta}_{LL}(x)$. Also, we say that $[\,Bias^0(\hat{\beta}_{LL})\,]^2 + Var(\hat{\beta}_{LL}) = h^4B^2(x) + V(x)/(nh^3) = MSE^0(\hat{\beta}_{LL}(x))$ is the leading MSE of $\hat{\beta}_{LL}(X)$. In the remaining part of the paper, the leading bias, variance and MSE of other local polynomial estimators are similarly defined.

The problem facing the econometrician is that one cannot compute $CV(h)$ defined by (2.2) because $\beta(x)$ is unknown. As an alternative, one can compute the leading bias and variance of $\hat{\beta}_{LL}(x)$, and choose the smoothing parameter h to minimize a weighted version of the integrated (leading) squared bias and variance of $\hat{\beta}_{LL}(x)$. This approach requires one to obtain initial estimates of $g(x)$ and $f(x)$ and their derivative functions up to the 3^{rd} order, which in turn requires one to use pilot smoothing parameters to estimate these unknown functions. This is called the 'plug-in' method (see Fan & Gijbels 1995). The 'plug-in' method of selecting the smoothing parameter is not completely automatic as it requires some initial choice of smoothing parameters. If the initial choices are far away from the optimal values, the 'plug-in' method may lead to poor selection of the smoothing parameters. Moreover, in the multivariate regression case or when there exists discrete covariates, this 'plug-in' method can be difficult to use as the 'plug-in' formulas are quite complex in these settings.

We propose a completely data-driven procedure to select h which is asymptotically equivalent to selecting an h that minimizes the infeasible objective function defined in (2.2). We construct our feasible objective function by replacing the unknown derivative function $\beta(x_j)$ by another consistent estimate of it, say $\hat{\beta}_{LP}(x_j)$, where subscript LP denotes an alternative local polynomial estimator. Hence, our objective function is based on

$$CV_{LP}(h) = \frac{1}{n}\sum_{j=1}^{n}\left[\hat{\beta}_{LL}(x_j) - \hat{\beta}_{LP}(x_j)\right]^2 M(x_j). \tag{2.5}$$

Our candidates for $\hat{\beta}_{LP}(\cdot)$ are from the set of local polynomial estimators, local-constant (LC), local-quadratic (LQ) and local-cubic (L-cubic) (or even higher order local polynomial estimators).

Following similar derivations as in Racine & Li (2004) and Henderson, Li &

Parmeter (2012), it can be shown that the leading term of $CV_{LP}(h)$ is given by

$$\begin{aligned} CV_{LP}^{0}(h) &= \int MSE^{0}\left[\hat{\beta}_{LL}(x) - \hat{\beta}_{LP}(x)\right] f(x)M(x)dx \\ &= \int \Big\{ \left[Bias^{0}(\hat{\beta}_{LL}(x) - \hat{\beta}_{LP}(x))\right]^{2} \\ &\quad + Var^{0}\Big(\hat{\beta}_{LL}(x) - \hat{\beta}_{LP}(x)\Big)\Big\} f(x)M(x)dx, \end{aligned}$$

where the superscript 0 denotes the leading term of $CV_{LP}(h)$. We want to select the local polynomial order of our estimation method such that (2.5) and (2.2) are asymptotically equivalent to each other. This may at first look like a formidable task, but in fact an easy solution to this problem exists, which we now detail.

We first define a d^{th} order local polynomial estimator of $\beta(x) = g'(x)$. We choose $b_0, b_1, \ldots, b_d$ to minimize the following objective function

$$\min_{b_0, b_1, \ldots, b_d} \sum_{j=1}^{n} \left(y_j - b_0 - b_1(x_j - x) - \cdots - b_d(x_j - x)^{d}\right)^{2} w\left(\frac{x_j - x}{h}\right), \qquad (2.6)$$

where $w(\cdot)$ is the kernel function and h is the smoothing parameter. The solution b_1 is the (d^{th}-order) local polynomial estimator of $\beta(x)$.

Using the notation $m'(x) = dm(x)/dx$, $m''(x) = d^2m(x)/dx^2$ and $m'''(x) = d^3m(x)/dx^3$ with $m(x) = g(x)$ or $m(x) = f(x)$, the d^{th} order local polynomial estimators' (with $0 \leq d \leq 3$) leading biases and variances are well established and given by[1]

$$\begin{aligned} & Bias^{0}\left(\hat{\beta}_{LC}(x)\right) \\ = \;& h^{2}\mu_{2}\frac{g''(x)f'(x)f(x) + 2g'(x)f''(x)f(x) + \frac{1}{2}g'''(x)f^{2}(x) - g'(x)[f'(x)]^{2}}{f^{2}(x)}, \end{aligned}$$

[1] For the local constant estimator, (2.6) does not give the derivative estimator directly. Rather, we have to take a derivative of $\hat{g}_{LC}(x)$ with respect to x to obtain a derivative estimator, i.e., $\hat{\beta}_{LC}(x) = \frac{d\hat{g}_{LC}(x)}{dx}$.

$$Var^0\left(\hat{\beta}_{LC}(x)\right) = \frac{1}{nh^3}\cdot\frac{\nu_0\sigma^2(x)}{f(x)},$$

$$Bias^0\left(\hat{\beta}_{LL}(x)\right) = h^2\left[\left(\frac{\mu_4-\mu_2^2}{2\mu_2}\right)\frac{g''(x)f'(x)}{f(x)} + \frac{\mu_4 g'''(x)}{6\mu_2}\right],$$

$$Var^0\left(\hat{\beta}_{LL}(x)\right) = \frac{1}{nh^3}\frac{\nu_2}{\mu_2^2}\frac{\sigma^2(x)}{f(x)},$$

$$Bias^0\left(\hat{\beta}_{LQ}(x)\right) = h^2\frac{\mu_4}{6\mu_2}g'''(x),$$

$$Var^0\left(\hat{\beta}_{LQ}(x)\right) = \frac{1}{nh^3}\cdot\frac{\nu_2\sigma^2(x)}{\mu_2^2 f(x)},$$

$$Bias^0\left(\hat{\beta}_{L\text{-}cubic}(x)\right) = O(h^4),$$

$$Var^0\left(\hat{\beta}_{L\text{-}cubic}(x)\right) = \frac{1}{nh^3}\frac{K_1}{K_2^2}\frac{\sigma^2(x)}{f(x)},$$

where $K_1 = (\mu_4\mu_6 - \mu_2^2\mu_6)^2\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)^2\nu_6 + 2(\mu_4\mu_6 - \mu_2^2\mu_6)(\mu_2^2\mu_4 - \mu_4^2)\nu_4$, $K_2 = \mu_2\mu_4\mu_6 - \mu_4^3 + \mu_2^2\mu_4^2 - \mu_2^3\mu_6$, $\mu_s = \int v^s w(v)dv$ and $\nu_s = \int v^s w^2(v)dv$, see Fan & Gijbels (1996, Theorem 3.1) and Henderson et al. (2012). Thus, (3.7) can be written as

$$\begin{aligned} CV^0(h) &= h^4\int\left[\left(\frac{\mu_4-\mu_2^2}{2\mu_2}\right)\frac{g''(x)f'(x)}{f(x)} + \frac{\mu_4 g'''(x)}{6\mu_2}\right]^2 f(x)M(x)dx \\ &\quad + \frac{1}{nh^3}\frac{\nu_2}{\mu_2^2}\int\sigma^2(x)M(x)dx \\ &= h^4\int\left[B_1(x)\right]^2 f(x)M(x)dx + \frac{1}{nh^3}V_1\int\sigma^2(x)M(x)dx, \end{aligned} \tag{2.7}$$

where $B_1(x) = \left(\frac{\mu_4-\mu_2^2}{2\mu_2}\right)\frac{g''(x)f'(x)}{f(x)} + \frac{\mu_4 g'''(x)}{6\mu_2}$ and $V_1 = \frac{\nu_2}{\mu_2^2}$.

We notice that variances of these local polynomial estimators are different from each other only by some multiplicative constants. In contrast, the biases are more complicated. They are distinct from each other by functions (including derivative functions) of x. This comparison motivates us to choose $LP = L\text{-}cubic$ given that $Bias^0\left(\hat{\beta}_{L\text{-}cubic}(x)\right) = O(h^4) = o(h^2)$ is negligible compared with the bias term of the local-linear estimator. Hence, the leading bias of $\hat{\beta}_{LL}(x) - \hat{\beta}_{LP}(x) \equiv \hat{\beta}_{LL}(x) - \hat{\beta}_{L\text{-}cubic}(x)$ is simply $Bias^0\left(\hat{\beta}_{LL}(x)\right)$. We still need to evaluate $Var[\hat{\beta}_{LL}(x) - \hat{\beta}_{L\text{-}cubic}(x)] = Var\left(\hat{\beta}_{LL}(x)\right) + Var\left(\hat{\beta}_{L\text{-}cubic}(x)\right) - 2Cov\left(\hat{\beta}_{LL}(x), \hat{\beta}_{L\text{-}cubic}(x)\right)$. Appendix A demonstrates that the leading term of the covariance between $\hat{\beta}_{L\text{-}cubic}(x)$ and $\hat{\beta}_{LL}(x)$ is given by

$$Cov^0\left(\hat{\beta}_{L\text{-}cubic}(x), \hat{\beta}_{LL}(x)\right) = \frac{1}{nh^3} \cdot \frac{(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4}{\mu_2 K_2} \cdot \frac{\sigma^2(x)}{f(x)} \tag{2.8}$$

Hence, the leading variance term of (3.9) with $LP = L\text{-}cubic$ is given by

$$\begin{aligned} & Var^0\left(\hat{\beta}_{LL}(x) - \hat{\beta}_{L\text{-}cubic}(x)\right) \\ = \; & Var^0\left(\hat{\beta}_{L\text{-}cubic}(x)\right) + Var^0\left(\hat{\beta}_{LL}(x)\right) - 2Cov^0\left(\hat{\beta}_{L\text{-}cubic}(x), \hat{\beta}_{LL}(x)\right) \\ = \; & \frac{1}{nh^3} \cdot \left[\frac{K_1}{K_2^2} + \frac{\nu_2}{\mu_2^2} - 2\frac{(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4}{\mu_2 K_2}\right] \cdot \frac{\sigma^2(x)}{f(x)}. \end{aligned}$$

Thus, by choosing $LP = L\text{-}cubic$, (3.9) can be written as

$$\begin{aligned} & CV^0_{L\text{-}cubic}(h) \\ = \; & h^4 \int \left[\left(\frac{\mu_4 - \mu_2^2}{2\mu_2}\right) \frac{g''(x) f'(x)}{f(x)} + \frac{\mu_4 g'''(x)}{6\mu_2}\right]^2 f(x) M(x) dx \\ & + \frac{1}{nh^3}\left[\frac{K_1}{K_2^2} + + \frac{\nu_2}{\mu_2^2} - 2\frac{(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4}{\mu_2 K_2}\right] \int \sigma^2(x) M(x) dx \\ = \; & h^4 \int [B_1(x)]^2 f(x) M(x) dx + \frac{1}{nh^3} V_{1,3} \int \sigma^2(x) M(x) dx. \end{aligned} \tag{2.9}$$

Let $h_{0,opt}$ and $h_{0,cubic}$ denote the values of h that minimizes (3.14) and (2.9), respectively, it is easy to see that

$$h_{0,opt} = \left[\frac{3V_1 \int \sigma^2(x) M(x) dx}{4 \int [B_1(x)]^2 f(x) M(x) dx} \right]^{1/7} n^{-1/7},$$

$$h_{0,cubic} = \left[\frac{3V_{1,3} \int \sigma^2(x) M(x) dx}{4 \int [B_1(x)]^2 f(x) M(x) dx} \right]^{1/7} n^{-1/7}.$$

Therefore, we have $h_{0,cubic} = (V_{1,3}/V_1)^{1/7} h_{0,opt}$. Letting $\tilde{h}_{cubic}$ denote the value of h that minimizes the feasible cross validation objective function (2.5) with $LP =$ *L-cubic*, we correct $\tilde{h}_{cubic}$ by multiplying it by a constant

$$\hat{h}_{cubic} = (V_1/V_{1,3})^{1/7} \ \tilde{h}_{cubic}.$$

It then follows that (under some regularity conditions similar to those given in Hall et al. (2007))

$$\hat{h}_{cubic}/h_{0,opt} \xrightarrow{p} 1. \tag{2.10}$$

Equation (2.10) follows from

$$\frac{\hat{h}_{cubic}}{h_{0,opt}} = \frac{(V_1/V_{1,3})^{1/7} \ \tilde{h}_{cubic}}{h_{0,opt}} = \frac{(V_1/V_{1,3})^{1/7} \ [h_{0,cubic} + o_p(h_{0,cubic})]}{h_{0,opt}} = 1 + o_p(1) \tag{2.11}$$

because $(V_1/V_{1,3})^{1/7} h_{0,cubic}/h_{0,opt} = 1$.

A rigorous proof of (2.11) follows similar proof arguments as Hall et al. (2007). We omit the detailed steps to save space.

It is straightforward to show that $(V_1/V_{1,3})^{1/7} = (16/15)^{1/7} \approx 1.009$ if we use the Gaussian kernel function and $(V_1/V_{1,3})^{1/7} = (308/945)^{1/7} \approx 0.852$ if we use the Epanechnikov kernel function. If a standard normal kernel is used in the local-linear and cubic estimations, there is hardly a need for adjustment of the optimally selected bandwidth.

2.2 The Multivariate Case

In the multivariate setting, we have $x = (x_1, \cdots, x_q)$ where $q > 1$. We want to choose $\mathbf{h} = (h_1, \cdots, h_q)$ optimally in the sense that they minimize the estimation mean squared error for the first order derivative functions of $g(x)$. Instead of considering the whole $q \times 1$ vector of the derivative function, we consider each partial derivative separately. We use the notation $\beta_s(x) = \partial g(x_1, \cdots, x_q)/\partial x_s$ for $s = 1, \ldots, q$ to denote the first order partial derivative functions. Without loss of generality we will focus on the case of $s = 1$. Similar to the univariate x case, ideally, we would like to choose $\mathbf{h}$ to minimize the following sample analog of the estimation mean squared error:

$$CV_1(\mathbf{h}) \overset{def}{=} \frac{1}{n} \sum_{j=1}^{n} \left[\hat{\beta}_{1,LL}(x_j) - \beta_1(x_j)\right]^2 M(x_j), \tag{2.12}$$

where $\hat{\beta}_{1,LL}(x)$ is the local-linear estimator of $\beta_1(x) = \partial g(x)/\partial x_1$ obtained from

$$\min_{a,b} \sum_{j=1}^{n} \left[y_j - a - b^T(x_j - x)\right]^2 W_{h,jx}. \tag{2.13}$$

where b estimates $(\partial g(x)/\partial x_1, \cdots, \partial g(x)/\partial x_q)^T$, the $q \times 1$ vector of first derivative functions. The first component of b in (2.13) is $\hat{\beta}_{1,LL}(x)$, the local linear estimator of $\beta_1(x)$. $W_{h,jx} = \prod_{s=1}^{q} h_s^{-1} w((x_{js} - x_s)/h_s)$ is the product kernel function.[2]

In practice $\beta_1(x_j)$ is unknown. We suggest replacing $\beta_1(x_j)$ in (2.12) by the local-cubic estimator $\hat{\beta}_{1,L\text{-}cubic}(x_i)$. In order to demonstrate how we use the local-cubic estimator of $\beta_1(x)$ with multivariate x, we need to introduce some additional

[2] A referee suggested that one may use a non-diagonal bandwidth matrix instead of the product kernel function. We conjecture that the main result of this paper remains valid when one uses a non-diagonal bandwidth matrix. However, this is beyond the scope of the current paper.

notation. Following Masry (1996), we define the following

i) For $l \in \{0,1,2,3\}$, let $N_l = \begin{pmatrix} l+q-1 \\ q-1 \end{pmatrix}$ and $\mathcal{N}_3 = \sum_{l=0}^{3} N_l$, where $\begin{pmatrix} a \\ b \end{pmatrix} = \frac{a!}{b!(a-b)!}$, a and b are positive integers ($a \geq b$).

ii) Let $\mathbf{k} = \{k_1, \ldots, k_q\}$ and $|\mathbf{k}| = \sum_{l=1}^{q} k_l$,

$$\begin{aligned} \mathbf{k}! &= k_1! \times \cdots \times k_q!, \quad x^{\mathbf{k}} = x_1^{k_1} \times \cdots \times x_q^{k_q}, \\ \sum_{0 \leq |\mathbf{k}| \leq 3} &= \sum_{l=0}^{3} \sum_{k_1=0}^{l} \cdots \sum_{\substack{k_q=0 \\ |\mathbf{k}|=k_1+\cdots+k_q=l}}^{l}, \end{aligned}$$

iii) $\left[b_0, b_{\mathbf{1}}^T, b_{\mathbf{2}}^T, b_{\mathbf{3}}^T\right]^T$ is an $\mathcal{N}_3 \times 1$ column vector where $b_{\mathbf{k}}$ is $N_{|\mathbf{k}|} \times 1$ column vector composed of $b_{\mathbf{k}}$ in lexicographical order.

$\hat{\beta}_{1,L\text{-}cubic}(x)$ is the first component of the solution $b_{\mathbf{1}}$ in the following minimization problem:

$$\min_{\{b_{\mathbf{0}},\ldots,b_{\mathbf{3}}\}} \sum_{j=1}^{n} \left(y_j - \sum_{0 \leq |\mathbf{k}| \leq 3} b_{\mathbf{k}} \left(x_j - x\right)^{\mathbf{k}} \right)^2 W_{h,jx}, \tag{2.14}$$

We suggest replacing $\beta_1(x_j)$ in (2.12) by $\hat{\beta}_{1,L\text{-}cubic}(x_j)$ and choose $\mathbf{h}$ to minimize the following feasible cross-validation function:

$$CV_{1,f}(\mathbf{h}) \overset{def}{=} \frac{1}{n} \sum_{j=1}^{n} \left[\hat{\beta}_{1,LL}(x_j) - \hat{\beta}_{1,L\text{-}cubic}(x_j)\right]^2 M(x_j). \tag{2.15}$$

In the multivariate case, the leading biases and variances of $\hat{\beta}_{1,LL}(x)$ and $\hat{\beta}_{1,L\text{-}cubic}(x)$ are given by

$$\begin{aligned} Bias^0(\hat{\beta}_{1,LL}(x)) = \; & \left[\left(\tfrac{\mu_4-\mu_2^2}{2\mu_2}\right)\tfrac{f_1(x)g_{11}(x)}{f(x)} + \tfrac{\mu_4}{6\mu_2} g_{111}(x)\right] h_1^2 + \tfrac{\mu_2}{2} \textstyle\sum_{s\neq 1}^{q} g_{1ss}(x) h_s^2 \\ & + \tfrac{\mu_2}{f(x)} \textstyle\sum_{s\neq 1}^{q} f_s(x) g_{1s}(x) h_s^2, \end{aligned}$$

$$Var^0(\hat{\beta}_{1,LL}(x)) = \frac{1}{nh_1^3 h_2 \cdots h_q} \frac{\nu_0^{q-1}\nu_2}{\mu_2^2} \frac{\sigma^2(x)}{f(x)},$$

$$Bias^0(\hat{\beta}_{1,L\text{-}cubic}(x)) = O(||h||^4), \quad \text{where } ||h||^2 = \sum_{s=1}^q h_s^2.$$

$$Var^0(\hat{\beta}_{1,L\text{-}cubic}(x)) = \frac{1}{nh_1^3 h_2 \cdots h_q} \frac{\nu_0^{q-1}K_1}{K_2^2} \frac{\sigma^2(x)}{f(x)},$$

where K_1, K_2, μ_s and ν_s are as defined in the univariate case, see Masry(1996). For the general q-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 $m_{stl}(x) = \frac{\partial^3 m(x)}{\partial x_s \partial x_t \partial x_l}$, where $m(x) = g(x)$ or $m(x) = f(x)$. Then, we have the leading term of $CV_1(h_1)$ given by

$$\begin{aligned}
CV_1^0(\mathbf{h}) &= \int \Big\{ \Big[Bias^0(\hat{\beta}_{1,LL}(x)) \Big]^2 + Var^0(\hat{\beta}_{1,LL}(x)) \Big\} M(x)f(x)dx \\
&= \int \Bigg\{ \Big[\Big(\frac{\mu_4 - \mu_2^2}{2\mu_2} \Big) \frac{f_1(x)g_{11}(x)}{f(x)} + \frac{\mu_4}{6\mu_2} g_{111}(x) \Big] h_1^2 \\
&\qquad + \frac{\mu_2}{2} \sum_{s\neq 1}^q g_{1ss}(x) h_s^2 + \frac{\mu_2}{f(x)} \sum_{s\neq 1}^q f_s(x) g_{1s}(x) h_s^2 \Bigg\}^2 f(x)M(x)dx \\
&\quad + \frac{\nu_0^{q-1}V_1}{nh_1^3 h_2 \cdots h_q} \int \sigma^2(x)M(x)dx, \qquad (2.16)
\end{aligned}$$

and as in the univariate setting $V_1 = \frac{\nu_2}{\mu_2^2}$.

To obtain the leading term of $CV_{1,f}(\mathbf{h})$, we need to calculate

$$Bias^0\left(\hat{\beta}_{1,LL}(x) - \hat{\beta}_{1,L\text{-}cubic}(x)\right) \quad \text{and} \quad Var^0\left(\hat{\beta}_{1,LL}(x) - \hat{\beta}_{1,L\text{-}cubic}(x)\right).$$

Since $Bias^0\left(\hat{\beta}_{1,cubic}(x)\right) = O(||h||^4) = o(||h||^2)$, we have $Bias^0\left(\hat{\beta}_{1,LL}(x) - \hat{\beta}_{1,L\text{-}cubic}(x)\right)$ $= Bias^0\left(\hat{\beta}_{1,LL}(x)\right) - Bias^0\left(\hat{\beta}_{1,L\text{-}cubic}(x)\right) = Bias^0\left(\hat{\beta}_{1,LL}(x)\right) + (s.o.)$. It can be shown that

$$\begin{aligned} & Cov^0\left(\hat{\beta}_{1,LL}(x), \hat{\beta}_{1,L\text{-}cubic}(x)\right) \\ = \ & \frac{1}{nh_1^3 h_2 \cdots h_q} \cdot \frac{\nu_0^{q-1}[(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4]}{\mu_2 K_2} \cdot \frac{\sigma^2(x)}{f(x)}. \end{aligned}$$

Thus,

$$\begin{aligned} Var^0\left(\hat{\beta}_{1,LL}(x) - \hat{\beta}_{1,L\text{-}cubic}(x)]\right) &= Var^0\left(\hat{\beta}_{1,LL}(x)\right) + Var^0\left(\hat{\beta}_{1,L\text{-}cubic}(x)\right) \\ &\quad -2Cov^0\left(\hat{\beta}_{1,LL}(x), \hat{\beta}_{1,L\text{-}cubic}(x)\right) \\ &= \frac{\nu_0^{q-1} V_{1,3}}{nh_1^3 h_2 \cdots h_q} \cdot \frac{\sigma^2(x)}{f(x)} \end{aligned}$$

where $V_{1,3} = \frac{K_1}{K_2^2} + \frac{\nu_2}{\mu_2^2} - 2\frac{(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4}{\mu_2 K_2}$. Then, we have that the leading term of $CV_{1,f}(h_1)$ (defined in (2.15)) is given by

$$\begin{aligned} & CV_{1,f}^0(\mathbf{h}) \\ = \ & \int \left\{ \left[Bias^0\left(\hat{\beta}_{1,LL}(x) - \hat{\beta}_{1,L\text{-}cubic}(x)\right)\right]^2 \right. \\ & \qquad \left. + Var^0\left(\hat{\beta}_{1,LL}(x) - \hat{\beta}_{1,L\text{-}cubic}(x)\right) \right\} M(x) f(x) dx \\ = \ & \int \left\{ \left[\left(\frac{\mu_4 - \mu_2^2}{2\mu_2}\right) \frac{f_1(x) g_{11}(x)}{f(x)} + \frac{\mu_4}{6\mu_2} g_{111}(x) \right] h_1^2 \right. \\ & \qquad \left. + \frac{\mu_2}{2} \sum_{s \neq 1}^{q} g_{1ss}(x) h_s^2 + \frac{\mu_2}{f(x)} \sum_{s \neq 1}^{q} f_s(x) g_{1s}(x) h_s^2 \right\}^2 f(x) M(x) dx \\ & + \frac{\nu_0^{q-1} V_{1,3}}{nh_1^3 h_2 \cdots h_q} \int \sigma^2(x) M(x) dx. \qquad (2.17) \end{aligned}$$

For expositional simplicity, we assume that $h_1 = h_2 = \cdots = h_q = h$. Let h_{opt} and

h_{cubic} denote the values of h that minimizes (2.16) and (2.17) respectively, then we have

$$h_{opt} = \left[\frac{3\nu_0^{q-1}V_1 \int \sigma^2(x)M(x)dx}{4\int B_2(x)B_3(x)f(x)M(x)dx}\right]^{1/(q+6)} n^{-1/(q+6)}$$

$$h_{cubic} = \left[\frac{3\nu_0^{q-1}V_{1,3} \int \sigma^2(x)M(x)dx}{4\int B_2(x)B_3(x)f(x)M(x)dx}\right]^{1/(q+6)} n^{-1/(q+6)}$$

where $B_2(x) = \left(\frac{\mu_4-\mu_2^2}{2\mu_2}\right)\frac{f_1(x)g_{11}(x)}{f(x)} + \frac{\mu_4 g_{111}(x)}{6\mu_2}$ and $B_3(x) = B_2(x) + \frac{\mu_2}{2}\sum_{s\neq 1}^q g_{1ss}(x) + \frac{\mu_2}{f(x)}\sum_{s\neq 1}^q f_s(x)g_{1s}(x)$. Thus, $h_{cubic} = (V_{1,3}/V_1)^{1/(q+6)}h_{opt}$. Let $\tilde{h}_{cubic}$ denote the value of h that minimizes (2.15), we correct it by

$$\hat{h}_{cubic} = (V_1/V_{1,3})^{1/(q+6)}\ \tilde{h}_{cubic}.$$

Then we have

$$\hat{h}_{cubic}/h_{opt} \xrightarrow{p} 1.$$

Note that if the gaussian kernel is used, the larger the q, the closer the factor $(V_1/V_{1,3})^{1/(q+6)}$ is to 1. For example, $(V_1/V_{1,3})^{1/(q+6)} = 1.009$ for $q = 1$, $(V_1/V_{1,3})^{1/(q+6)} = 1.008$ if $q = 2$, and $(V_1/V_{1,3})^{1/(q+6)} = 1.007$ if $q = 3$. Therefore, if a normal kernel is used, there is hardly a need for multiplying by the adjustment constant.

3. GRADIENT BASED CROSS-VALIDATION METHOD FOR SINGLE INDEX MODEL DERIVATIVE ESTIMATIONS

From section 2 we can see when extended to multivariate case, gradient based cross-validation method may have computational complication. Just take a look at of the case q = 2 and case q = 3, we need to estimate 10 parameters from (2.14) for q = 2 and 20 parameters from (2.14) for q = 3. However, we do not have such problem when the model is a single index model and this model is widely used by applied econometricians. Semiparametric single index models arise naturally in binary choice settings. Let y_i denote a binary dependent variable whose value is determined by a single index and an error term as follows:

$$y_i = \begin{cases} 1 \quad \text{if} \quad y_i^* = X_i^\top \gamma - \epsilon_i \geq 0; \\ 0 \quad \text{if} \quad y_i^* = X_i^\top \gamma - \epsilon_i < 0. \end{cases} \tag{3.1}$$

where y^* is a latent variable and ϵ is a continuously distributed random variable independent of X and whose distribution in our case is unknown. Thus the conditional mean function

$$\begin{aligned} E(y_i|X_i) &= 1 * Prob(y_i = 1|X_i) + 0 * Prob(y_i = 0|X_i) \\ &= Prob(y_i = 1|X_i) \\ &= Prob(\epsilon_i \leq X_i^\top \gamma | X_i) \\ &\overset{def}{=} g(X_i^\top \gamma), \end{aligned} \tag{3.2}$$

which gives us the single index model

$$\begin{aligned} y_i &= E(y_i|X_i) + u_i \\ &= g(X_i^\top \gamma) + u_i. \end{aligned} \tag{3.3}$$

In binary choice setting, $g(\cdot)$ is the cumulative density function (cdf) of ϵ and $\beta(\cdot) = \partial g(\cdot)/\partial z$ is the probability density function (pdf) of ϵ. As the property of cdf and pdf functions, $g(\cdot)$ is nondecreasing and $\beta(\cdot)$ is non-negative. However, in practice these two properties are not guaranteed in estimation if we do not select the smoothing parameters appropriately, making the estimated conditional mean function $\hat{g}(\cdot)$ and derivative function $\hat{\beta}(\cdot)$ hard to be interpreted as a cdf and pdf function.

In practice, we often arrive at an undersmoothed function curve by using the traditional smoothing parameter selection method. Chen, Gao and Li (2013) gives us an empirical example that fits into this scenario. It chooses the optimal bandwidth as the one that minimizes the mean squared estimation error which is defined by

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\left[Y_{it} - \hat{g}_i^{(-t)}(X_{it}^\top \hat{\theta}^{(-t)})\right]^2$$

in a panel data setting, where $\hat{g}_i^{(-t)}(X_{it}^\top \hat{\theta}^{(-t)})$ is the leave-one-out estimator of $g(X_{it}^\top \hat{\theta})$. This bandwidth selection method often leads to wiggly fitted curves. That is, it often leads to undersmoothing.

Targeted at ameliorating such problems, We discuss in this section how we can apply gradient based cross-validation method to single index model, which is optimal for derivative estimation and expected to mitigate the “undersmoothing” problems. We will elaborate it in the following subsections.

3.1 The Case That γ_0 Is Known

If γ_0 is known, actually we can apply the smoothing parameter selection method discussed in section 2 directly. We explain it briefly in this subsection.

For the pseudo case in which γ_0 is known, we can generate the regressor $z_i = X_i^\top \gamma_0$ and consider the following nonparametric model:

$$y_i = g(z_i) + u_i, \tag{3.4}$$

where the functional form of $g(\cdot)$ is not specified, the error term u_i satisfies $E(u_i|X_i) = 0$. Let $\beta(z) = \partial g(z)/\partial z$ denote the first order derivative function of $g(\cdot)$ with respect to z. Let $\hat{\beta}_{LL}(z)$ be the local linear estimator of $\beta(z)$, ideally, one would like to choose smoothing parameter h to minimize the estimation mean squared error $E\{[\hat{\beta}_{LL}(z) - \beta(z)]^2\}$, or minimize a sample analogue of it:

$$CV(h) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}(z_i) - \beta(z_i)]^2 M(z_i), \tag{3.5}$$

where $M(\cdot)$ is a weight function that trims out data near the boundary of the support of z.

Following the same arguments as in Racine and Li (2004), and Hall, Li and Racine (2007), one can show that

$$CV(h) = \int E[\hat{\beta}_{LL}(z) - \beta(z)]^2 M(z) f(z) dz + (s.o.), \tag{3.6}$$

where $(s.o.)$ denote terms having smaller orders than $\int E[\hat{\beta}_{LL}(z) - \beta(z)]^2 M(z) f(z) dz$.

Let $Bias^0(\hat{\beta}(z))$ and $Var^0(\hat{\beta}(z))$ denote the leading bias and leading variance terms of $\hat{\beta}_{LL}(z)$. Then the leading term of $CV(h)$ is given by

$$CV^0(h) \overset{def}{=} \int \left\{ \left[Bias^0(\hat{\beta}_{LL}(z)) \right]^2 + Var^0(\hat{\beta}_{LL}(z)) \right\} M(z) f(z) dz. \tag{3.7}$$

The difficulty is that $\beta(z_i)$ is unknown so that one cannot compute $CV(h)$ defined in (3.5) in practice. Gradient based cross-validation method proposes to use the local cubic estimate to replace the unknown $\beta(z_i)$ in the criterion function (3.5) and choose h to minimize the following feasible objective function:

$$CV_f(h) \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i) \right]^2 M(z_i). \tag{3.8}$$

where $\hat{\beta}_{Cubic}(\cdot)$ denotes the local cubic estimator of $\beta(\cdot)$.

Following same derivations as shown in Racine and Li (2004) and Henderson et al. (2012), it can be shown that the leading term of $CV_f(h)$ is given by

$$\begin{aligned} CV_f^0(h) &= \int MSE \left[\hat{\beta}_{LL}(z) - \hat{\beta}_{Cubic}(z) \right] f(z) M(z) dz \\ &= \int \left\{ \left[Bias^0(\hat{\beta}_{LL}(z) - \hat{\beta}_{Cubic}(z)) \right]^2 \right. \\ &\quad \left. + Var^0\Big(\hat{\beta}_{LL}(z) - \hat{\beta}_{Cubic}(z) \Big) \right\} f(z) M(z) dz + (s.o.), \end{aligned} \tag{3.9}$$

where the superscript 0 denotes the leading term of $CV_f(h)$.

For the local linear and local cubic estimators, their leading biases and leading variances are well established:

$$Bias^0(\hat{\beta}_{LL}(z)) = h^2 \left[(\frac{\mu_4 - \mu_2^2}{2\mu_2}) \frac{g''(z) f'(z)}{f(z)} + \frac{\mu_4 g'''(z)}{6\mu_2} \right] \tag{3.10}$$

$$Var^0(\hat{\beta}_{LL}(z)) = \frac{1}{nh^3}\frac{\nu_2}{\mu_2^2}\frac{\sigma^2(z)}{f(z)} \tag{3.11}$$

$$Bias^0(\hat{\beta}_{Cubic}(z)) = O(h^4) \tag{3.12}$$

$$Var^0(\hat{\beta}_{Cubic}(z)) = \frac{1}{nh^3}\frac{D}{|A|^2}\frac{\sigma^2(z)}{f(z)} \tag{3.13}$$

where $D = (\mu_4\mu_6 - \mu_2^2\mu_6)^2\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)^2\nu_6 + 2(\mu_4\mu_6 - \mu_2^2\mu_6)(\mu_2^2\mu_4 - \mu_4^2)\nu_4$ and $|A| = \mu_2\mu_4\mu_6 - \mu_4^3 + \mu_2^2\mu_4^2 - \mu_2^3\mu_6$, see Fan & Gijbels (1996) Theorem 3.1 and Henderson et al. (2012). Thus by using (3.10) and (3.11), (3.7) could be written as

$$\begin{aligned} CV^0(h) &= h^4 \int \left[\left(\frac{\mu_4 - \mu_2^2}{2\mu_2} \right) \frac{g''(z)f'(z)}{f(z)} + \frac{\mu_4 g'''(z)}{6\mu_2} \right]^2 f(z)M(z)dz \\ &\quad + \frac{1}{nh^3}\frac{\nu_2}{\mu_2^2} \int \sigma^2(z)M(z)dz \\ &= h^4 \int [B_1(z)]^2 f(z)M(z)dz + \frac{1}{nh^3} V_1 \int \sigma^2(z)M(z)dz, \end{aligned} \tag{3.14}$$

where $B_1(z) = \left(\frac{\mu_4 - \mu_2^2}{2\mu_2}\right)\frac{g''(z)f'(z)}{f(z)} + \frac{\mu_4 g'''(z)}{6\mu_2}$ and $V_1 = \nu_2/\mu_2^2$.

Note that since $Bias^0(\hat{\beta}_{Cubic}(z)) = O(h^4) = o(h^2)$ is negligible, the leading bias of $\hat{\beta}_{LL}(z) - \hat{\beta}_{Cubic}(z)$ is simply $Bias^0(\hat{\beta}_{LL}(z))$. It can be shown that the leading covariance between $\hat{\beta}_{LL}(z)$ and $\hat{\beta}_{Cubic}(z)$ is

$$Cov^0(\hat{\beta}_{LL}(z), \hat{\beta}_{Cubic}(z)) = \frac{1}{nh^3} \cdot \frac{(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4}{\mu_2|A|} \cdot \frac{\sigma^2(z)}{f(z)} \tag{3.15}$$

Hence, the leading variance term is given by

$$Var^0(\hat{\beta}_{LL}(z) - \hat{\beta}_{Cubic}(z))$$

$$
\begin{aligned}
&= Var^0(\hat{\beta}_{LL}(z)) + Var^0(\hat{\beta}_{Cubic}(z)) - 2Cov^0(\hat{\beta}_{LL}(z), \hat{\beta}_{Cubic}(z)) \\
&= \frac{1}{nh^3} \cdot \left[\frac{D}{|A|^2} + \frac{\nu_2}{\mu_2^2} - 2\frac{(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4}{\mu_2|A|} \right] \cdot \frac{\sigma^2(z)}{f(z)}. \qquad (3.16)
\end{aligned}
$$

Using (3.10) and (3.16) , (3.9) could be written as

$$
\begin{aligned}
CV_f^0(h) &= h^4 \int \left[\left(\frac{\mu_4 - \mu_2^2}{2\mu_2} \right) \frac{g''(z)f'(z)}{f(z)} + \frac{\mu_4 g'''(z)}{6\mu_2} \right]^2 f(z)M(z)dz \\
&\quad + \frac{1}{nh^3} \left[\frac{D}{|A|^2} + + \frac{\nu_2}{\mu_2^2} - 2\frac{(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4}{\mu_2|A|} \right] \int \sigma^2(z)M(z)dz \\
&= h^4 \int [B_1(z)]^2 f(z)M(z)dz + \frac{1}{nh^3} V_{1,3} \int \sigma^2(z)M(z)dz \qquad (3.17)
\end{aligned}
$$

where

$$
V_{1,3} = \frac{D}{|A|^2} + + \frac{\nu_2}{\mu_2^2} - 2\frac{(\mu_4\mu_6 - \mu_2^2\mu_6)\nu_2 + (\mu_2^2\mu_4 - \mu_4^2)\nu_4}{\mu_2|A|}.
$$

Let $h_{0,Cubic}$ and $h_{0,opt}$ denote the values of h that minimizes (3.17) and (3.14), respectively, it is easy to see that

$$
h_{0,Cubic} = \left[\frac{3V_{1,3} \int \sigma^2(z)M(z)dz}{4 \int [B_1(z)]^2 f(z)M(z)dz} \right]^{1/7} n^{-1/7} \qquad (3.18)
$$

$$
h_{0,opt} = \left[\frac{3V_1 \int \sigma^2(z)M(z)dz}{4 \int [B_1(z)]^2 f(z)M(z)dz} \right]^{1/7} n^{-1/7}. \qquad (3.19)
$$

Thus, $h_{0,Cubic} = (V_{1,3}/V_1)^{1/7} h_{0,opt}$. Let h_{opt} and $\tilde{h}_{Cubic}$ denote the value of h that minimize (3.5) and (3.8) respectively, define $\hat{h}_{Cubic} \overset{def}{=} (V_1/V_{1,3})^{1/7}\, \tilde{h}_{Cubic}$, then we have $\hat{h}_{Cubic}/h_{opt} \xrightarrow{p} 1$. We summarize the above results in the following lemma:

Lemma 3.1.1 *(a) Let* $\hat{\beta}_{LL}(\cdot)$ *and* $\hat{\beta}_{Cubic}(\cdot)$ *denote the local linear and local cubic estimator, we have* $\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i) = O_p\left(h^2 + \frac{1}{\sqrt{nh^3}}\right)$*; (b) Let* h_{opt} *and* $\tilde{h}_{Cubic}$ *denote the value of* h *that minimize (3.5) and (3.8) respectively, define* $\hat{h}_{Cubic} \overset{def}{=}$

$(V_1/V_{1,3})^{1/7}\ \tilde{h}_{Cubic}$*, then we have* $\hat{h}_{Cubic}/h_{opt} \xrightarrow{p} 1$.

3.2 The Case That γ_0 Is Unknown

In last subsection, We show that, if γ_0 is known, we can actually select smoothing parameter h that asymptotically approaches the optimal smoothing parameter value that minimizes the estimation mean square error of the derivative estimation. However, in single index model, coefficient γ is unknown to us. Thus the smoothing parameter selection method shown in subsection 3.1 is not ready to use. In this subsection, We will show that by combining with the minimum average variance (MAV) estimation method (Xia & Härdle, 2006), we can achieve similar results as shown in lemma 3.1.1. The proof is quite tedious and We put them in appendix.

Since γ_0 is unknown, $z_i = X_i^\top \gamma_0$ can not be directly used for the smoothing parameter selection. For this reason, we need to estimate the coefficient γ first. We choose the minimum average variance (MAV) estimate $\hat{\gamma}$ to replace γ_0 and generate $\hat{z}_i = X_i^\top \hat{\gamma}$. Xia & Härdle (2006) shows that

$$\hat{\gamma} - \gamma_0 = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{3.20}$$

This property of $\hat{\gamma}$ is key to our success of smoothing parameter selection.

Like in subsection 3.1, ideally one would like to choose smoothing parameter h to minimize

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_{LL}(\hat{z}_i) - \beta(z_i)\right]^2 M(X_i) \tag{3.21}$$

where $\hat{\beta}_{LL}(\cdot)$ is the local linear estimator of $\beta(\cdot) = \partial g(\cdot)/\partial z$, $\hat{z}_i = X_i^\top \hat{\gamma}$, $z_i = X_i^\top \gamma_0$ and $\hat{\gamma}$ is the MAV estimate of coefficient γ.

The difficulty is that we do not observe $\beta(z_i)$ and do not have any natural observable approximation for it. To make the criterion function feasible, We use the local cubic estimator of $\beta(\cdot)$ evaluated at $\hat{z}_i = X_i^\top \hat{\gamma}$ to replace the oracle in (3.21) and arrive at the following objective function:

$$CV_f(h) = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{Cubic}(\hat{z}_i) \right]^2 M(X_i) \tag{3.22}$$

where $\hat{\beta}_{Cubic}(\cdot)$ denotes the local cubic estimator.

One can see that if we replace the generated regressor $\hat{z}_i = X_i^\top \hat{\gamma}$ by $z_i = X_i^\top \gamma_0$ in (3.21) and (3.22), they are the same as (3.5) and (3.8) in subsection 3.1. If (3.21) and (3.22) is asymptotically equivalent to (3.5) and (3.8) respectively, then we can follow the proof line as in subsection 3.1. Actually this is true, because the MAV estimator $\hat{\gamma}$ has a faster convergent rate than the local polynomial estimator $\hat{\beta}_{LL}(\cdot)$ and $\hat{\beta}_{Cubic}(\cdot)$. These facts are summarized in the following two lemmas.

Lemma 3.2.1 *$CV_f(h) = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{Cubic}(\hat{z}_i) \right]^2 M(X_i)$ is asymptotically equivalent to $\frac{1}{n} \sum_{i=1}^{n} \left[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i) \right]^2 M(X_i)$, where $\hat{\beta}_{LL}(\cdot)$ is the local linear estimator, $\hat{z}_i = X_i^\top \hat{\gamma}$, $z_i = X_i^\top \gamma_0$, $\hat{\gamma}$ is the minimum average variance (MAV) estimator of γ (see Xia & Härdle 2006) and γ_0 is the true value of γ.*

Proof: See the appendix.

Lemma 3.2.2 *$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\beta}_{LL}(\hat{z}_i) - \beta(z_i) \right]^2 M(X_i)$ is asymptotically equivalent to*

$\frac{1}{n} \sum_{i=1}^{n} \left[\hat{\beta}_{LL}(z_i) - \beta(z_i) \right]^2 M(X_i)$, where $\hat{\beta}_{LL}(\cdot)$ is the local linear estimator, $\hat{z}_i = X_i^\top \hat{\gamma}$, $z_i = X_i^\top \gamma_0$, $\hat{\gamma}$ is the minimum average variance (MAV) estimator of γ (see Xia & Härdle 2006) and γ_0 is the true value of γ..

Proof: See the appendix.

By using lemma 3.2.1 and lemma 3.2.2, we can follow the same proof line as in subsection 3.1 and obtain similar results as in lemma 3.1.1. Below We first make some regularity assumptions.

Assumption 3.2.3 *(i) The data* $\{X_i, y_i\}_{i=1}^n$ *are independent and identically distributed (i.i.d.),* $z_i = X_i^\top \gamma_0$ *admits a density function* $f(z)$*. (ii) Let* $g(z_i) = E(y_i|z_i = X_i^\top \gamma_0)$*.* $g(z)$ *has continuous partial derivative functions up to fourth-order on* $X \in \mathcal{M}$*, where* $\mathcal{M}$ *is the support of the trimming function (*$\mathcal{M}$ *is a compact subset of* $\mathcal{R}^q$*). (iii)* $f(z)$ *has continuous partial derivatives up to second-order on* $X \in \mathcal{M}$*.*

Assumption 3.2.4 *(i) Let* $u_i = y_i - g(z_i)$*. Then* $\sigma^2(z) = E(u_i^2|z_i = z)$ *is a continuous function on* $X \in \mathcal{M}$*. (ii) Define* $\mu_m(z_i) = E(u_i^m|z_i)$*,* $\mu_m(z)$ *is bounded on* $X \in \mathcal{M}$ *for all finite positive* m*.*

Assumption 3.2.5 *(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 3.2.6 $h \in H_n$*, where* $H_n = \{h : \; c_1 n^{-1/(1+\delta_1)} \leq h \leq c_2 n^{-1/(1+6+\delta_2)}\}$*, for some small positive constant* $\delta_1 > 0$*, and large positive constant* $\delta_2 > 0$*, where* c_1 *and* c_2 *are positive constants.*

Under assumption 3.2.3 to 3.2.6, we have the follow similar result as in lemma 3.1.1:

THEOREM 3.2.1 *Let* $h_{si,opt}$ *and* $\tilde{h}_{si}$ *denote the value of* h *that minimize (3.21) and (3.22) respectively. Define* $\hat{h}_{si} \overset{def}{=} (V_1/V_{1,3})^{1/7}\tilde{h}_{si}$*. Under assumption 3.2.3 to 3.2.6, we have*

$$\hat{h}_{si}/h_{si,opt} \xrightarrow{p} 1. \tag{3.23}$$

Proof: By lemma 3.1.1, lemma 3.2.1 and lemma 3.2.2, theorem 3.2.1 is proved.

This result shows that by using the generated regressor we can apply smoothing parameter selection method developed in section 2 in single index model derivative estimations. The selected smoothing parameter, modified by a constant $(V_1/V_{1,3})^{1/7}$, asymptotically approaches the optimal smoothing parameter value that minimizes the estimation mean square error.

4. SIMULATION STUDY

We use Monte Carlo simulations to assess the finite sample performance of our proposed GBCV bandwidth selection mechanism. Here we will know the true unknown gradient and so a comparison to the oracle setting is feasible. We present both univariate and bivariate results to discern the impact that the dimensionality has on our proposed method.

4.1 Univariate Simulations

We consider the nonparametric model in (B.8) with homoskedastic error

$$y_j = g(x_j) + u_j,$$

for three different function specifications for $g(x)$:

DGP 1 $g\left(x\right) = 2 + \sin\left(1.5x\right)$

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

DGP 3 $g(x) = (x^4 - 0.1x^3 - 4.64x^2 + 1.324x + 0.408)/4.$

We use sample sizes of $n = 200$, 400 and 800 with 500 replications per experiment. Our covariate x is generated from $N(0, 0.8^2)$ and u is distributed $N(0, 0.5^2)$. We trim the top and bottom 2.5% of the data for all simulations when calculating the optimal bandwidth. That is, in (2.5) we have $M(x_j) = 1\left\{q_{\alpha/2} \leq x_j \leq q_{1-\alpha/2}\right\}$ where q_α is the α^{th} quantile of the data.[1]

[1] We solve the optimization using Powell's direction set algorithm with with a maximum of 100 function evaluations.

Our performance criteria is average squared error (ASE),

$$ASE(\hat{\beta}_{LL,A}) = n^{-1}\sum_{j=1}^{n}\left(\hat{\beta}_{LL}(x_j) - \beta_A(x_j)\right)^2,$$

where $\hat{\beta}_{LL}$ is the local-linear estimator of $\beta = dg(\cdot)/dx$ and β_A is one of the estimators from: (i) the local constant estimator, (ii) the local-quadratic estimator, (iii) the local-cubic estimator and (iv) the true gradient function. ASE is evaluated at the sample points for each simulation. We use the trimmed data points that are excluded when we engage in bandwidth selection when calculating average squared error.

Table 4.1 presents percentiles of ASE for the bandwidths selected by GBCV using local-constant, local-quadratic, local-cubic and the infeasible estimator over the 500 simulations for DGP 1. Each entry in the table provides the 10th, 50th and 90th percentile ASE in brackets for the method listed. The median ASEs provide insight into the general behavior of the bandwidth selection method while the extreme deciles provide insight into the tail performance of a given method across the simulations.

Table 4.1: Relative ASE for DGP 1 for GBCV selected bandwidths over 500 Simulations. Numbers in brackets are the 10th, 50th and 90th percentile of ASE across 500 simulations, respectively.

	$n = 200$	$n = 400$	$n = 800$
L-Constant	[0.072, 0.164, 0.424]	[0.069, 0.152, 0.362]	[0.059, 0.122, 0.267]
L-Quadratic	[0.071, 0.195, 0.589]	[0.059, 0.121, 0.389]	[0.047, 0.098, 0.339]
L-Cubic	[0.043, 0.080, 0.167]	[0.029, 0.055, 0.094]	[0.020, 0.036, 0.060]
Inf. True β	[0.038, 0.071, 0.138]	[0.026, 0.051, 0.085]	[0.018, 0.033, 0.057]

For DGP 1, we see that for all sample sizes, GBCV using local-cubic dominates both local-constant and local-quadratic at the median, and is very close to GBCV using the true gradients at the median. Further, we see that using local-cubic dominates (with respect to ASE) both local-constant and local-quadratic across all of the simulations. These gains increase as the sample size increases. An interesting pattern emerges amongst the lower and upper decile ASE ratios. As the sample size increases, the relative ASE between the local cubic and the local quadratic bandwidth estimators remains roughly constant at the lower decile while the relative ASE increases at the upper decile. Further, across all three local polynomial methods, the local cubic appears to uniformly dominate for this DGP.

Table 4.2 presents the same information for DGP 2. Here we see superior performance again of local-cubic GBCV, but not as great as with DGP 1. However, as with DGP 1, as the sample size increases local-cubic GBCV approaches the truth and still possesses gains over both local-constant and local-quadratic GBCV. The median ratio of ASE between local cubic and local quadratic hovers around 1.6 as the sample size increases while the upper decile ratio stays the same and the lower decile increases, a somewhat different pattern than was observed with DGP 1. This is to be expected as changes in curvature of the unknown regression function influences the cross validation criterion function.

Lastly, our performance metrics for DGP 3 appear in Table 4.3. As with the previous results, the bandwidth selected using local-cubic GBCV produces estimates which dominates those estimates using bandwidths produced with local-constant and local-quadratic GBCV. Compared with the local-constant and local-linear based estimators, the local-cubic based estimator displays large gains in performance, especially at the upper decile. The performance of bandwidths selected using the local-cubic estimator is roughly double at the median and almost triple at the upper decile in this setting.

Table 4.2: Relative ASE for DGP 2 for GBCV selected bandwidths over 500 Simulations. Numbers in brackets are the 10th, 50th and 90th percentile of ASE across 500 simulations, respectively.

	$n = 200$	$n = 400$	$n = 800$
L-Constant	[0.146, 0.276, 0.598]	[0.147, 0.244, 0.490]	[0.127, 0.213, 0.434]
L-Quadratic	[0.138, 0.313, 0.787]	[0.122, 0.209, 0.464]	[0.123, 0.196, 0.401]
L-Cubic	[0.107, 0.188, 0.348]	[0.093, 0.150, 0.212]	[0.088, 0.124, 0.173]
Inf. True β	[0.100, 0.180, 0.289]	[0.090, 0.143, 0.203]	[0.086, 0.122, 0.166]

Table 4.3: Relative ASE for DGP 3 for GBCV selected bandwidths over 500 Simulations. Numbers in brackets are the 10th, 50th and 90th percentile of ASE across 500 simulations, respectively.

	$n = 200$	$n = 400$	$n = 800$
L-Constant	[0.065, 0.160, 0.570]	[0.054, 0.115, 0.306]	[0.044, 0.094, 0.219]
L-Quadratic	[0.058, 0.136, 0.580]	[0.045, 0.101, 0.288]	[0.037, 0.075, 0.233]
L-Cubic	[0.044, 0.104, 0.199]	[0.035, 0.069, 0.127]	[0.029, 0.048, 0.085]
Inf. True β	[0.042, 0.094, 0.182]	[0.031, 0.064, 0.117]	[0.027, 0.046, 0.078]

Overall, our univariate simulation results confirm our theoretical conclusions. GBCV using the local-cubic estimator delivers bandwidths which behave as though one deployed the infeasible, known gradient of the unknown conditional mean. Further, the asymptotic biases of the local constant estimator and the local quadratic estimator have the same order as that of the local linear estimator. Hence, local constant and local quadratic estimators cannot be modified easily into an efficient estimator Indeed throughout the range of ASE produced in our simulations, not just at the median suggesting uniformly better performance when deploying the local cubic estimator.

4.2 Bivariate Simulations

Here we consider the bivariate nonparametric model:

$$y_j = g(x_{1j}, x_{2j}) + u_j, \quad j = 1, 2, \ldots, n.$$

where the function $g(x)$ is specified as

DGP 4 $g(x_1, x_2) = 1 + \sin(1.5x_1) + \frac{3e^{-3x_2}}{1+e^{-3x_2}}$.

We use sample sizes of $n = 200$, 400 and 800 with 500 replications per experiment. For all simulations x_{1j} and x_{2j} are generated from $N(0, 0.6^2)$ while the error term u_j is distributed $N(0, 0.5^2)$. Again, we trim the top and bottom 2.5% of the data (over both x_1 and x_2) for all simulations when calculating the optimal bandwidth, but we will use these points when calculating average squared error. To control the signal to noise ratio we construct our unknown function as $g^*(x_{1j}, x_{2j}) = g(x_{1j}, x_{2j})/\sigma(g(x_{1j}, x_{2j}))$. With the above distributional assumptions, this leads to a signal to noise ratio of approximately 0.80.

Table 4.4 presents the median and extreme decile ASEs for DGP 4 across the three local polynomial methods as well as the infeasible method. We note that the speed at which the local-cubic approaches the infeasible estimator is slower than in the univariate case, which is expected given the dimensionality. However, the main feature is that local-cubic is dominant compared to the both local-constant and local-quadratic and approaches the infeasible estimator.

Table 4.4: Relative ASE for Bivariate DGP 4 for GBCV over 500 Simulations.

	$n = 200$	$n = 400$	$n = 800$
L-Constant	[0.122, 0.190, 1.345]	[0.128, 0.185, 1.035]	[0.038, 0.086, 0.174]
L-Quadratic	[0.049, 0.108, 0.266]	[0.037, 0.074, 0.137]	[0.032, 0.055, 0.098]
L-Cubic	[0.049, 0.115, 0.220]	[0.035, 0.070, 0.123]	[0.025, 0.045, 0.075]
Inf. True β	[0.032, 0.070, 0.133]	[0.026, 0.050, 0.088]	[0.020, 0.036, 0.060]

We observe that for $n = 200$, the local-cubic estimator has a larger median ASE than that of the local quadratic estimator. This is a finite sample result because the local cubic estimator estimates more parameters using local data, hence, it may have a larger variance than that of a local quadratic estimator when sample size is not large. However, as expected, as the sample size increases, the gains of the local cubic estimator become apparent. In fact, while all three methods display decreases in the ASE as the sample size increases, the local cubic bandwidths produce an estimator whose ASE approaches that of the infeasible estimator the fastest. For example, for $n = 200$, the relative median ASE between the local constant estimator and the infeasible estimator is approximately 2.70 while for $n = 800$ this ratio is approximately 2.4. Alternatively, for the local cubic estimator, the $n = 200$ relative median ASE is 1.64 compared to $n = 800$ which produces a relative difference of 1.24. The reduction in the relative ASE of the local cubic bandwidth selection mechanism is roughly double the reduction in the relative ASE of the local constant estimator. Consistent with the theoretical underpinnings detailed above, the local cubic estimator behaves asymptotically the same as the infeasible estimator.

5. CONCLUSIONS

In this dissertation we propose a novel approach to select bandwidths in nonparametric kernel regression and extend it to semiparametric single index model. In contrast to previous research on bandwidth selection focusing on the unknown conditional mean, we are primarily concerned with estimation of the gradient function. Uncovering gradients nonparametrically is important in many areas of economics such as determining risk premium or recovering distributions of individual preferences. Estimation of gradients is often of more practical interest as studying 'marginal effects' is a cornerstone of applied econometric analysis. Our procedure is shown to deliver bandwidths which behave asymptotically equivalently to the infeasible selection procedure where the true gradient is used. Our simulations show that determining the optimal bandwidth by using the local-cubic estimator to construct an estimate of the unknown gradient delivers finite sample performance on par with the bandwidth selected using the actual, unknown gradient.

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

REFERENCES

P. Bajari and M. E. Kahn. Estimating housing demand with an application to explaining racial segregation in cities. *Journal of Business and Economic Statistics*, 23(1):20–33, 2005.

Z. Cai, J. Fan, and Q. Yao. Functional coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956, 2000.

R. Charnigo, B. Hall, and C. Srinivasan. A generalized c_p criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011.

J. Chen, J. Gao, and D. Li. Estimation in single-index panel data models with heterogeneous link functions. *Econometric Reviews*, 32(8):928–955, 2013.

P. A. Chiapporis, A. Gandhi, B. Salanié, and F. Salanié. Identifying preferences under risk from discrete choices. *American Economic Review*, 99(2):356–362, 2009.

J. Fan and I. Gijbels. 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, 1995.

J. Fan and I. Gijbels. *Local Polynomial Modeling and its Applications.* Chapman and Hall, London, 1996.

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

P. Hall, Q. Li, and J. S. Racine. Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics*, 89(4):784–789, 2007.

J. J. Heckman, R. L. Matzkin, and L. Nesheim. Nonparametric identification and estimation of nonadditive hedonic models. *Econometrica*, 78(5):1569–1591, 2010.

D. J. Henderson, Q. Li, and C. F. Parmeter. Gradient based smoothing parameter selection for nonparametric regression estimation. University of Miami, Department of Economics Working Paper WP2014-01, 2012.

E. Masry. Multivariate regression estimation Local polynomial fitting for time series. *Stochastic Processes and their Applications*, 65(1):81–101, 1996.

H.-G. Müller, U. Stadmüller, and T. Schmitt. Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, 74(4):743–749, 1987.

J. S. Racine and Q. Li. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130, 2004.

J. A. Rice. Bandwidth choice for differentiation. *Journal of Multivariate Analysis*, 19:251–264, 1986.

D. Ruppert. Empirical-bias bandwidth for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92(439):1049–1062, 1997.

G. Wahba and Y. Wang. When is the optimal regularization parameter insensitive to the choice of loss function? *Communications in Statistics*, 19:1685–1700, 1990.

Y. Xia and W. Härdle. Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97(5):1162–1184, 2006.

APPENDIX A

PROOF OF EQUATION (2.8)

In this appendix we show the calculation of the covariance $Cov(\hat{\beta}_{LL}(x), \hat{\beta}_{L\text{-}cubic}(x))$ as defined in (2.8). For this we need to derive local-cubic and local-linear estimator of $\beta(x) = \partial g(x)/\partial x$. We begin with the local-cubic estimator. Taking a Taylor expansion of $g(x_j)$ at x, we can rewrite (2.1) as

$$\begin{aligned} y_j &= g(x_j) + u_j \\ &= g(x) + g'(x)(x_j - x) + \frac{1}{2}g''(x)(x_j - x)^2 + \frac{1}{6}g'''(x)(x_j - x)^3 + R_{jx} + u_j \end{aligned} \tag{A.1}$$

where $g'(x) = \partial g(x)/\partial x$, $g''(x) = \partial^2 g(x)/\partial x^2$ and $g'''(x) = \partial^3 g(x)/\partial x^3$. The local-cubic estimator of $(g(x), g'(x)h, \frac{1}{2}g''(x)h^2, \frac{1}{6}g'''(x)h^3)^T$ is obtained by choosing $(a, b, c, d)^T$ to minimize the following objective function

$$\min_{a,b,c,d} \sum_{j=1}^{n} \left[y_j - a - b\frac{x_j - x}{h} - c\frac{(x_j - x)^2}{h^2} - d\frac{(x_j - x)^3}{h^3} \right]^2 W_{h,jx} \tag{A.2}$$

The first order condition (normal equation) to the minimization problem $(A.2)$ is:

$$\sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j - x}{h} \\ \frac{(x_j - x)^2}{h^2} \\ \frac{(x_j - x)^3}{h^3} \end{pmatrix} \left\{ y_j - \left[1, \frac{x_j - x}{h}, \frac{(x_j - x)^2}{h^2}, \frac{(x_j - x)^3}{h^3}\right] \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \right\} = 0$$

which leads to the closed form solution of

$(a,b,c,d)^T = (g(x), g'(x)h, \frac{1}{2}g''(x)h^2, \frac{1}{6}g'''(x)h^3)^T$ given by

$$\begin{aligned}
\begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \\ \hat{d} \end{pmatrix} &= \begin{pmatrix} \hat{g}(x) \\ \widehat{g'}(x)h \\ \frac{1}{2}\widehat{g''}(x)h^2 \\ \frac{1}{6}\widehat{g'''}(x)h^3 \end{pmatrix} \\
&= \left\{ \sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j-x}{h} \\ \frac{(x_j-x)^2}{h^2} \\ \frac{(x_j-x)^3}{h^3} \end{pmatrix} \left[1, \frac{x_j-x}{h}, \frac{(x_j-x)^2}{h^2}, \frac{(x_j-x)^3}{h^3}\right] \right\}^{-1} \\
&\quad \times \sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j-x}{h} \\ \frac{(x_j-x)^2}{h^2} \\ \frac{(x_j-x)^3}{h^3} \end{pmatrix} y_j. \qquad \text{(A.3)}
\end{aligned}$$

Substitute y_j in (A.3) with (A.1), and re-arrange terms, leads to

$$\begin{pmatrix} \hat{g}(x) - g(x) \\ \left[\widehat{g'}(x) - g'(x)\right] h \\ \frac{1}{2}\left[\widehat{g''}(x) - g''(x)\right] h^2 \\ \frac{1}{6}\left[\widehat{g'''}(x) - g'''(x)\right] h^3 \end{pmatrix} = A_{2,x}^{-1} A_{1,x},$$

where

$$A_{1,x} = \frac{1}{n} \sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j-x}{h} \\ \frac{(x_j-x)^2}{h^2} \\ \frac{(x_j-x)^3}{h^3} \end{pmatrix} (R_{jx} + u_j), \qquad \text{(A.4)}$$

and

$$A_{2,x} = \frac{1}{n}\sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j - x}{h} \\ \frac{(x_j-x)^2}{h^2} \\ \frac{(x_j-x)^3}{h^3} \end{pmatrix} \left[1, \frac{x_j - x}{h}, \frac{(x_j-x)^2}{h^2}, \frac{(x_j-x)^3}{h^3}\right].$$

Using the standard kernel estimation uniform convergence proof techniques, we have

$$\begin{aligned} A_{2,x} &= \begin{pmatrix} f(x) & h\mu_2 f'(x) & \mu_2 f(x) & h\mu_4 f'(x) \\ h\mu_2 f'(x) & \mu_2 f(x) & h\mu_4 f'(x) & \mu_4 f(x) \\ \mu_2 f(x) & h\mu_4 f'(x) & \mu_4 f(x) & h\mu_6 f'(x) \\ h\mu_4 f'(x) & \mu_4 f(x) & h\mu_6 f'(x) & \mu_6 f(x) \end{pmatrix} + o(h) \\ &= H_x + hF_x + o(h) \end{aligned}$$

uniformly in $x \in \mathcal{M}$, where $\mathcal{M}$ is the (bounded) support of the trimming function,

$$H_x = f(x)\begin{pmatrix} 1 & 0 & \mu_2 & 0 \\ 0 & \mu_2 & 0 & \mu_4 \\ \mu_2 & 0 & \mu_4 & 0 \\ 0 & \mu_4 & 0 & \mu_6 \end{pmatrix},$$

and

$$F_x = f'(x)\begin{pmatrix} 0 & \mu_2 & 0 & \mu_4 \\ \mu_2 & 0 & \mu_4 & 0 \\ 0 & \mu_4 & 0 & \mu_6 \\ \mu_4 & 0 & \mu_6 & 0 \end{pmatrix}.$$

Using the identity $\{H_x + hF_x + o(h)\}^{-1} = H_x^{-1} - hH_x^{-1}F_xH_x^{-1} + o(h)$, we obtain

$$\begin{aligned} \left[\widehat{g'}(x) - g'(x)\right] h &= (0,1,0,0)A_{2,x}^{-1}A_{1,x} \\ &= (0,1,0,0)\left[H_x^{-1} - hH_x^{-1}F_xH_x^{-1}\right]A_{1,x} + (s.o.) \end{aligned}$$

$$= \; (\, D_{1,x}h, \quad D_{2,x}, \quad D_{3,x}h, \quad D_{4,x} \,)\, A_{1,x} + (s.o.),$$

where $(s.o.)$ denotes negligible smaller order terms. Then by using (A.4) we obtain

$$\begin{aligned} & \widehat{g}'(x) - g'(x) \\ = \; & \frac{1}{nh}\sum_{j=1}^{n} W_{h,jx}\left[D_{1,x}h + D_{2,x}\frac{x_j - x}{h} + D_{3,x}\frac{(x_j-x)^2}{h} + D_{4,x}\frac{(x_j-x)^3}{h^3}\right](R_{jx} + u_j) \end{aligned} \tag{A.5}$$

where

$$D_{1,x} = -(\mu_2\mu_4\mu_6 - \mu_4^3)(\mu_2\mu_4\mu_6 - \mu_2^3\mu_6 + \mu_2^2\mu_4^2 - \mu_4^3)\frac{f'(x)}{f^2(x)K_2^2},$$

$$D_{2,x} = (\mu_4\mu_6 - \mu_2^2\mu_6)\frac{1}{f(x)K_2},$$

$$D_{3,x} = -(\mu_2\mu_4^2 - \mu_2^2\mu_6)(\mu_2\mu_4\mu_6 - \mu_2^3\mu_6 + \mu_2^2\mu_4^2 - \mu_4^3)\frac{f'(x)}{f^2(x)K_2^2},$$

$$D_{4,x} = (\mu_2^2\mu_4 - \mu_4^2)\frac{1}{f(x)K_2},$$

$$K_2 = \mu_2\mu_4\mu_6 - \mu_4^3 + \mu_2^2\mu_4^2 - \mu_2^3\mu_6.$$

Next, we derive the leading terms of the local-linear estimator. Again, taking a Taylor expansion of $g(x_j)$ at x, we can rewrite (2.1) as

$$\begin{aligned} y_j &= g(x_j) + u_j \\ &= g(x) + g'(x)(x_j - x) + \eta_{jx} + u_j \end{aligned} \tag{A.6}$$

where $g'(x) = \partial g(x)/\partial x$. The local-linear estimator of $(g(x), g'(x)h)^T$ is obtained by choosing $(a,b)^T$ to minimize the following objective function

$$\min_{a,b} \sum_{j=1}^{n} \left(y_j - a - b\frac{x_j - x}{h} \right)^2 W_{h,jx}. \tag{A.7}$$

The first order condition (normal equation) to the minimization problem $(A.7)$ is:

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

which leads to the closed form solution of $(a,b)^T = (g(x), g'(x)h)^T$ given by

$$\begin{aligned} \begin{pmatrix} \tilde{a} \\ \tilde{b} \end{pmatrix} &= \begin{pmatrix} \tilde{g}(x) \\ \widetilde{g}'(x)h \end{pmatrix} \\ &= \left[\sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j-x}{h} \end{pmatrix} \left(1, \frac{x_j - x}{h}\right) \right]^{-1} \sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j-x}{h} \end{pmatrix} y_j. \end{aligned} \tag{A.8}$$

Substitute y_j in (A.8) with (A.6) and re-arrange terms, leads to

$$\begin{pmatrix} \tilde{g}(x) - g(x) \\ \left[\widetilde{g}'(x) - g'(x)\right] h \end{pmatrix} = G_{2,x}^{-1} G_{1,x},$$

where

$$G_{1,x} = \frac{1}{n} \sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j-x}{h} \end{pmatrix} (\eta_{jx} + u_j), \tag{A.9}$$

and

$$G_{2,x} = \frac{1}{n} \sum_{j=1}^{n} W_{h,jx} \begin{pmatrix} 1 \\ \frac{x_j-x}{h} \end{pmatrix} \left(1, \frac{x_j - x}{h}\right).$$

Using the standard kernel estimation uniform convergence proof techniques, we have

$$
\begin{aligned}
G_{2,x} &= \begin{pmatrix} f(x) & h\mu_2 f'(x) \\ h\mu_2 f'(x) & \mu_2 f(x) \end{pmatrix} + o(h) \\
&= J_x + hL_x + o(h),
\end{aligned}
$$

where

$$J_x = f(x) \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix},$$

and

$$L_x = \mu_2 f'(x) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Using the identity $\{J_x + hL_x + o(h)\}^{-1} = J_x^{-1} - hJ_x^{-1}L_xJ_x^{-1} + o(h)$, we obtain

$$
\begin{aligned}
\left[\widetilde{g}'(x) - g'(x)\right] h &= (0,1)G_{2,x}^{-1}G_{1,x} \\
&= (0,1)\left[J_x^{-1} - hJ_x^{-1}L_xJ_x^{-1}\right] G_{1,x} + (s.o.) \\
&= (0,1) \begin{pmatrix} \frac{1}{f(x)} & -h\frac{f'(x)}{f^2(x)} \\ -h\frac{f'(x)}{f^2(x)} & \frac{1}{\mu_2 f(x)} \end{pmatrix} G_{1,x} + (s.o.) \\
&= \begin{pmatrix} -hC_{1,x}, & C_{2,x} \end{pmatrix} G_{1,x} + (s.o.),
\end{aligned}
$$

where $C_{1,x} = \frac{f'(x)}{f^2(x)}$ and $C_{2,x} = \frac{1}{\mu_2 f(x)}$. Thus by using (A.9) we get

$$\widetilde{g}'(x) - g'(x) = \frac{1}{n}\sum_{j=1}^{n} W_{h,jx}\left(C_{2,x}\frac{x_j - x}{h^2} - C_{1,x}\right)(\eta_{jx} + u_j). \tag{A.10}$$

In equations (A.5) and (A.10), R_{jx} and η_{jx} are associated with the bias term and u_j is associated with the variance. The leading covariance term comes from the terms associated with u_j in (A.5) and (A.10). Hence, we have

$$Cov(\hat{\beta}_{LL}(x), \hat{\beta}_{L\text{-}cubic}(x))$$

$$
\begin{aligned}
&= E\left\{ \frac{1}{n}\sum_{j=1}^{n} W_{h,jx}\left(C_{2,x}\frac{x_j-x}{h^2}-C_{1,x}\right)u_j \right.\\
&\qquad \left. \times \frac{1}{nh}\sum_{k=1}^{n} W_{h,kx}\left[D_{1,x}h+D_{2,x}\frac{x_k-x}{h}+D_{3,x}\frac{(x_k-x)^2}{h}+D_{4,x}\frac{(x_k-x)^3}{h^3}\right]u_k\right\}\\
&= \frac{1}{nh}E\left\{W_{h,jx}^2u_j^2\left(C_{2,x}\frac{x_j-x}{h^2}-C_{1,x}\right)\right.\\
&\qquad \left. \times\left[D_{1,x}h+D_{2,x}\frac{x_k-x}{h}+D_{3,x}\frac{(x_k-x)^2}{h}+D_{4,x}\frac{(x_k-x)^3}{h^3}\right]\right\}\\
&= \frac{1}{nh^2}\int m(x+hv)W^2(v)(C_{2,x}h^{-1}v-C_{1,x})\left[(D_{2,x}v+D_{4,x}v^3)+(D_{1,x}+D_{3,x}v^2)h\right]dv\\
&= \frac{1}{nh^2}m(x)\int W^2(v)C_{2,x}h^{-1}(D_{2,x}v^2+D_{4,x}v^4)dv+(s.o.)\\
&= \frac{1}{nh^3}f(x)\sigma^2(x)C_{2,x}(D_{2,x}\nu_2+D_{4,x}\nu_4)+(s.o.)\\
&= \frac{1}{nh^3}\cdot\frac{(\mu_4\mu_6-\mu_2^2\mu_6)\nu_2+(\mu_2^2\mu_4-\mu_4^2)\nu_4}{\mu_2K_2}\cdot\frac{\sigma^2(x)}{f(x)}+(s.o.),
\end{aligned}
$$

where $m(x)=f(x)\sigma^2(x)$, $\sigma^2(x)=E(u^2|x)$ and $K_2=\mu_2\mu_4\mu_6-\mu_4^3+\mu_2^2\mu_4^2-\mu_2^3\mu_6$.

APPENDIX B

PROOF OF LEMMA 3.2.1

We first decompose $CV_f(h)$ as follow

$$\begin{aligned}
CV_f(h) &= \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{Cubic}(\hat{z}_i)\right]^2 M(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big\{[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{LL}(z_i)] - [\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)] \\
&\qquad + [\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)]\Big\}^2 M(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{LL}(z_i)]^2 M(x_i) + \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)]^2 M(x_i) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)]^2 M(x_i) \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)] M(x_i) \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)] M(x_i) \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)] M(x_i) \\
&= C_1 + C_2 + C_3 - C_4 - C_5 + C_6 \qquad \text{(B.1)}
\end{aligned}$$

where

$$C_1 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{LL}(z_i)]^2 M(x_i) \qquad \text{(B.2)}$$

$$C_2 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)]^2 M(x_i) \qquad \text{(B.3)}$$

$$C_3 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)]^2 M(x_i) \tag{B.4}$$

$$C_4 = \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)]M(x_i) \tag{B.5}$$

$$C_5 = \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)]M(x_i) \tag{B.6}$$

$$C_6 = \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)]M(x_i) \tag{B.7}$$

Lemma B.1 $\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{LL}(z_i) = O_p\left(\frac{1}{\sqrt{n}}\right)$, *where* $\hat{\beta}_{LL}(\cdot)$ *is the local linear estimator,* $\hat{z}_i = x_i^\top\hat{\gamma}$, $z_i = x_i^\top\gamma_0$, $\hat{\gamma}$ *is the minimum average variance (MAV) estimator of* γ *(see Xia & Härdle 2006) and* γ_0 *is the true value of* γ.

Proof of Lemma B.1: We begin with the single index model

$$y_j = g(z_j) + u_j \tag{B.8}$$

where $z_j = X_j^\top\gamma$ is a scalar.

We define a 2×1 vector $\delta(z)$ by

$$\delta(z) = \begin{pmatrix} g(z) \\ \beta(z) \end{pmatrix}, \tag{B.9}$$

where the first component of $\delta(z)$ is $g(z)$ and the second component is the first derivative of $g(z)$ w.r.t. z. Taking a Taylor series expansion of $g(z_j)$ at z_i, we get

$$g(z_j) = g(z_i) + (z_j - z_i)\beta(z_i) + R_{ji}, \tag{B.10}$$

Note that (B.10) defines R_{ji}, i.e.,

$$R_{ji} = g(z_j) - g(z_i) - (z_j - z_i)\beta(z_i). \tag{B.11}$$

Using (B.10) we can re-write (B.8) as

$$y_j = g(z_i) + (z_j - z_i)\beta(z_i) + R_{ji} + u_j = (1, (z_j - z_i))\ \delta(z_i) + R_{ji} + u_j. \tag{B.12}$$

The local linear estimator of $\delta(z) = (g(z), \beta(z))^\top$ is obtained by choosing $(a, b)^\top \in \mathcal{R}^2$ to minimize the following objective function

$$\min_{a,b} \sum_{j=1}^{n} [\, y_j - a - (z_j - z)b\,]^2\ W_{h,jz}, \tag{B.13}$$

where $W_{h,jz} = h^{-1} w\left(\frac{z_j - z}{h}\right)$ is a univariate kernel function.

The first-order condition (normal equations) to the minimization problem (B.13) is:

$$\sum_{j=1}^{n} \begin{pmatrix} 1 \\ z_j - z \end{pmatrix} [\, y_j - a - (z_j - z)\, b\,]\ W_{h,jz} = 0, \tag{B.14}$$

which leads to the closed form solution of $\hat{\delta}_{LL}(z) = (\hat{a}, \hat{b})^\top \equiv (\hat{g}_{LL}(z), \hat{\beta}_{LL}(z))^\top$ given by

$$\hat{\delta}_{LL}(z) = \begin{pmatrix} \hat{g}_{LL}(z) \\ \hat{\beta}_{LL}(z) \end{pmatrix} = \left[\sum_{j=1}^{n} W_{h,jz} \begin{pmatrix} 1, & z_j - z \\ z_j - z, & (z_j - z)^2 \end{pmatrix} \right]^{-1} \sum_{l=1}^{n} W_{h,jz} \begin{pmatrix} 1 \\ z_j - z \end{pmatrix} y_j, \tag{B.15}$$

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

$$\hat{\delta}_{-i,LL}(z_i) = \begin{pmatrix} \hat{g}_{-i,LL}(z_i) \\ \hat{\beta}_{-i,LL}(z_i) \end{pmatrix} = \left[\sum_{j\neq i}^{n} W_{h,ji} \begin{pmatrix} 1, & z_j - z_i \\ z_j - z_i, & (z_j - z_i)^2 \end{pmatrix} \right]^{-1} \sum_{j\neq i}^{n} W_{h,ji} \begin{pmatrix} 1 \\ z_j - z_i \end{pmatrix} y_j, \tag{B.16}$$

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

$$\begin{aligned} y_j &= g(z_j) + u_j = g(z_i) + (z_j - z_i)\beta(z_i) + R_{ji} + u_j \\ &= (\,1, (z_j - z_i)\,) \begin{pmatrix} g(z_i) \\ \beta(z_i) \end{pmatrix} + R_{ji} + u_j. \end{aligned} \tag{B.17}$$

Substituting y_j in (B.16) with (B.17), leads to

$$\hat{\delta}_{-i,LL}(z_i) = \delta(z_i) + \left[\sum_{j\neq i}^{n} W_{h,ji} \begin{pmatrix} 1, & z_j - z_i \\ z_j - z_i, & (z_j - z_i)^2 \end{pmatrix} \right]^{-1} \sum_{j\neq i}^{n} W_{h,ji} \begin{pmatrix} 1 \\ z_j - z_i \end{pmatrix} (R_{ji} + u_j) \tag{B.18}$$

Multiple (B.18) by 1×2 matrix (0,1), we get

$$\hat{\beta}_{-i,LL}(z_i) = \beta(z_i) + B(z_i) \tag{B.19}$$

where

$$B(z_i) = (0,1) \left[\sum_{j\neq i}^{n} W_{h,ji} \begin{pmatrix} 1, & z_j - z_i \\ z_j - z_i, & (z_j - z_i)^2 \end{pmatrix} \right]^{-1} \sum_{j\neq i}^{n} W_{h,ji} \begin{pmatrix} 1 \\ z_j - z_i \end{pmatrix} (R_{ji} + u_j) \tag{B.20}$$

It is well established that $\hat{\beta}_{-i,LL}(z_i) - \beta(z_i) = B(z_i) = O_p\left(h^2 + \frac{1}{\sqrt{nh^3}}\right) = o_p(1)$, see Cai, Fan & Yao (2000), Henderson, Li & Parmeter (2012) and Fan & Gijbels (1996) Theorem 3.1.

Substituting z_i for $\hat{z}_i$ in (B.19), we have

$$\hat{\beta}_{-i,LL}(\hat{z}_i) = \beta(\hat{z}_i) + B(\hat{z}_i) \tag{B.21}$$

Taking a Taylor expansion of $B(\hat{z}_i)$ at z_i, we get

$$B(\hat{z}_i) = B(z_i) + B'(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.22}$$

where $B'(z_i) = dB(z_i)/dz_i$ and (s.o.) denotes smaller order terms.

Thus we have

$$\hat{\beta}_{-i,LL}(\hat{z}_i) = \beta(\hat{z}_i) + B(z_i) + B'(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.23}$$

Deducting (B.19) from (B.23), we get

$$\hat{\beta}_{-i,LL}(\hat{z}_i) - \hat{\beta}_{-i,LL}(z_i) = \beta(\hat{z}_i) - \beta(z_i) + B'(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.24}$$

Next, we analyze $\beta(\hat{z}_i) - \beta(z_i)$ and $B'(z_i)(\hat{z}_i - z_i)$ one by one. First, we analyze $\beta(\hat{z}_i) - \beta(z_i)$. Taking a Taylor expansion of $\beta(\hat{z}_i)$ at z_i, we get

$$\beta(\hat{z}_i) = \beta(z_i) + \beta'(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.25}$$

where $\beta'(z_i) = d\beta(z_i)/dz_i$.

Thus

$$\beta(\hat{z}_i) - \beta(z_i) \quad = \quad \beta'(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.26}$$

$$= \beta'(z_i)X_i^\top(\hat{\gamma} - \gamma_0) + (s.o.) \tag{B.27}$$

$$= O_p\left(\frac{1}{\sqrt{n}}\right) \tag{B.28}$$

where the last equality uses the fact that the MAV estimator $\hat{\gamma}$ is a $\sqrt{n}$-consistent estimator, i.e. $\hat{\gamma} - \gamma_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$.

Next, we analyze $B'(z_i)(\hat{z}_i - z_i)$. We note that $B'(z_i) = o_p(1)$ or $O_p(1)$. Thus

$$\begin{aligned} B'(z_i)(\hat{z}_i - z_i) &= B'(z_i)X_i^\top(\hat{\gamma} - \gamma_0) && \text{(B.29)} \\ &= o_p\left(\frac{1}{\sqrt{n}}\right) \text{ or } O_p\left(\frac{1}{\sqrt{n}}\right) && \text{(B.30)} \end{aligned}$$

Summarizing (B.24), (B.28) and (B.30), we have

$$\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{LL}(z_i) = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{B.31}$$

Lemma B.2 *$\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i) = O_p(\frac{1}{\sqrt{n}})$, where $\hat{\beta}_{Cubic}(\cdot)$ is the local cubic estimator, $\hat{z}_i = x_i^\top\hat{\gamma}$, $z_i = x_i^\top\gamma_0$, $\hat{\gamma}$ is the minimum average variance (MAV) estimator of γ (see Xia & Härdle 2006) and γ_0 is the true value of γ.*

Proof of Lemma B.2: Again, we begin with the single index model:

$$y_j = g(z_j) + u_j \tag{B.32}$$

where $z_j = X_j^\top\gamma$ is a scalar.

We define a 4×1 vector $\delta^*(z)$ by

$$\delta^*(z) = \begin{pmatrix} g(z) \\ \beta(z) \\ \alpha(z) \\ \theta(z) \end{pmatrix}, \tag{B.33}$$

where the first component of $\delta^*(z)$ is $g(z)$; the second component is $\frac{\partial g(z)}{\partial z}$; the third component is $\frac{1}{2}\frac{\partial g^{(2)}(z)}{\partial z^2}$; the fourth component is $\frac{1}{6}\frac{\partial g^{(3)}(z)}{\partial z^3}$. Taking a Taylor series expansion of $g(z_i)$ at z_i, we get

$$g(z_j) = g(z_i) + \beta(z_i)(z_j - z_i) + \alpha(z_i)(z_j - z_i)^2 + \theta(z_i)(z_j - z_i)^3 + R^*_{ji} \tag{B.34}$$

where $\beta(z) = \frac{\partial g(z)}{\partial z}$; $\alpha(z) = \frac{1}{2}\frac{\partial g^{(2)}(z)}{\partial z^2}$; $\theta(z) = \frac{1}{6}\frac{\partial g^{(3)}(z)}{\partial z^3}$.

Note that (B.34) defines R^*_{ji}, i.e.,

$$R^*_{ji} = g(z_j) - g(z_i) - \beta(z_i)(z_j - z_i) - \alpha(z_i)(z_j - z_i)^2 - \theta(z_i)(z_j - z_i)^3 + R^*_{ji}. \tag{B.35}$$

Using (B.10) we can re-write (B.32) as

$$\begin{aligned} y_j &= g(z_i) + \beta(z_i)(z_j - z_i) + \alpha(z_i)(z_j - z_i)^2 + \theta(z_i)(z_j - z_i)^3 + R^*_{ji} + u_j \\ &= (1, (z_j - z_i), (z_j - z_i)^2, (z_j - z_i)^3)\ \delta^*(z_i) + R^*_{ji} + u_j. \end{aligned} \tag{B.36}$$

The local cubic estimator of $\delta^*(z) = (g(z), \beta(z), \alpha(z), \theta(z))^\top$ is obtained by choosing $(a, b, c, d)^\top \in \mathcal{R}^4$ to minimize the following objective function

$$\min_{a,b,c,d} \sum_{j=1}^{n} \left[y_j - a - (z_j - z)b - (z_j - z)^2 c - (z_j - z)^3 d \right]^2 \ W_{h,jz}, \tag{B.37}$$

where $W_{h,jz} = h^{-1} w\left(\frac{z_j - z}{h}\right)$ is a univariate kernel function.

The first-order condition (normal equations) to the minimization problem (B.37) is:

$$\sum_{j=1}^{n} \begin{pmatrix} 1 \\ z_j - z \\ (z_j - z)^2 \\ (z_j - z)^3 \end{pmatrix} \left[y_j - a - (z_j - z)b - (z_j - z)^2 c - (z_j - z)^3 d \right] W_{h,jz} = 0, \quad \text{(B.38)}$$

which leads to the closed form solution of

$\hat{\delta}^*_{Cubic}(z) = (\hat{a}, \hat{b}, \hat{c}, \hat{d})^\top \equiv (\hat{g}_{Cubic}(z), \hat{\beta}_{Cubic}(z), \hat{\alpha}_{Cubic}(z), \hat{\theta}_{Cubic}(z))^\top$ given by

$$\begin{aligned} \hat{\delta}^*_{Cubic}(z) &= \begin{pmatrix} \hat{g}_{Cubic}(z) \\ \hat{\beta}_{Cubic}(z) \\ \hat{\alpha}_{Cubic}(z) \\ \hat{\theta}_{Cubic}(z) \end{pmatrix} && \text{(B.39)} \\ &= \left[\sum_{j=1}^{n} W_{h,jz} \begin{pmatrix} 1, & z_j - z, & (z_j - z)^2, & (z_j - z)^3 \\ z_j - z, & (z_j - z)^2, & (z_j - z)^3, & (z_j - z)^4 \\ (z_j - z)^2, & (z_j - z)^3, & (z_j - z)^4, & (z_j - z)^5 \\ (z_j - z)^3, & (z_j - z)^4, & (z_j - z)^5, & (z_j - z)^6 \end{pmatrix} \right]^{-1} \\ &\times \sum_{l=1}^{n} W_{h,jz} \begin{pmatrix} 1 \\ z_j - z \\ (z_j - z)^2 \\ (z_j - z)^3 \end{pmatrix} y_j, && \text{(B.40)} \end{aligned}$$

A leave-one-out local cubic kernel estimator of $\delta^*(z_i)$ is obtained by replacing z with z_i and replacing $\sum_{j=1}^{n}$ by $\sum_{j \neq i}^{n}$.

$$
\begin{aligned}
\hat{\delta}^*_{-i,Cubic}(z) &= \begin{pmatrix} \hat{g}_{-i,Cubic}(z) \\ \hat{\beta}_{-i,Cubic}(z) \\ \hat{\alpha}_{-i,Cubic}(z) \\ \hat{\theta}_{-i,Cubic}(z) \end{pmatrix} && \text{(B.41)} \\
&= \left[\sum_{j=1}^{n} W_{h,ji} \begin{pmatrix} 1, & z_j - z_i, & (z_j - z_i)^2, & (z_j - z_i)^3 \\ z_j - z_i, & (z_j - z_i)^2, & (z_j - z_i)^3, & (z_j - z_i)^4 \\ (z_j - z_i)^2, & (z_j - z_i)^3, & (z_j - z_i)^4, & (z_j - z_i)^5 \\ (z_j - z_i)^3, & (z_j - z_i)^4, & (z_j - z_i)^5, & (z_j - z_i)^6 \end{pmatrix} \right]^{-1} \\
&\times \sum_{l=1}^{n} W_{h,ji} \begin{pmatrix} 1 \\ z_j - z_i \\ (z_j - z_i)^2 \\ (z_j - z_i)^3 \end{pmatrix} y_j, && \text{(B.42)}
\end{aligned}
$$

Recall that $R^*_{ji} = g(z_j) - g(z_i) - \beta(z_i)(z_j - z_i) - \alpha(z_i)(z_j - z_i)^2 - \theta(z_i)(z_j - z_i)^3 + R^*_{ji}$. We can write y_j as

$$
\begin{aligned}
y_j &= g(z_j) + u_j \\
&= g(z_i) + \beta(z_i)(z_j - z_i) + \alpha(z_i)(z_j - z_i)^2 + \theta(z_i)(z_j - z_i)^3 + R^*_{ji} + u_j \\
&= (\, 1, z_j - z_i, (z_j - z_i)^2, (z_j - z_i)^3 \,) \begin{pmatrix} g(z_i) \\ \beta(z_i \\ \alpha(z_i) \\ \theta(z_i)) \end{pmatrix} + R^*_{ji} + u_j. && \text{(B.43)}
\end{aligned}
$$

Substituting y_j in (B.42) with (B.43), leads to

$$
\begin{aligned}
&\hat{\delta}^*_{-i,Cubic}(z_i) \\
= \quad & \delta^*(z_i) + \left[\sum_{j=1}^{n} W_{h,ji} \begin{pmatrix} 1, & z_j - z_i, & (z_j - z_i)^2, & (z_j - z_i)^3 \\ z_j - z_i, & (z_j - z_i)^2, & (z_j - z_i)^3, & (z_j - z_i)^4 \\ (z_j - z_i)^2, & (z_j - z_i)^3, & (z_j - z_i)^4, & (z_j - z_i)^5 \\ (z_j - z_i)^3, & (z_j - z_i)^4, & (z_j - z_i)^5, & (z_j - z_i)^6 \end{pmatrix} \right]^{-1} \\
& \times \sum_{l=1}^{n} W_{h,ji} \begin{pmatrix} 1 \\ z_j - z_i \\ (z_j - z_i)^2 \\ (z_j - z_i)^3 \end{pmatrix} (R^*_{ji} + u_j) \qquad \text{(B.44)}
\end{aligned}
$$

Multiple (B.44) by 1×4 matrix (0,1,0,0), we get

$$\hat{\beta}_{-i,Cubic}(z_i) = \beta(z_i) + B^*(z_i) \qquad \text{(B.45)}$$

where

$$
\begin{aligned}
&B^*(z_i) \\
= \quad & (0,1,0,0) \left[\sum_{j \neq i}^{n} W_{h,ji} \begin{pmatrix} 1, & z_j - z_i, & (z_j - z_i)^2, & (z_j - z_i)^3 \\ z_j - z_i, & (z_j - z_i)^2, & (z_j - z_i)^3, & (z_j - z_i)^4 \\ (z_j - z_i)^2, & (z_j - z_i)^3, & (z_j - z_i)^4, & (z_j - z_i)^5 \\ (z_j - z_i)^3, & (z_j - z_i)^4, & (z_j - z_i)^5, & (z_j - z_i)^6 \end{pmatrix} \right]^{-1} \\
& \times \sum_{l=1}^{n} W_{h,ji} \begin{pmatrix} 1 \\ z_j - z_i \\ (z_j - z_i)^2 \\ (z_j - z_i)^3 \end{pmatrix} (R^*_{ji} + u_j) \qquad \text{(B.46)}
\end{aligned}
$$

It is well established that $\hat{\beta}_{-i,Cubic}(z_i) - \beta(z_i) = B^*(z_i) = O_p\left(h^4 + \frac{1}{\sqrt{nh^3}}\right) = o_p(1)$, see Fan & Gijbels (1996) Theorem 3.1.

Substituting z_i for $\hat{z}_i$ in (B.45), we have

$$\hat{\beta}_{-i,Cubic}(\hat{z}_i) = \beta(\hat{z}_i) + B^*(\hat{z}_i) \tag{B.47}$$

Taking a Taylor expansion of $B^*(\hat{z}_i)$ at z_i, we get

$$B^*(\hat{z}_i) = B^*(z_i) + B^{*\prime}(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.48}$$

where $B^{*\prime}(z_i) = dB^*(z_i)/dz_i$ and (s.o.) denotes smaller order terms.

Thus we have

$$\hat{\beta}_{-i,Cubic}(\hat{z}_i) = \beta(\hat{z}_i) + B^*(z_i) + B^{*\prime}(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.49}$$

Deducting (B.45) from (B.49), we get

$$\hat{\beta}_{-i,Cubic}(\hat{z}_i) - \hat{\beta}_{-i,Cubic}(z_i) = \beta(\hat{z}_i) - \beta(z_i) + B^{*\prime}(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.50}$$

Next, we analyze $\beta(\hat{z}_i) - \beta(z_i)$ and $B^{*\prime}(z_i)(\hat{z}_i - z_i)$ one by one. First, we analyze $\beta(\hat{z}_i) - \beta(z_i)$. Taking a Taylor expansion of $\beta(\hat{z}_i)$ at z_i, we get

$$\beta(\hat{z}_i) = \beta(z_i) + \beta'(z_i)(\hat{z}_i - z_i) + (s.o.) \tag{B.51}$$

where $\beta'(z_i) = d\beta(z_i)/dz_i$.

Thus

$$\begin{aligned} \beta(\hat{z}_i) - \beta(z_i) &= \beta'(z_i)(\hat{z}_i - z_i) + (s.o.) && \text{(B.52)} \\ &= \beta'(z_i)X_i^\top(\hat{\gamma} - \gamma_0) + (s.o.) && \text{(B.53)} \\ &= O_p\left(\frac{1}{\sqrt{n}}\right) && \text{(B.54)} \end{aligned}$$

where the last equality uses the fact that the MAV estimator $\hat{\gamma}$ is a $\sqrt{n}$-consistent estimator, i.e. $\hat{\gamma} - \gamma_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$.

Next, we analyze $B^{*\prime}(z_i)(\hat{z}_i - z_i)$. We note that $B^{*\prime}(z_i) = o_p(1)$ or $O_p(1)$. Thus

$$\begin{aligned} B^{*\prime}(z_i)(\hat{z}_i - z_i) &= B^{*\prime}(z_i)X_i^\top(\hat{\gamma} - \gamma_0) && \text{(B.55)} \\ &= o_p\left(\frac{1}{\sqrt{n}}\right) \text{ or } O_p\left(\frac{1}{\sqrt{n}}\right) && \text{(B.56)} \end{aligned}$$

Summarizing (B.50), (B.54) and (B.56), we have

$$\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i) = O_p\left(\frac{1}{\sqrt{n}}\right) \qquad \text{(B.57)}$$

From lemma (3.1.1) we know $\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i) = O_p\left(h^2 + \frac{1}{\sqrt{nh^3}}\right)$. Then by lemma (B.1) and lemma (B.2) we have

$$C_1 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{LL}(z_i)]^2 M(X_i) = O_p\left(\frac{1}{n}\right) \qquad \text{(B.58)}$$

$$C_2 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)]^2 M(X_i) = O_p\left(\frac{1}{n}\right) \tag{B.59}$$

$$C_3 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)]^2 M(X_i) = O_p\left(h^4 + \frac{1}{nh^3}\right) \tag{B.60}$$

$$C_4 = \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)]M(X_i) = O_p\left(\frac{1}{n}\right) \tag{B.61}$$

$$\begin{aligned} C_5 &= \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{Cubic}(\hat{z}_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)]M(X_i) \\ &= O_p\left(\frac{1}{\sqrt{n}}(h^2 + \frac{1}{\sqrt{nh^3}})\right). \end{aligned} \tag{B.62}$$

$$\begin{aligned} C_6 &= \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)][\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)]M(X_i) \\ &= O_p\left(\frac{1}{\sqrt{n}}(h^2 + \frac{1}{\sqrt{nh^3}})\right). \end{aligned} \tag{B.63}$$

From lemma (3.1.1) we know $h = O_p(n^{-\frac{1}{7}})$. Thus the leading term of $CV_f(h)$ is C_3, i.e. $CV_f(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{Cubic}(\hat{z}_i)\right]^2 M(X_i)$ is asymptotically equivalent to

$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_{LL}(z_i) - \hat{\beta}_{Cubic}(z_i)\right]^2 M(X_i)$.

APPENDIX C

PROOF OF LEMMA 3.2.2

We first decompose $CV(h)$ as follow

$$\begin{aligned} CV(h) &= \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_{LL}(\hat{z}_i)-\beta(z_i)\right]^2 M(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{[\hat{\beta}_{LL}(\hat{z}_i)-\hat{\beta}_{LL}(z_i)]+[\hat{\beta}_{LL}(z_i)-\beta(z_i)]\right\}^2 M(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i)-\hat{\beta}_{LL}(z_i)]^2 M(X_i)+\frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i)-\beta(z_i)]^2 M(X_i) \\ &+ \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i)-\hat{\beta}_{LL}(z_i)][\hat{\beta}_{LL}(z_i)-\beta(z_i)]M(X_i) \\ &= C_1+C_7+C_8 \end{aligned} \tag{C.1}$$

where

$$C_1=\frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i)-\hat{\beta}_{LL}(z_i)]^2 M(X_i) \tag{C.2}$$

$$C_7=\frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i)-\beta(z_i)]^2 M(X_i) \tag{C.3}$$

$$C_8=\frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i)-\hat{\beta}_{LL}(z_i)][\hat{\beta}_{LL}(z_i)-\beta(z_i)]M(X_i) \tag{C.4}$$

It is well established that $\hat{\beta}_{-i,LL}(z_i)-\beta(z_i)=O_p\left(h^2+\frac{1}{\sqrt{nh^3}}\right)=o_p(1)$, see Henderson, Li & Parmeter (2012) and Fan & Gijbels (1996) Theorem 3.1. Using this fact and lemma (B.1) we have

$$C_1 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{LL}(z_i)]^2 M(X_i) = O_p\left(\frac{1}{n}\right) \tag{C.5}$$

$$C_7 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(z_i) - \beta(z_i)]^2 M(X_i) = O_p\left(h^4 + \frac{1}{nh^3}\right) \tag{C.6}$$

$$\begin{aligned} C_8 &= \frac{2}{n}\sum_{i=1}^{n}[\hat{\beta}_{LL}(\hat{z}_i) - \hat{\beta}_{LL}(z_i)][\hat{\beta}_{LL}(z_i) - \beta(z_i)]M(X_i) \\ &= O_p\left(\frac{1}{\sqrt{n}}(h^2 + \frac{1}{\sqrt{nh^3}})\right) \end{aligned} \tag{C.7}$$

By lemma (3.1.1) we know $h = O_p(n^{-\frac{1}{7}})$. Thus the leading term of $CV(h)$ is C_7, i.e.

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_{LL}(\hat{z}_i) - \beta(z_i)\right]^2 M(X_i)$$

is asymptotically equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_{LL}(z_i) - \beta(z_i)\right]^2 M(X_i).$$