Journal of Machine Learning Research 16 (2015) 2617-2641 Submitted 5/14; Revised 11/14; Published 7/15

Derivative Estimation Based on Difference Sequence via Locally Weighted Least Squares Regression

WenWu Wang WENGEWSH@SINA.COM
Lu Lin LINLU@SDU.EDU.CN
Qilu Securities Institute for Financial Studies & School of Mathematics
Shandong University
Jinan, 250100, China

Editor: Francois Caron

Abstract

A new method is proposed for estimating derivatives of a nonparametric regression function. By applying Taylor expansion technique to a derived symmetric difference sequence, we obtain a sequence of approximate linear regression representation in which the derivative is just the intercept term. Using locally weighted least squares, we estimate the derivative in the linear regression model. The estimator has less bias in both valleys and peaks of the true derivative function. For the special case of a domain with equispaced design points, the asymptotic bias and variance are derived; consistency and asymptotic normality are established. In simulations our estimators have less bias and mean square error than its main competitors, especially second order derivative estimator.

Keywords: nonparametric derivative estimation, locally weighted least squares, bias-correction, symmetric difference sequence, Taylor expansion

1. Introduction

In nonparametric regressions, it is often of interest to estimate mean functions. Many estimation methodologies and relevant theoretical properties have been rigorously investigated, see, for example, Fan and Gijbels (1996), Härdle et al. (2004), and Horowitz (2009). Nonparametric derivative estimation has never attracted much attention as one usually gets the derivative estimates as "by-products" from a local polynomial or spline fit, as Newell and Einbeck (2007) mentioned. However, applications of derivative estimation are important and wide-ranging. For example, in the analysis of human growth data, first and second derivatives of the height as a function of time are important parameters (Müller, 1988; Ramsay and Silverman, 2002): the first derivative has the interpretation of speed and the second derivative acceleration. Another field of application is the change point problems, including exploring the structures of curves (Chaudhuri and Marron, 1999; Gijbels and Goderniaux, 2005), detecting the extremum of derivative (Newell et al., 2005), characterizing submicroscopic nanoparticle (Charnigo et al., 2007) and comparing regression curves (Park and Kang, 2008). Other needs arise in nonparametric regressions themselves, for example, in the construction of confidence intervals (Eubank and Speckman, 1993), in the computation of bias and variance, and in the bandwidth selection (Ruppert et al., 1995).

There are three main approaches of nonparametric derivative estimation in the literature: smoothing spline, local polynomial regression (LPR), and difference-based method. As for smoothing spline, the usual way of estimating derivatives is to take derivatives of spline estimate. Stone (1985) showed that spline derivative estimators achieve the optimal L_2 rate of convergence. Zhou and Wolfe (2000) derived asymptotic bias, variance, and established normality properties. Heckman and Ramsay (2000) considered a penalized version. In the case of LPR, a polynomial obtained by Taylor Theorem is fitted locally by kernel regression. Ruppert and Wand (1994) derived the leading bias and variance terms for general multivariate kernel weights using locally weighted least squares theory. Fan and Gijbels (1996) established its asymptotic properties. Delecroix and Rosa (2007) showed its uniform consistency. In the context of difference-based derivative estimation, Müller et al. (1987) and Härdle (1990) proposed a cross-validation technique to estimate the first derivative by combining difference quotients with kernel smoothing. But the variance of the estimator is proportional to n^2 in the case of equidistant design. Charnigo et al. (2011) employed a variance-reducing linear combination of symmetric quotients called empirical derivative, quantified the asymptotic variance and bias, and proposed a generalized C_p criterion for derivative estimation. De Brabanter et al. (2013) derived L_1 and L_2 rates and established consistency of the empirical derivative.

LPR relies on Taylor expansion—a local approximation, and the main term of Taylor series is the mean rather than the derivatives. The convergence rates of the mean estimation and the derivative estimations are different in LPR. When the mean estimator achieves the optimal rate of convergence, the derivative estimators do not (see Table 3 in Appendix I). Empirical derivative can eliminate the main term of the approximation, but it seems that their asymptotic bias and variance properties have not been well studied. Also large biases may exist in valleys and peaks of the derivative function, and boundary problem caused by estimation variance is still an unsolved problem. Motivated by Tong and Wang (2005) and Lin and Li (2008), we propose a new method to estimate derivatives in the interior. By applying Taylor expansion to a derived symmetric difference sequence, we obtain a sequence of approximate linear regression representation in which the derivative is just the intercept term. Then we estimate the derivative in the linear regression model via locally weighted least squares. The asymptotic bias and variance of the new estimator are derived, consistency and asymptotic normality are established. Theoretical properties and simulation results illustrate that our estimators have less bias, especially higher order derivative estimator. In the theory frame of locally weighted least squares regression, the empirical first derivative is our special case: local constant estimator. In addition, one-side locally weighted least squares regression is proposed to solve the boundary problem of first order derivative estimation.

This paper is organized as follows. Section 2 introduces the motivation and methodology of this paper. Section 3 presents theoretical results of the first order derivative estimator, including the asymptotic bias and variance, consistency and asymptotic normality. Further, we describe the behavior at the boundaries of first order derivative estimation and propose a correction method. Section 4 generalizes the idea to higher order derivative estimation. Simulation studies are given in Section 5, and the paper concludes by some discussions in Section 6. All proofs are given in Appendices A-H, respectively.

2. Motivation and Estimation Methodology for the First Order Derivative

In this section, we first show that where the bias and variance of derivative estimation come from, and then propose a new method for the first order derivative estimation.

2.1 Motivation

Consider the following nonparametric regression model

$$Y_i = m(x_i) + \epsilon_i, \quad 1 \le i \le n, \tag{1}$$

where x_i's are equidistantly designed, that is, $x_i = i/n$, Y_i's are random response variables, $m(\cdot)$ is an unknown smooth mean function, ϵ_i's are independent and identically distributed random errors with $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$.

If errors ϵ_i's are not present in (1), the model can be expressed as

$$Y_i = m(x_i), \quad 1 \le i \le n. \tag{2}$$

In this case, the observed Y_i's are actually the true values of the mean function at x_i's. Derivative estimation in model (2) can be viewed as a numerical computation problem. Assume that $m(\cdot)$ is three times continuously differentiable on $[0, 1]$. Then Taylor expansions of $m(x_{i\pm j})$ at x_i are given by

$$\begin{aligned}
m(x_{i+j}) &= m(x_i) + m^{(1)}(x_i)\frac{j}{n} + \frac{m^{(2)}(x_i)}{2!}\frac{j^2}{n^2} + \frac{m^{(3)}(x_i)}{3!}\frac{j^3}{n^3} + o\left(\frac{j^3}{n^3}\right), \\
m(x_{i-j}) &= m(x_i) - m^{(1)}(x_i)\frac{j}{n} + \frac{m^{(2)}(x_i)}{2!}\frac{j^2}{n^2} - \frac{m^{(3)}(x_i)}{3!}\frac{j^3}{n^3} + o\left(\frac{j^3}{n^3}\right).
\end{aligned}$$

In order to eliminate the dominant term $m(x_i)$, we employ a linear combination of $m(x_{i-j})$ and $m(x_{i+j})$ subject to

$$a_{ij} \cdot m(x_{i+j}) + b_{ij} \cdot m(x_{i-j}) = 0 \cdot m(x_i) + 1 \cdot m^{(1)}(x_i) + O\left(\frac{j}{n}\right).$$

It is equivalent to solving the equations

$$\begin{cases} a_{ij} + b_{ij} = 0, \\ (a_{ij} - b_{ij})\dfrac{j}{n} = 1, \end{cases}$$

whose solution is

$$\begin{cases} a_{ij} = \dfrac{n}{2j}, \\ b_{ij} = -\dfrac{n}{2j}. \end{cases}$$

So we obtain

$$m^{(1)}(x_i) = \frac{m(x_{i+j}) - m(x_{i-j})}{2j/n} - \frac{m^{(3)}(x_i)}{6}\frac{j^2}{n^2} + o\left(\frac{j^2}{n^2}\right). \tag{3}$$

As j increases , the bias will also increase. To minimize the bias, set $j = 1$. Then the first order derivative $m^{(1)}(x_i)$ is estimated by

$$\hat{m}^{(1)}(x_i) = \frac{m(x_{i+1}) - m(x_{i-1})}{2/n}.$$

Here the estimation bias is only the remainder term in Taylor expansion.

We now consider the true regression model (1). Symmetric (about i) difference quotients (Charnigo et al., 2011; De Brabanter et al., 2013) are defined as

$$Y_{ij}^{(1)} = \frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}}, \quad 1 \le j \le k, \tag{4}$$

where k is a positive integer. Under model (1), we can decompose $Y_{ij}^{(1)}$ into two parts as

$$Y_{ij}^{(1)} = \frac{m(x_{i+j}) - m(x_{i-j})}{2j/n} + \frac{\epsilon_{i+j} - \epsilon_{i-j}}{2j/n}, \quad 1 \le j \le k. \tag{5}$$

On the right hand side of (5), the first term includes the bias, and the second term contains the information of the variance.

From (3) and (5), we have

$$Y_{ij}^{(1)} = m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6}\frac{j^2}{n^2} + o\left(\frac{j^2}{n^2}\right) + \frac{\epsilon_{i+j} - \epsilon_{i-j}}{2j/n}. \tag{6}$$

Taking expectation on (6), we have

$$\begin{aligned} E[Y_{ij}^{(1)}] &= m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6}\frac{j^2}{n^2} + o\left(\frac{j^2}{n^2}\right) \\ &\doteq m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6}\frac{j^2}{n^2}. \end{aligned}$$

For any fixed $k = o(n)$,

$$E[Y_{ij}^{(1)}] \doteq m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6} d_j, \quad 1 \le j \le k, \tag{7}$$

where $d_j = \frac{j^2}{n^2}$. We treat (7) as a linear regression with d_j and $Y_{ij}^{(1)}$ as the independent variable and dependent variable respectively, and then estimate $m^{(1)}(x_i)$ as the intercept using the locally weighted least squares regression.

2.2 Estimation Methodology

For a fixed x_i, express equation (6) in the following form:

$$Y_{ij}^{(1)} = \beta_{i0} + \beta_{i1} d_{1j} + \delta_{ij}, \quad 1 \le j \le k,$$

where $\beta_{i0} = m^{(1)}(x_i)$, $\beta_{i1} = \frac{m^{(3)}(x_i)}{6}$, $d_{1j} = \frac{j^2}{n^2}$, and $\delta_{ij} = o\left(\frac{j^2}{n^2}\right) + \frac{\epsilon_{i+j} - \epsilon_{i-j}}{2j/n}$ are independent across j. The above expression takes a regression form, in which the independent variable is d_{1j} and the dependent variable $Y_{ij}^{(1)}$, and the error term satisfies

$$E[\delta_{ij}] = o\left(\frac{j^2}{n^2}\right) \doteq 0, \quad Var[\delta_{ij}] = \frac{n^2\sigma^2}{2j^2}.$$

To reduce the variance and combine the information for all j, we use the locally weighted least squares regression (LWLSR) to estimate coefficients as

$$\begin{aligned}\hat{\beta}_i &= \arg\min_{\beta_{i0},\beta_{i1}} \sum_{j=1}^{k} (Y_{ij}^{(1)} - \beta_{i0} - \beta_{i1} d_{1j})^2 w_{ij} \\ &= (D^\top W D)^{-1} D^\top W Y_i^{(1)},\end{aligned}$$

where $w_{ij} = \frac{\sigma^2/2}{Var[\delta_{ij}]} = \frac{j^2}{n^2}$, $\hat{\beta}_i = (\hat{\beta}_{i0}, \hat{\beta}_{i1})^\top$, superscript $\top$ denotes the transpose of a matrix,

$$D = \begin{pmatrix} 1 & 1^2/n^2 \\ 1 & 2^2/n^2 \\ \vdots & \vdots \\ 1 & k^2/n^2 \end{pmatrix}, Y_i^{(1)} = \begin{pmatrix} Y_{i1}^{(1)} \\ Y_{i2}^{(1)} \\ \vdots \\ Y_{ik}^{(1)} \end{pmatrix}, W = \begin{pmatrix} 1^2/n^2 & 0 & \cdots & 0 \\ 0 & 2^2/n^2 & \cdots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & k^2/n^2 \end{pmatrix}.$$

Therefore, the estimator is obtained as

$$\hat{m}^{(1)}(x_i) = \hat{\beta}_{i0} = e_1^\top \hat{\beta}_i, \tag{8}$$

where $e_1 = (1, 0)^\top$.

3. Properties of the First Order Derivative Estimation

In this section, we study asymptotic properties of our first order derivative estimator (8) in interior points, and reveal that empirical first derivative is our special case: local constant estimator. For boundary points, we propose one-side LWLSR to reduce estimation variance.

3.1 Asymptotic Results

The following theorems provide asymptotic results on bias and variance, and establish pointwise consistency and asymptotic normality of the first order derivative estimators.

Theorem 1 (**Uniform Asymptotic Variance**) *Assume that the nonparametric model* (1) *holds with equidistant design and the unknown smooth function* $m(\cdot)$ *is three times continuously differentiable on* $[0, 1]$. *Furthermore, assume that the third order derivative* $m^{(3)}(\cdot)$ *is finite on* $[0, 1]$. *Then the variance of the first order derivative estimator in* (8) *is*

$$Var[\hat{m}^{(1)}(x_i)] = \frac{75\sigma^2}{8}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right)$$

uniformly for $k + 1 \le i \le n - k$.

Theorem 1 shows that the variance of the derivative estimator is constant as x changes, while the following theorem shows that the bias changes with x.

Theorem 2 (**Pointwise Asymptotic Bias**) *Assume that the nonparametric model* (1) *holds with equidistant design and the unknown smooth function* $m(\cdot)$ *is five times continuously differentiable on* $[0,1]$. *Furthermore, assume that the fifth order derivative* $m^{(5)}(\cdot)$ *is finite on* $[0,1]$. *Then the bias of the first order derivative estimator in* (8) *is*

$$Bias[\hat{m}^{(1)}(x_i)] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right)$$

for $k+1 \leq i \leq n-k$.

Using Theorems 1 and 2, we have that if $nk^{-3/2} \to 0$ and $n^{-1}k \to 0$, then our estimator has the consistency property

$$\hat{m}^{(1)}(x_i) \xrightarrow{P} m^{(1)}(x_i).$$

Furthermore, we establish asymptotic normality in the following theorem.

Theorem 3 (**Asymptotic Normality**) *Under the assumptions of Theorem 2, if* $k \to \infty$ *as* $n \to \infty$ *such that* $nk^{-3/2} \to 0$ *and* $n^{-1}k \to 0$, *then*

$$\frac{k^{3/2}}{n}\left(\hat{m}^{(1)}(x_i) - m^{(1)}(x_i) + \frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4}\right) \xrightarrow{d} N\left(0, \frac{75\sigma^2}{8}\right)$$

for $k+1 \leq i \leq n-k$. *Further, if* $k \to \infty$ *as* $n \to \infty$ *such that* $nk^{-3/2} \to 0$ *and* $n^{-1}k^{11/10} \to 0$, *then*

$$\frac{k^{3/2}}{n}\left(\hat{m}^{(1)}(x_i) - m^{(1)}(x_i)\right) \xrightarrow{d} N\left(0, \frac{75\sigma^2}{8}\right)$$

for $k+1 \leq i \leq n-k$.

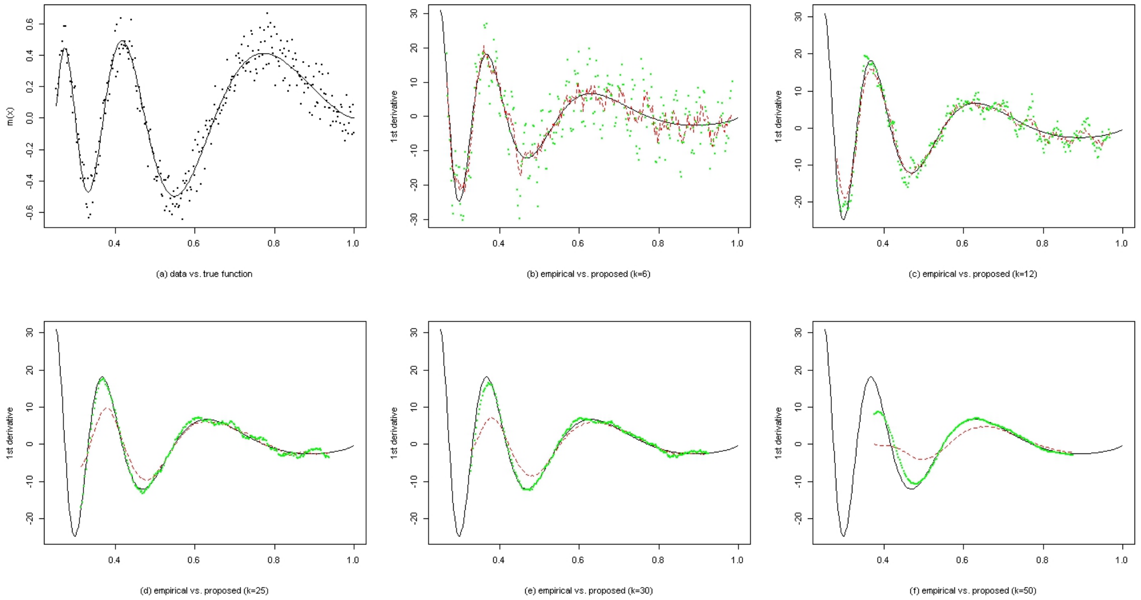
Theorem 3 shows that with suitable choice of k our first order derivative estimator is asymptotically normally distributed, even asymptotically unbiased. Using the asymptotic normality property, we can construct confidence intervals and confidence bands. From the above theorems, the following corollary follows naturally.

Corollary 4 *Under the assumptions of Theorem 2, the optimal choice of* k *that minimizes the asymptotic mean square error of the first order derivative estimator in* (8) *is*

$$k_{opt} \doteq 3.48\left(\frac{\sigma^2}{(m^{(5)}(x_i))^2}\right)^{1/11} n^{10/11}.$$

With the optimal choice of k, *the asymptotic mean square error of the first order derivative estimator in* (8) *can be expressed as*

$$AMSE[\hat{m}^{(1)}(x_i)] \doteq 0.31\left(\sigma^{16}(m^{(5)}(x_i))^6\right)^{1/11} n^{-8/11}.$$



(a) data vs. true function

(b) empirical vs. proposed (k=6)

(c) empirical vs. proposed (k=12)

(d) empirical vs. proposed (k=25)

(e) empirical vs. proposed (k=30)

(f) empirical vs. proposed (k=50)

Figure 1: (a) Simulated data set of size 300 from model (1) with equidistant $x_i \in [0.25, 1]$, $m(x) = \sqrt{x(1-x)}\sin((2.1\pi)/(x+0.05))$, $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$, and the true mean function (bold line). (b)-(f) The proposed first order derivative estimators (green dots) and the empirical first derivatives (red dashed lines) for $k \in \{6, 12, 25, 30, 50\}$. As a reference, the true first order derivative is also plotted (bold line).

Now we briefly examine the finite sample behavior of our estimator and compare it with the empirical first derivative given by Charnigo et al. (2011) and De Brabanter et al. (2013). Their estimator has the following form:

$$Y_i^{[1]} = \sum_{j=1}^{k_1} w_{ij} Y_{ij}^{(1)}, \quad k_1 + 1 \leq i \leq n - k_1, \tag{9}$$

where k_1 is a positive integer, $w_{ij} = \frac{j^2/n^2}{\sum_{j=1}^{k_1} j^2/n^2}$, and $Y_{ij}^{(1)}$ is defined in (4).

Figure 1 displays our proposed first order derivative estimators (8) and empirical first derivatives (9) with $k_1 = k \in \{6, 12, 25, 30, 50\}$, for a data set of size 300 generated from model (1) with $x_i \in [0.25, 1]$, $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$, and $m(x) = \sqrt{x(1-x)}\sin((2.1\pi)/(x+0.05))$. This $m(x)$ is borrowed from De Brabanter et al. (2013). When k is small (see Figure 1 (b) and (c)), the proposed estimators are noise corrupted versions of the true first order derivatives, while the performance of the empirical derivatives is better except that there are large biases near local peaks and valleys of the true derivative function. As k becomes bigger (see Figure 1 (d)- (f)), our estimators have much less biases than empirical derivative estimators near local peaks and valleys of the true derivative. The balance between the estimation bias and variance is clear even visually. Furthermore, if we combine the left part of Figure 1 (d), the middle part of (e) and the right part of (f), more accurate derivative estimators are obtained.

Actually, empirical first derivative and our estimator have a close relationship. Express equation (6) in simple linear regression form

$$Y_{ij}^{(1)} = \beta_{i0} + \eta_{ij}, \quad 1 \leq j \leq k,$$

where $\beta_{i0} = m^{(1)}(x_i)$, $\eta_{ij} = O\left((\frac{j}{n})^2\right) + \frac{\epsilon_{i+j} - \epsilon_{i-j}}{2j/n}$ with

$$E[\eta_{ij}] \doteq 0, \quad Var[\eta_{ij}] = \frac{n^2\sigma^2}{2j^2}.$$

This is called the local constant-truncated estimator. By the LWLSR, we get

$$\hat{m}^{(1)}(x_i) = Y_i^{[1]},$$

which is exactly the empirical first derivative. On three times continuous differentiability, we have the following bias and variance

$$Bias[Y_i^{[1]}] = \frac{m^{(3)}(x_i)}{10}\frac{k^2}{n^2}, \quad Var[Y_i^{[1]}] = \frac{3\sigma^2}{2}\frac{n^2}{k^3}.$$

For empirical first derivative and our estimator, symmetric difference sequence eliminates the even-order terms in Taylor expansion of mean function. This is an important advantage, i.e., if the mean function is two times continuously differentiable, then the second-order term is eliminated so that the bias is smaller than the second-order term. In De Brabanter et al. (2013), the bias is $O(k/n)$ which is obtained via a inequality (See Appendix B, De Brabanter et al. 2013). In fact, the bias should be

$$Bias[Y_i^{[1]}] < O(k/n) \quad or \quad Bias[Y_i^{[1]}] = o(k/n),$$

which does not have exact and explicit expression on two times continuous differentiability. In order to obtain explicit expression, we make the stronger smoothing condition—three times continuous differentiability.

In addition, the smoothing assumptions on bias and variance are different in Theorem 1 and Theorem 2. For empirical first derivative and ours, the variance term only needs one time continuous differentiability; whereas the bias term needs three times and five times respectively. From the viewpoint of Taylor expansion, it seems we pay a serious price. However, in practical applications bias-correction is needed especially in the cases of immense oscillation of mean function. From the viewpoint of *Weierstrass approximation theorem,* even if a continuous function is nondifferentiable we still can correct the bias.

3.2 Behavior at the Boundaries

Recall that for the boundary region ($2 \leq i \leq k$ and $n - k + 1 \leq i \leq n - 1$) the weights in empirical first derivative (9) are slightly modified by normalizing the weight sum. Whereas our estimator can be obtained directly from the LWLSR without any modification, the only difference is that the smoothing parameter is $i-1$ instead of k.

For the boundary ($3 \leq i \leq k$), the bias and variance for our estimator are

$$Bias[\hat{m}^{(1)}(x_i)] = -\frac{m^{(5)}(x_i)}{504}\frac{(i-1)^4}{n^4}, \quad Var[\hat{m}^{(1)}(x_i)] = \frac{75\sigma^2}{8}\frac{n^2}{(i-1)^3}. \tag{10}$$

Hence, the variance will be the largest for $i=3$ and decrease for growing i till $i=k$, whereas the bias will be smallest for $i=3$ and increase for growing i till $i=k$. A similar analysis for $n-k+1 \leq i \leq n-2$ shows the same results.

For the modified estimator (De Brabanter et al., 2013), the bias and variance in the theory frame of the LWLSR are

$$Bias[\hat{m}^{(1)}(x_i)] = \frac{m^{(3)}(x_i)}{10}\frac{(i-1)^2}{n^2}, \quad Var[\hat{m}^{(1)}(x_i)] = \frac{3\sigma^2}{2}\frac{n^2}{(i-1)^3},$$

Which have the analogue change trend like (10) above. Although our estimator has less bias ($O(1/n^4)$) than empirical first derivative ($O(1/n^2)$), the variances both are big enough ($O(n^2)$). So the two estimators are inaccurate and the boundary problem still exists.

In order to reduce the variance, we propose the one-side locally weighted least squares regression method which consists of two cases: left-side locally weighted least squares regression (LSLWLSR) and right-side locally weighted least squares regression (RSLWLSR). These estimation methods can be used for the boundary: LSLWLSR is for $n-k+1 \leq i \leq n$ and RSLWLSR is for $1 \leq i \leq k$. On two times continuous differentiability, the estimation bias is $O(k/n)$ and variance is $O(n^2/k^3)$.

Assume that $m(\cdot)$ is two times continuously differentiable on $[0,1]$. For $1 \leq i \leq n-k$, define right-side lag-j first-order difference sequence

$$Y_{ij}^{<1>} = \frac{Y_{i+j} - Y_i}{x_{i+j} - x_i}, \quad 1 \leq j \leq k. \tag{11}$$

Decompose $Y_{ij}^{<1>}$ into two parts and simplify from (11) such as

$$\begin{aligned} Y_{ij}^{<1>} &= \frac{m(x_{i+j}) - m(x_i)}{j/n} + \frac{\epsilon_{i+j} - \epsilon_i}{j/n} \\ &= m^{(1)}(x_i) + \frac{m^{(2)}(x_i)}{2!}\frac{j^1}{n^1} + o\left(\frac{j^1}{n^1}\right) + \frac{\epsilon_{i+j} - \epsilon_i}{j/n}. \end{aligned} \tag{12}$$

For some fixed i, ϵ_i is constant as j increases. Thus we express equation (12) in the following form:

$$Y_{ij}^{<1>} = \beta_{i0} + \beta_{i1}d_{1j} + \delta_{ij}, \quad 1 \leq j \leq k,$$

where $\beta_{i0} = m^{(1)}(x_i)$, $\beta_{i1} = -\epsilon_i$, $d_{1j} = \frac{n}{j}$, and $\delta_{ij} = \frac{m^{(2)}(x_i)}{2!}\frac{j^1}{n^1} + o\left(\frac{j^1}{n^1}\right) + \frac{\epsilon_{i+j}}{j/n}$ are independent across j with

$$E[\delta_{ij}|\epsilon_i] = \frac{m^{(2)}(x_i)}{2!}\frac{j^1}{n^1} + o\left(\frac{j^1}{n^1}\right) \doteq 0, \quad Var[\delta_{ij}|\epsilon_i] = \frac{n^2\sigma^2}{j^2}.$$

So we use the LWLSR to estimate regression coefficients as

$$\begin{aligned}\hat{\beta}_i &= \arg\min_{\beta_{i0},\beta_{i1}} \sum_{j=1}^{k} (Y_{ij}^{(1)} - \beta_{i0} - \beta_{i1} d_{1j})^2 w_{ij} \\ &= (D^\top W D)^{-1} D^\top W Y_i^{<1>},\end{aligned}$$

where $w_{ij} = \frac{\sigma^2}{Var[\delta_{ij}]} = \frac{j^2}{n^2}$, $\hat{\beta}_i = (\hat{\beta}_{i0}, \hat{\beta}_{i1})^\top$,

$$D = \begin{pmatrix} 1 & n^1/1^1 \\ 1 & n^1/2^1 \\ \vdots & \vdots \\ 1 & n^1/k^1 \end{pmatrix}, Y_i^{<1>} = \begin{pmatrix} Y_{i1}^{<1>} \\ Y_{i2}^{<1>} \\ \vdots \\ Y_{ik}^{<1>} \end{pmatrix}, W = \begin{pmatrix} 1^2/n^2 & 0 & \cdots & 0 \\ 0 & 2^2/n^2 & \cdots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & k^2/n^2 \end{pmatrix}.$$

Therefore, the estimator is obtained as

$$\hat{m}^{(1)}(x_i) = \hat{\beta}_{i0} = e_1^\top \hat{\beta}_i, \tag{13}$$

where $e_1 = (1, 0)^\top$.

Following Theorem 1-4 above, we have the similar theorems for the right-side first-order derivative estimator in (13). Here we only give the asymptotic bias and variance as follows.

Theorem 5 *Assume that the nonparametric model (1) holds with equidistant design and the unknown smooth function $m(\cdot)$ is two times continuously differentiable on $[0,1]$. Furthermore, assume that the second order derivative $m^{(2)}(\cdot)$ is finite on $[0,1]$. Then the bias and variance of the right-side first-order derivative estimator in (13) are*

$$\begin{aligned} Bias[\hat{m}^{(1)}(x_i)|\epsilon_i] &= \frac{m^{(2)}(x_i)}{2}\frac{k^1}{n^1} + o\left(\frac{k^1}{n^1}\right) \\ Var[\hat{m}^{(1)}(x_i)|\epsilon_i] &= 12\sigma^2 \frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right) \end{aligned}$$

correspondingly for $1 \leq i \leq n-k$.

From Theorem 5 above, we can see that the variance and bias for the right-side first-order derivative estimator in (13) is $O(n^2/k^3)$ and $O(k/n)$, which is the same rate as De Brabanter et al. (2013) deduced on two times continuous differentiability. For further bias-correction, high-order Taylor expansion may be needed. A similar analysis for left-side lag-j first-order difference sequence obtains the same results.

3.3 The Choice of k

From the tradeoff between bias and variance, we have two methods for the choice of k: adaptive method and uniform method. The adaptive k based on asymptotic mean square error is

$$k_a = 3.48 \left(\frac{\sigma^2}{(m^{(5)}(x_i))^2}\right)^{1/11} n^{10/11}.$$

To choose k globally, we consider the mean averaged square error (MASE) criterion

$$\begin{aligned}
MASE &= \frac{1}{n-2k}\sum_{i=k+1}^{n-k} MSE(\hat{m}^{(1)}(x_i)) \\
&= \frac{1}{n-2k}\sum_{i=k+1}^{n-k}(\frac{75\sigma^2}{8}\frac{n^2}{k^3}+\frac{(m^{(5)}(x_i))^2}{504^2}\frac{k^8}{n^8}) \\
&= \frac{75\sigma^2}{8}\frac{n^2}{k^3}+\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\frac{(m^{(5)}(x_i))^2}{504^2}\frac{k^8}{n^8} \\
&\to \frac{75\sigma^2}{8}\frac{n^2}{k^3}+\frac{L_5}{504^2}\frac{k^8}{n^8},
\end{aligned}$$

where $L_5 = \int_0^1 (m^{(5)}(x))^2 \mathrm{d}x$. Minimizing the $MASE$ with respect to k, the uniform k is

$$k_u = 3.48(\frac{\sigma^2}{L_5})^{1/11} n^{10/11}.$$

Since the k_a and k_u are unknown in practice, a rule of thumb estimator may be preferable. The error variance σ^2 can be estimated by Hall et al. (1990), the fifth order derivative $m^{(5)}(x_i)$ can be estimated by local polynomial regression (R-package: locpol), and L_5 is estimated by Seifert et al. (1993).

However, questions still remain. First, $k = O(n^{10/11})$, which requires n to be large enough to ensure $k < n$; Second, 'the higher the derivative, the wilder the behavior' (Ramsay, 1998), thus the estimations of $m^{(5)}(x_i)$ and L_5 are inaccurate. The most important is that when the bias is very small or large we can't balance the bias and variance via only increasing or decreasing the value of k. From the expression of adaptive k, uniform k and simulations, we put forward the following considerations.

- On the whole, k should satisfy $k < n/2$ or else the needed data size $2k$ goes over the total size n so that we can't estimate any derivative. In addition, we can't leave more boundary points than interior points, so k needs to satisfy the condition $k < n/4$.

- The choice of k relies on Taylor expansion which is a local concept. There exists some maximum value of k suitable for a fixed mean function, denoted by k_{max}. However, adaptive and uniform k is determined by many factor: variance, sample size, frequency and amplitude of mean function. Thus it is possible to obtain too big k in the cases of large variance, and now cross validation could be an alternative. As frequency and amplitude increase, the uniform and adaptive k decrease. *This is the reason why our estimator adopting different k for different oscillation has better performance in Figure 1.* In addition, as the order of Taylor expansion increases, the k_{max} becomes large. So our estimator needs a larger k than empirical derivative.

- When the third-order and fifth-order derivatives are close to zero, the values of k_a and k_u are too big even $k > n/2$. Thus we can't balance bias and variance via increasing the value of k when bias is very small. Meanwhile we can't balance bias and variance via decreasing the value of k when bias is too big. It is better to correct bias by higher-order Taylor expansion.

4. Higher Order Derivative Estimations

In this section, we generalize the idea of the first order derivative estimation to higher order. Different difference sequences are adopted for first and second order derivative estimation.

4.1 Second Order Derivative Estimation

As for the second order derivative estimation, we can show by a similar technique as in (3) that

$$m^{(2)}(x_i) = \frac{m(x_{i-j}) - 2m(x_i) + m(x_{i+j})}{j^2/n^2} - \frac{m^{(4)}(x_i)}{12}\frac{j^2}{n^2} + o\left(\frac{j^2}{n^2}\right).$$

Define

$$Y_{ij}^{(2)} = \frac{Y_{i-j} - 2Y_i + Y_{i+j}}{j^2/n^2}. \tag{14}$$

Just as in equation (5), decompose (14) into two parts as

$$Y_{ij}^{(2)} = \frac{m(x_{i-j}) - 2m(x_i) + m(x_{i+j})}{j^2/n^2} + \frac{\epsilon_{i-j} - 2\epsilon_i + \epsilon_{i+j}}{j^2/n^2}, \quad 1 \leq j \leq k.$$

Note that i is fixed as j changes. Thus the conditional expectation of $Y_{ij}^{(2)}$ given ϵ_i is

$$\begin{aligned} E[Y_{ij}^{(2)}|\epsilon_i] &= \frac{m(x_{i-j}) - 2m(x_i) + m(x_{i+j})}{j^2/n^2} + (-2\epsilon_i)\frac{n^2}{j^2} \\ &\doteq m^{(2)}(x_i) + \frac{m^{(4)}(x_i)}{12}\frac{j^2}{n^2} + (-2\epsilon_i)\frac{n^2}{j^2}. \end{aligned}$$

Therefore, the new regression model is given by

$$Y_{ij}^{(2)} = \beta_{i0} + \beta_{i1}d_{1j} + \beta_{i2}d_{2j} + \delta_{ij}, \quad 1 \leq j \leq k,$$

where the regression coefficient vector $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})^\top = (m^{(2)}(x_i), \frac{m^{(4)}(x_i)}{12}, -2\epsilon_i)^\top$, covariates $d_{1j} = \frac{j^2}{n^2}$ and $d_{2j} = \frac{n^2}{j^2}$, and the error term $\delta_{ij} = \frac{\epsilon_{i+j}+\epsilon_{i-j}}{j^2/n^2} + o\left(\frac{j^2}{n^2}\right)$, with

$$E[\delta_{ij}|\epsilon_i] \doteq 0, \quad Var[\delta_{ij}|\epsilon_i] = \frac{2\sigma^2 n^4}{j^4}.$$

Now the locally weighted least squares estimator of β_i can be expressed as

$$\hat{\beta}_i = (D^\top W D)^{-1} D^\top W Y_i^{(2)},$$

where

$$D = \begin{pmatrix} 1 & 1^2/n^2 & n^2/1^2 \\ 1 & 2^2/n^2 & n^2/2^2 \\ \vdots & \vdots & \vdots \\ 1 & k^2/n^2 & n^2/k^2 \end{pmatrix}, Y_i^{(2)} = \begin{pmatrix} Y_{i1}^{(2)} \\ Y_{i2}^{(2)} \\ \vdots \\ Y_{ik}^{(2)} \end{pmatrix}, W = \begin{pmatrix} 1^4/n^4 & 0 & \cdots & 0 \\ 0 & 2^4/n^4 & \cdots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & k^4/n^4 \end{pmatrix}.$$

Therefore,

$$\hat{m}^{(2)}(x_i) = \hat{\beta}_{i0} = e_1^\top \hat{\beta}_i, \tag{15}$$

where $e_1 = (1, 0, 0)^\top$.

The following three theorems provide asymptotic results on bias, variance and mean square error, and establish pointwise consistency and asymptotic normality of the second order derivative estimator.

Theorem 6 *Assume that the nonparametric model* (1) *holds with equidistant design and the unknown smooth function $m(\cdot)$ is six times continuously differentiable on* $[0, 1]$. *Furthermore, assume that the sixth order derivative $m^{(6)}(\cdot)$ is finite on* $[0, 1]$. *Then the variance of the second order derivative estimator in* (15) *is*

$$Var[\hat{m}^{(2)}(x_i)|\epsilon_i] = \frac{2205\sigma^2}{8}\frac{n^4}{k^5} + o(\frac{n^4}{k^5})$$

uniformly for $k + 1 \leq i \leq n - k$, and the bias is

$$Bias[\hat{m}^{(2)}(x_i)|\epsilon_i] = -\frac{m^{(6)}(x_i)}{792}\frac{k^4}{n^4} + o(\frac{k^4}{n^4})$$

for $k + 1 \leq i \leq n - k$.

From Theorem 6, we can see that if $nk^{-5/4} \to 0$ and $n^{-1}k \to 0$, then our estimator is consistent

$$\hat{m}^{(2)}(x_i) \xrightarrow{P} m^{(2)}(x_i).$$

Moreover, we establish asymptotic normality, derive the asymptotic mean square error and the optimal k value.

Corollary 7 *Under the assumptions of Theorem 6, if $k \to \infty$ as $n \to \infty$ such that $nk^{-5/4} \to 0$ and $n^{-1}k \to 0$, then*

$$\frac{k^{5/2}}{n^2}\left(\hat{m}^{(2)}(x_i) - m^{(2)}(x_i) + \frac{m^{(6)}(x_i)}{792}\frac{k^4}{n^4}\right) \xrightarrow{d} N\left(0, \frac{2205\sigma^2}{8}\right).$$

Moreover, if $nk^{-5/4} \to 0$ and $n^{-1}k^{13/12} \to 0$, then

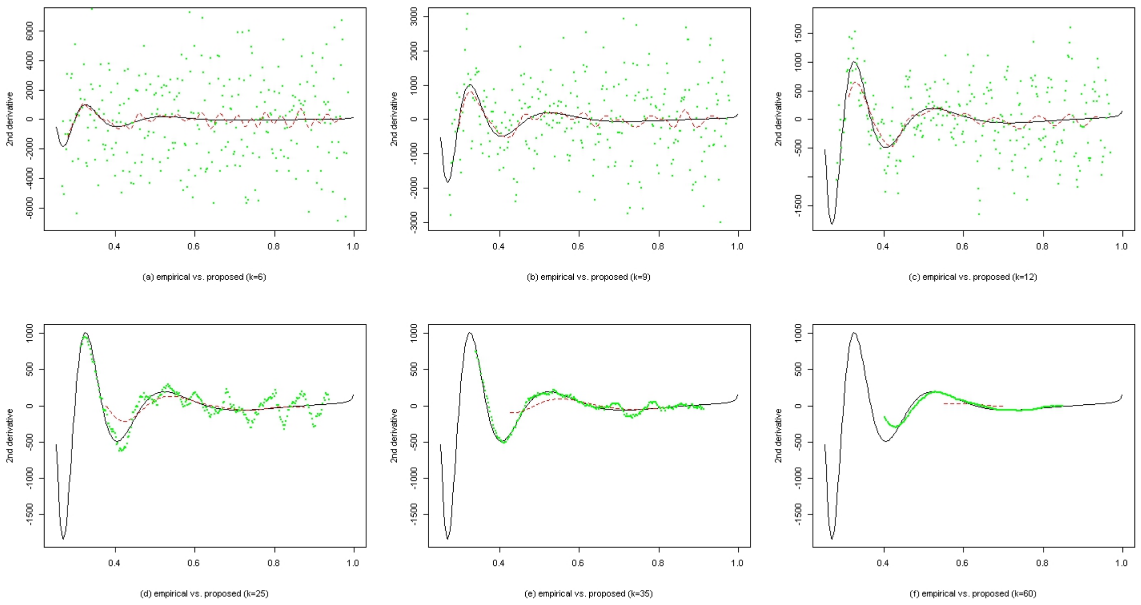
$$\frac{k^{5/2}}{n^2}\left(\hat{m}^{(2)}(x_i) - m^{(2)}(x_i)\right) \xrightarrow{d} N\left(0, \frac{2205\sigma^2}{8}\right).$$

Corollary 8 *Under the assumptions of Theorem 6, the optimal k value that minimizes the asymptotic mean square error of the second order derivative estimator in* (15) *is*

$$k_{opt} \doteq 4.15\left(\frac{\sigma^2}{(m^{(6)}(x_i))^2}\right)^{1/13} n^{12/13}.$$

With the optimal choice of k, the asymptotic mean square error of the second order derivative estimator in (15) *can be expressed as*

$$AMSE[\hat{m}^{(1)}(x_i)] \doteq 0.36\left(\sigma^{16}(m^{(6)}(x_i))^{10}\right)^{1/13} n^{-8/13}.$$



(a) empirical vs. proposed (k=6)

(b) empirical vs. proposed (k=9)

(c) empirical vs. proposed (k=12)

(d) empirical vs. proposed (k=25)

(e) empirical vs. proposed (k=35)

(f) empirical vs. proposed (k=60)

Figure 2: (a)-(f) The proposed second order derivative estimators (green points) and the empirical second derivatives (red dashed line) for $k \in \{6, 9, 12, 25, 35, 60\}$ based on the simulated data set from Figure 1. As a reference, the true second order derivative curve is also plotted (bold line).

Here we also use a simple simulation to examine the finite sample behavior of the new estimator and compare it with the empirical second derivative given by Charnigo et al. (2011) and De Brabanter et al. (2013). Their estimator has the following form:

$$Y_i^{[2]} = \sum_{j=1}^{k_2} w_{ij} Y_{ij}^{(2)}, \quad k_1 + k_2 + 1 \leq i \leq n - k_1 - k_2, \tag{16}$$

where $w_{ij} = \frac{j/n}{\sum_{j=1}^{k_2} j/n}$, $Y_{ij}^{(2)} = (Y_{i+j}^{(1)} - Y_{i-j}^{(1)})/(2j/n)$, k_1 is the same as in (9), and k_2 is a positive integer. Figure 2 displays our second order derivative estimators and empirical second derivatives (16) at interior point for the data from Figure 1, where $k_1 = k_2 = k \in \{6, 9, 12, 25, 35, 60\}$. The performance of the our second derivative estimator and empirical second derivative is parallel to the first derivative's case. Note that the k values used here are larger than the counterparts in the first order derivative estimation.

4.2 Higher Order Derivative Estimation

We generalize the method aforementioned to higher order derivatives $m^{(l)}(x_i)$ $(l > 2)$. The method includes two main steps: the first step is to construct a sequence of symmetric difference quotients in which the derivative is the intercept of the linear regression derived by Taylor expansion, and the second step is to estimate the derivative using the LWLSR.

The construction of a difference sequence is particularly important because it determines the estimation accuracy.

When l is odd, set $d = \frac{l+1}{2}$. We linearly combine $m(x_{i\pm j})$'s subject to

$$\sum_{h=1}^{d}[a_{i,jd+h}m(x_{i+jd+h}) + a_{i,-(jd+h)}m(x_{i-(jd+h)})] = m^{(l)}(x_i) + O\left(\frac{j}{n}\right), \quad 0 \leq j \leq k,$$

where k is a positive integer. We can derive $2d$ equations through Taylor expansion and solve out the $2d$ unknown parameters. Define

$$Y_{ij}^{(l)} = \sum_{h=1}^{d}[a_{i,jd+h}Y_{i+jd+h} + a_{i,-(jd+h)}Y_{i-(jd+h)}].$$

and consider the linear regression

$$Y_{ij}^{(l)} = m^{(l)}(x_i) + \delta_{ij}, \quad 1 \leq j \leq k,$$

where $\delta_{ij} = \sum_{h=1}^{d}[a_{i,jd+h}\epsilon_{i+jd+h} + a_{i,-(jd+h)}\epsilon_{i-(jd+h)}] + O(\frac{j}{n})$.

When l is even, set $d = \frac{l}{2}$. We linearly combine $m(x_{i\pm j})$'s subject to

$$b_{i,j}m(x_i)+\sum_{h=1}^{d}[a_{i,jd+h}m(x_{i+jd+h})+a_{i,-(jd+h)}m(x_{i-(jd+h)})] = m^{(l)}(x_i)+O\left(\frac{j}{n}\right), \quad 0 \leq j \leq k,$$

where k is a positive integer. We can derive $2d+1$ equations through Taylor expansion and solve out the $2d+1$ unknown parameters. Define

$$Y_{ij}^{(l)} = b_{i,j}m(x_i) + \sum_{h=1}^{d}[a_{i,jd+h}Y_{i+jd+h} + a_{i,-(jd+h)}Y_{i-(jd+h)}].$$

and consider the linear regression

$$Y_{ij}^{(l)} = m^{(l)}(x_i) + b_{i,j}\epsilon_i + \delta_{ij}, \quad 1 \leq j \leq k,$$

where $\delta_{ij} = \sum_{h=1}^{d}[a_{i,jd+h}\epsilon_{i+jd+h} + a_{i,-(jd+h)}\epsilon_{i-(jd+h)}] + O(\frac{j}{n})$.

If k is large enough, it is better to keep the j^2/n^2 term like (7) to reduce the estimation bias. With the regression models defined above, we can obtain the higher order derivative estimators and deduce their asymptotic results by similar arguments as in the previous subsection; the details are omitted here.

5. Simulations

In addition to the simple simulations in the previous sections, we conduct more simulation studies in this section to further evaluate the finite-sample performances of the proposed method and compare it with two other well-known methods. To get more comprehensive comparisons, we use estimation curves and mean absolute errors to assess the performances of different methodologies.

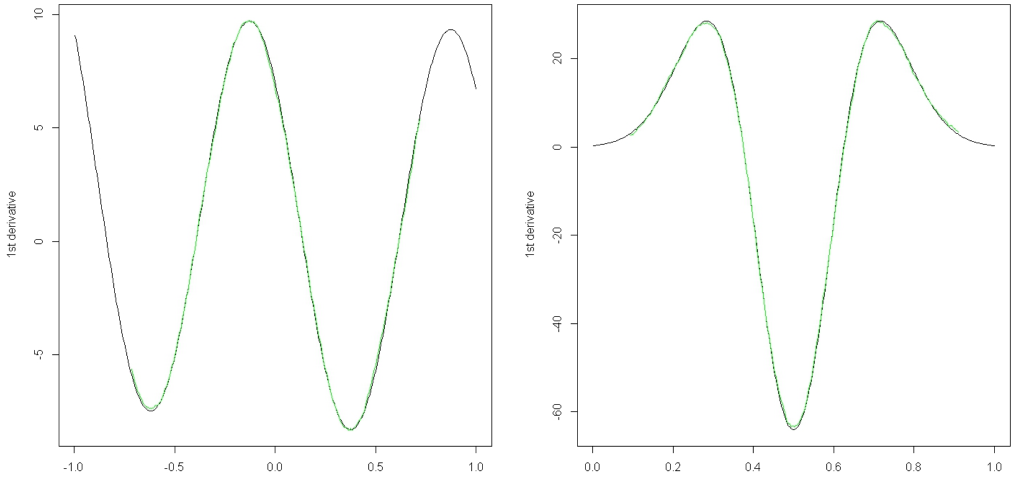
5.1 Finite Sample Results of the First Order Derivative Estimation

We first consider the following two regression functions

$$m(x) = \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x), \quad x \in [-1, 1], \tag{17}$$

and

$$m(x) = 32e^{-8(1-2x)^2}(1 - 2x), \quad x \in [0, 1]. \tag{18}$$



(a) proposed (k=70) vs. true function

(b) proposed (k=45) vs. true function

Figure 3: (a) The true first order derivative function (bold line) and our first order derivative estimations (green dashed line). Simulated data set of size 500 from model (1) with equispaced $x_i \in [-1, 1]$, $m(x) = \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$, and $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$. (b) The true first order derivative function (bold line) and our first order derivative estimations (green dashed line). Simulated data set of size 500 from model (1) with equispaced $x_i \in [0, 1]$, $m(x) = 32e^{-8(1-2x)^2}(1 - 2x)$, and $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$.

These two functions were considered by Hall (2010) and De Brabanter et al. (2013), respectively. The data sets are of size $n = 500$ and generated from model (1) with $\epsilon \sim N(0, \sigma^2)$ for $\sigma = 0.1$. Figure 3 presents the first order derivative estimations of regression functions (17) and (18). It shows that our estimation curves of the first order derivative fit the true curves accurately, although a comparison with the other estimators is not given in the figure.

We now evaluate our estimator with empirical first derivative. Since the oscillation of the periodic function depends on frequency and amplitude, in our simulations we choose the mean function

$$m(x) = A \sin(2\pi f x), \quad x \in [0, 1],$$

			Ours	Empirical	Ours	Empirical	Ours	Empirical
A	f	σ	n=50	n=50	n=200	n=200	n=1000	n=1000
1	1	0.1	0.28(0.09)	0.36(0.08)	0.14(0.04)	0.24(0.04)	0.07(0.02)	0.16(0.03)
		0.5	1.38(0.45)	1.01(0.28)	0.69(0.23)	0.61(0.16)	0.30(0.10)	0.38(0.07)
		2	5.54(1.87)	2.35(0.90)	2.73(0.89)	1.39(0.48)	1.18(0.37)	0.85(0.24)
	2	0.1	0.58(0.15)	1.00(0.13)	0.34(0.07)	0.61(0.07)	0.19(0.03)	0.39(0.04)
		0.5	1.76(0.57)	2.33(0.52)	1.08(0.31)	1.52(0.28)	0.60(0.17)	0.97(0.16)
		2	5.59(1.88)	4.91(1.60)	2.96(1.05)	3.36(0.95)	1.63(0.52)	2.11(0.48)
10	1	0.1	0.41(0.09)	0.98(0.12)	0.24(0.05)	0.67(0.07)	0.13(0.03)	0.42(0.04)
		0.5	1.45(0.47)	2.46(0.44)	0.80(0.22)	1.65(0.25)	0.42(0.10)	1.05(0.14)
		2	5.52(1.76)	5.27(1.24)	2.71(0.89)	3.55(0.76)	1.28(0.35)	2.31(0.38)
	2	0.1	1.15(0.17)	2.90(0.20)	0.64(0.09)	1.66(0.12)	0.35(0.05)	1.06(0.06)
		0.5	3.72(0.79)	6.44(0.77)	2.06(0.39)	4.18(0.42)	1.16(0.19)	2.65(0.24)
		2	9.38(2.78)	13.3(2.40)	5.66(1.38)	9.14(1.35)	3.17(0.72)	5.74(0.65)

Table 1: Adjusted Mean Absolute Error for the first order derivative estimation.

with design points $x_i = i/n$, and the errors are independent and identically normal distribution with zero mean and variance σ^2. We consider three sample sites $n = 50, 200, 1000$, corresponding to small, moderate, and large sample sizes, three standard deviations $\sigma = 0.1, 0.5, 2$, two frequencies $f = 1, 2$, and two amplitudes $A = 1, 10$. The number of repetitions is set as 1000. We consider two criterion: adjusted mean absolute error (AMAE) and mean averaged square error, and find that they have similar performance. For the sake of simplicity and robustness, we choose the AMAE as a measure of comparison. It is defined as

$$AMAE(k) = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} |\hat{m}'(x_i) - m'(x_i)|,$$

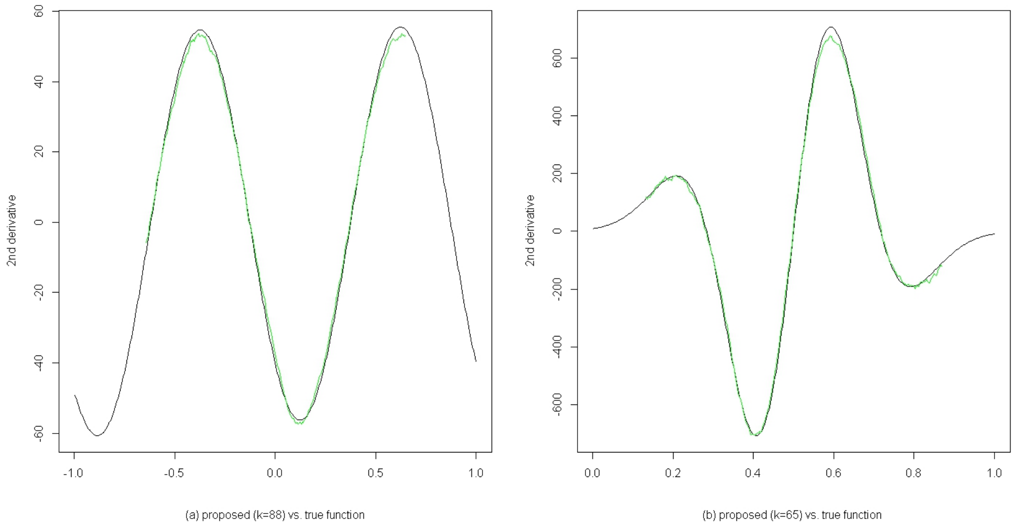
here the boundary effects are excluded. According to the condition $k < n/4$ and the AMAE criterion, we choose k as follows

$$\hat{k} = \min\{\arg\min_{k} AMAE(k), \frac{n}{4}\}.$$

Table 1 reports the simulation results. The numbers outside and inside the brackets are the mean and standard deviation of the AMAE. It indicates our estimator performs better than empirical first derivative in most cases *except that the adoptive k is much less the theoretically uniform k.*

5.2 Finite Sample Results of the Second Order Derivative Estimation

Consider the same functions as in Subsection 5.1. Figure 4 presents the estimation curves and the true curves of the second order derivatives of (17) and (18). It shows that our estimators track the true curves closely.



(a) proposed (k=88) vs. true function

(b) proposed (k=65) vs. true function

Figure 4: (a)-(b) The true second order derivative function (bold line) and proposed second order derivative estimations (green dashed line) based on the simulated data sets from Figure 3 correspondingly.

	method	n=50	n=100	n=250	n=500
$\sigma = 0.02$	ours	1.03(0.18)	0.79(0.11)	0.62(0.068)	0.47(0.07)
	locpol	1.58(0.45)	0.98(0.22)	0.70(0.11)	0.54(0.08)
	pspline	1.05(0.87)	0.80(0.82)	0.41(0.18)	0.60(0.78)
$\sigma = 0.1$	ours	2.40(0.55)	2.03(0.39)	1.46(0.27)	1.26(0.25)
	locpol	3.90(1.53)	2.93(1.71)	1.79(0.46)	1.52(1.14)
	pspline	2.53(2.32)	3.54(8.33)	1.86(2.79)	2.36(3.91)
$\sigma = 0.5$	ours	6.63(2.05)	5.05(1.61)	4.08(0.96)	3.27(0.90)
	locpol	9.48(2.70)	8.16(5.07)	5.80(3.47)	4.38(2.20)
	pspline	8.23(11.9)	8.00(15.1)	7.52(12.3)	4.77(11.8)

Table 2: Adjusted Mean Absolute Error for the second order derivative estimation.

We evaluate our method with two other well-known methods by Monte Carlo studies, that is local polynomial regression with $p = 5$ (R packages `locpol`, Cabrera, 2012) and penalized smoothing splines with $norder = 6$ and $method = 4$ (R packages `pspline`, Ramsay and Ripley, 2013) in model (1). For the sake of simplicity, we set the mean function

$$m(x) = \sin(2\pi x), \quad x \in [-1, 1].$$

We consider four sample sizes, $n \in \{50, 100, 250, 500\}$, and three standard deviations, $\sigma \in \{0.02, 0.1, 0.5\}$. The number of repetitions is set as 100. Table 2 indicates that our estimator is superior to the others in both mean and standard deviation.

6. Discussion

In this paper we propose a new methodology to estimate derivatives in nonparametric regression. The method includes two main steps: construct a sequence of symmetric difference quotients, and estimate the derivative using locally weighted least squares regression. The construction of a difference sequence is particularly important, since it determines the estimation accuracy. We consider three basic principles to construct a difference sequence. First, we eliminate the terms before the derivative of interest through linear combinations, the derivative is thus put in the important place. Second, we adopt every dependent variable only once, which keeps the independence of the difference sequence's terms. Third, we retain one or two terms behind the derivative of interest in the derived linear regression, which reduces estimation bias.

Our method and the local polynomial regression (LPR) have a close relationship. Both methods rely on Taylor expansion and employ the idea of locally weighted fitting. However, there are important differences between them. The first difference is the aim of estimation. The aim of LPR is to estimate the mean, the derivative estimation is only a "by-product", while the aim of our method is to estimate the derivative directly. The second difference is the method of weighting. LPR is kernel-weighted, the farther the distance, the lower the weight; our weight is based on variance, which can be computed exactly. Our simulation studies show that our estimator is more efficient than the LPR in most cases.

All results have been derived for equidistant design with independent identical distributed errors, and extension to more general designs is left to further research. Also, the boundary problem deserve further consideration.

Acknowledgments

We would like to thank Ping Yu at the University of Hong Kong for useful discussions, and two anonymous reviewers and associate editor for their constructive comments on improving the quality of the paper. The research was supported by NNSF projects (11171188, 11571204, and 11231005) of China.

Appendix A. Proof of Theorem 1

For (8), we yield $Var[\hat{\beta}_i] = Var[(D^\top WD)^{-1}D^\top WY_i^{(1)}] = \frac{\sigma^2}{2}(D^\top WD)^{-1}$. We can compute

$$D^\top WD = \begin{pmatrix} I_2/n^2 & I_4/n^4 \\ I_4/n^4 & I_6/n^6 \end{pmatrix},$$

where $I_l = \sum_{j=1}^{k} j^l$, l is an integer. Using the formula for the inverse of a matrix, we have

$$(D^\top WD)^{-1} = \frac{n^8}{I_2 I_6 - I_4^2} \begin{pmatrix} I_6/n^6 & -I_4/n^4 \\ -I_4/n^4 & I_2/n^2 \end{pmatrix}.$$

Therefore the variance of $\hat{\beta}_{i0}$ is

$$Var[\hat{\beta}_{i0}] = \frac{\sigma^2}{2} e_1^\top (D^\top WD)^{-1} e_1 = \frac{75\sigma^2}{8}\frac{n^2}{k^3} + o(\frac{n^2}{k^3}).$$

Appendix B. Proof of Theorem 2

From (8), we yield $E[\hat{\beta}_i] = E[(D^\top WD)^{-1}D^\top WY_i^{(1)}] = \beta + (D^\top WD)^{-1}D^\top WE[\delta_i]$. So we have

$$Bias[\hat{\beta}_i] = (D^\top WD)^{-1}D^\top WE[\delta_i].$$

Since m is five times continuously differentiable, the following Taylor expansions are valid for $m(x_{i\pm j})$ around x_i

$$\begin{aligned} m(x_{i\pm j}) &= m(x_i) + m^{(1)}(x)(\frac{\pm j}{n}) + \frac{m^{(2)}(x)}{2!}(\frac{\pm j}{n})^2 + \frac{m^{(3)}(x)}{3!}(\frac{\pm j}{n})^3 + \frac{m^{(4)}(x)}{4!}(\frac{\pm j}{n})^4 \\ &\quad + \frac{m^{(5)}(x)}{5!}(\frac{\pm j}{n})^5 + o\left((\frac{\pm j}{n})^5\right). \end{aligned}$$

We have

$$\begin{aligned} Y_{ij}^{(1)} &= \frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}} \\ &= \frac{m^{(1)}(x_i)(\frac{2j}{n}) + \frac{m^{(3)}(x_i)}{3}(\frac{j}{n})^3 + \frac{m^{(5)}(x_i)}{60}(\frac{j}{n})^5 + o\left((\frac{j}{n})^5\right) + (\epsilon_{i+j} - \epsilon_{i-j})}{2j/n} \\ &= m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6}(\frac{j}{n})^2 + \frac{m^{(5)}(x_i)}{120}(\frac{j}{n})^4 + o\left((\frac{j}{n})^4\right) + \frac{\epsilon_{i+j} - \epsilon_{i-j}}{2j/n} \\ &= m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6}(\frac{j}{n})^2 + \delta_{ij}. \end{aligned}$$

So

$$\begin{aligned} E[\delta_i] &= \frac{m^{(5)}(x_i)}{120}\begin{pmatrix} 1^4/n^4 \\ 2^4/n^4 \\ \vdots \\ k^4/n^4 \end{pmatrix} + o(\begin{pmatrix} 1^4/n^4 \\ 2^4/n^4 \\ \vdots \\ k^4/n^4 \end{pmatrix}), \\ Bias[\hat{\beta}_i] &= \frac{m^{(5)}(x_i)}{120}\frac{k^4}{n^4}\begin{pmatrix} -5/21 \\ 10/9 \end{pmatrix} + o(\frac{k^4}{n^4}). \end{aligned}$$

The estimation bias is $Bias[\hat{\beta}_{i0}] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + o(\frac{k^4}{n^4})$.

Appendix C. Proof of Theorem 3

Using the asymptotic theory of least squares and the fact that $\{\delta_{ij}\}_{j=1}^k$ are independent distributed with mean zeros and variance $\{\frac{n^2\sigma^2}{2j^2}\}_{j=1}^k$, it follows that the asymptotic normality is proved.

Appendix D. Proof of Corollary 4

For the first derivative estimation, the mean square error is given by

$$\begin{aligned} MSE[\hat{m}^{(1)}(x_i)] &= (Bias[\hat{m}^{(1)}(x_i)])^2 + Var[\hat{m}^{(1)}(x_i)] \\ &= \frac{(m^{(5)}(x_i))^2}{254016}\frac{k^8}{n^8} + \frac{75\sigma^2}{8}\frac{n^2}{k^3} + o(\frac{k^8}{n^8}) + o(\frac{n^2}{k^3}). \end{aligned}$$

Ignoring higher order terms, we obtain the asymptotic mean square error

$$AMSE(\hat{m}^{(1)}(x_i)) = \frac{(m^{(5)}(x_i))^2}{254016}\frac{k^8}{n^8} + \frac{75\sigma^2}{8}\frac{n^2}{k^3}. \tag{19}$$

To minimize (19) with respect to k, we take the first derivative of (19) and yield the gradient

$$\frac{d[AMSE(\hat{m}^{(1)}(x_i))]}{dk} = \frac{(m^{(5)}(x_i))^2}{31752}\frac{k^7}{n^8} - \frac{225\sigma^2}{8}\frac{n^2}{k^4},$$

our optimization problem is to solve $\frac{d[AMSE(\hat{m}^{(1)}(x_i))]}{dk} = 0$. So we obtain

$$k_{opt} = \left(\frac{893025\sigma^2}{(m^{(5)}(x_i))^2}\right)^{1/11} n^{10/11} \doteq 3.48\left(\frac{\sigma^2}{(m^{(5)}(x_i))^2}\right)^{1/11} n^{10/11},$$

and

$$AMSE(\hat{m}^{(1)}(x_i)) \doteq 0.31(\sigma^{16}(m^{(5)}(x_i))^6)^{1/11} n^{-8/11}.$$

Appendix E. Proof of Theorem 5

The conditional variance of $\hat{\beta}_{i0}$ is $Var[\hat{\beta}_{i0}|\epsilon_i] = \sigma^2(D^\top WD)^{-1} = 12\sigma^2\frac{n^2}{k^3} + o(\frac{n^2}{k^3})$.

Since the conditional bias of δ_{ij} is

$$E[\delta_{ij}|\epsilon_i] = \frac{m^{(2)}(x_i)}{2!}\frac{j^1}{n^1} + o\left(\frac{j^1}{n^1}\right).$$

Thus the conditional bias of $\hat{\beta}_{i0}$ is

$$\begin{aligned} Bias[\hat{\beta}_{i0}|\epsilon_i] &= (D^\top WD)^{-1}D^\top WE[\delta_i] \\ &= \frac{m^{(2)}(x_i)}{2}\frac{k^1}{n^1} + o\left(\frac{k^1}{n^1}\right). \end{aligned}$$

Appendix F. Proof of Theorem 6

The conditional variance is given by $Var[\hat{\beta}_i|\epsilon_i] = 2\sigma^2(D^\top WD)^{-1}$. We can compute

$$D^\top WD = \begin{pmatrix} I_4/n^4 & I_6/n^6 & I_2/n^2 \\ I_6/n^6 & I_8/n^8 & I_4/n^4 \\ I_2/n^2 & I_4/n^4 & I_0/n^0 \end{pmatrix}.$$

The determinant of $D^\top WD$ is

$$|D^\top WD| = \frac{I_0I_4I_8 + 2I_2I_4I_6 - I_0I_6^2 - I_4^3 - I_2^2I_8}{n^{12}},$$

and the adjoint matrix is

$$(D^\top WD)^\star = \begin{pmatrix} (I_0I_8 - I_4^2)/n^8 & (I_2I_4 - I_0I_6)/n^6 & (I_4I_6 - I_2I_8)/n^{10} \\ (I_2I_4 - I_0I_6)/n^6 & (I_0I_4 - I_2^2)/n^4 & (I_2I_6 - I_4^2)/n^8 \\ (I_4I_6 - I_2I_8)/n^{10} & (I_2I_6 - I_4^2)/n^8 & (I_4I_8 - I_6^2)/n^{12} \end{pmatrix}.$$

Based on the formula for the inverse of a matrix $A^{-1} = \frac{1}{|A|}A^{\star}$, we have

$$Var[\hat{\beta}_{i0}|\epsilon_i] = 2\sigma^2 e_1^\top (D^\top W D)^{-1} e_1 = \frac{2205\sigma^2}{8}\frac{n^4}{k^5} + o(\frac{n^4}{k^5}).$$

Revisit the sixth order Taylor approximation for $m(x_{i\pm j})$ around x_i

$$\begin{aligned} m(x_{i\pm j}) &= m(x_i) + m^{(1)}(x_i)(\frac{\pm j}{n}) + \frac{m^{(2)}(x_i)}{2!}(\frac{\pm j}{n})^2 + \frac{m^{(3)}(x_i)}{3!}(\frac{\pm j}{n})^3 + \frac{m^{(4)}(x_i)}{4!}(\frac{\pm j}{n})^4 \\ &+ \frac{m^{(5)}(x_i)}{5!}(\frac{\pm j}{n})^5 + \frac{m^{(6)}(x_i)}{6!}(\frac{\pm j}{n})^6 + o\left((\frac{\pm j}{n})^6\right). \end{aligned}$$

We have

$$\begin{aligned} Y_{ij}^{(2)} &= \frac{m(x_{i-j}) - 2m(x_i) + m(x_{i+j})}{j^2/n^2} + \frac{\epsilon_{i-j} - 2\epsilon_i + \epsilon_{i+j}}{j^2/n^2} \\ &= m^{(2)}(x_i) + \frac{m^{(4)}(x_i)}{12}\frac{j^2}{n^2} + (-2\epsilon_i)\frac{n^2}{j^2} + \frac{m^{(6)}(x_i)}{360}\frac{j^4}{n^4} + o\left(\frac{j^4}{n^4}\right) + \frac{\epsilon_{i-j} + \epsilon_{i+j}}{j^2/n^2}. \end{aligned}$$

So the conditional mean is

$$E[\delta_i|\epsilon_i] = \frac{m^{(6)}(x_i)}{360}\begin{pmatrix} 1^4/n^4 \\ 2^4/n^4 \\ \vdots \\ k^4/n^4 \end{pmatrix} + o(\begin{pmatrix} 1^4/n^4 \\ 2^4/n^4 \\ \vdots \\ k^4/n^4 \end{pmatrix}),$$

and the conditional bias is

$$\begin{aligned} Bias[\hat{\beta}_i|\epsilon_i] &= (D^\top W D)^{-1} D^\top W E[\delta_i|\epsilon_i] \\ &= \frac{m^{(6)}(x_i)}{360}\begin{pmatrix} 5/11 & & \\ & 15/11 & \\ & & 5/231 \end{pmatrix}\begin{pmatrix} k^4/n^4 \\ k^2/n^2 \\ k^6/n^6 \end{pmatrix} + o(\begin{pmatrix} k^4/n^4 \\ k^2/n^2 \\ k^6/n^6 \end{pmatrix}). \end{aligned}$$

We get

$$Bias[\hat{\beta}_{i0}|\epsilon_i] = -\frac{m^{(6)}(x_i)}{792}\frac{k^4}{n^4} + o(\frac{k^4}{n^4}).$$

Appendix G. Proof of Corollary 7

Using the asymptotic theory of least square and the fact that $\{\delta_{ij}\}_{j=1}^k$ are independent distributed with conditional mean zeros and conditional variance $\{\frac{2\sigma^2 n^4}{j^4}\}_{j=1}^k$, it follows that the asymptotic normality is proved.

Appendix H. Proof of Corollary 8

For the second derivative estimation, the MSE is

$$\begin{aligned} MSE[\hat{m}^{(2)}(x_i)|\epsilon_i] &= Bias[\hat{m}^{(2)}(x_i)]^2 + Var[\hat{m}^{(2)}(x_i)] \\ &= \frac{(m^{(6)}(x_i))^2}{627264}\frac{k^8}{n^8} + \frac{2205\sigma^2}{8}\frac{n^4}{k^5} + o(\frac{n^4}{k^5}) + o(\frac{k^8}{n^8}). \end{aligned}$$

Ignoring higher order terms, we get AMSE

$$AMSE[\hat{m}^{(2)}(x_i)|\epsilon_i] = \frac{(m^{(6)}(x_i))^2}{627264}\frac{k^8}{n^8} + \frac{2205\sigma^2}{8}\frac{n^4}{k^5}. \tag{20}$$

To minimize (20) with respect to k, take the first derivative of (20) and yield the gradient

$$\frac{d[AMSE[\hat{m}^{(2)}(x_i)|\epsilon_i]]}{dk} = \frac{(m^{(6)}(x_i))^2}{78408}\frac{k^7}{n^8} - \frac{11025\sigma^2}{8}\frac{n^4}{k^6},$$

our optimization problem is to solve $\frac{d[AMSE[\hat{m}^{(2)}(x_i)|\epsilon_i]]}{dk} = 0$. So we obtain

$$k_{opt} = \left(\frac{108056025\sigma^2}{(m^{(6)}(x_i))^2}\right)^{1/13} n^{12/13} \doteq 4.15\left(\frac{\sigma^2}{(m^{(6)}(x_i))^2}\right)^{1/13} n^{12/13},$$

and

$$AMSE(\hat{m}^{(1)}(x_i)) \doteq 0.36\left(\sigma^{16}(m^{(6)}(x_i))^{10}\right)^{1/13} n^{-8/13}.$$

Appendix I. Convergence Rates

In Table 3, we give the convergence rate of mean estimator and the first order derivative estimator in LPR. $p = 1$ means that the order of LPR is 1. Var_0 represents the convergence rate of the variance of the mean estimator, Var_1 represents the convergence rate of the variance of the first order derivative estimator. $\widetilde{MSE}_1$ stands for the convergence rate of the mean square error of first order derivative estimator when $k = k_0$,

	Var_0	$Bias_0^2$	k_0	MSE_0	Var_1	$Bias_1^2$	k_1	MSE_1	$\widetilde{MSE}_1$
p=1	$1/k$	k^4/n^4	$n^{4/5}$	$n^{-4/5}$	n^2/k^3	k^4/n^4	$n^{6/7}$	$n^{-4/7}$	$n^{-2/5}$
p=2	$1/k$	k^8/n^8	$n^{8/9}$	$n^{-8/9}$	n^2/k^3	k^4/n^4	$n^{6/7}$	$n^{-4/7}$	$n^{-4/9}$
p=3	$1/k$	k^8/n^8	$n^{8/9}$	$n^{-8/9}$	n^2/k^3	k^8/n^8	$n^{10/11}$	$n^{-8/11}$	$n^{-2/3}$
p=4	$1/k$	k^{12}/n^{12}	$n^{12/13}$	$n^{-12/13}$	n^2/k^3	k^8/n^8	$n^{10/11}$	$n^{-8/11}$	$n^{-8/13}$

Table 3: The convergence rates for mean estimator and the first order derivative estimator.

References

J.L.O. Cabrera. locpol: Kernel local polynomial regression. R packages version 0.6-0, 2012. URL `http://mirrors.ustc.edu.cn/CRAN/web/packages/locpol/index.html`.

R. Charnigo, M. Francoeur, M.P. Mengüç, A. Brock, M. Leichter, and C. Srinivasan. Derivatives of scattering profiles: tools for nanoparticle characterization. *Journal of the Optical Society of America*, 24(9):2578–2589, 2007.

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

P. Chaudhuri and J.S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

K. De Brabanter, J. De Brabanter, B. De Moor, and I. Gijbels. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14(1):281–301, 2013.

M. Delecroix and A.C. Rosa. Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *Journal of Nonparametric Statistics*, 6(4):367–382, 2007.

R.L. Eubank and P.L. Speckman. Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301, 1993.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London, 1996.

I. Gijbels and A.-C. Goderniaux. Data-driven discontinuity detection in derivatives of a regression function. *Communications in Statistics-Theory and Methods*, 33(4):851–871, 2005.

B. Hall. *Nonparametric Estimation of Derivatives with Applications.* PhD thesis, University of Kentucky, Lexington, Kentucky, 2010.

P. Hall, J.W. Kay, and D.M. Titterington. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528, 1990.

W. Härdle. *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, 1990.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models: An Introduction.* Springer, Berlin, 2004.

N.E. Heckman and J.O. Ramsay. Penalized regression with model-based penalties. *The Canadian Journal of Statistics*, 28(2):241–258, 2000.

J.L. Horowitz. *Semiparametric and Nonparametric Methods in Econometrics.* Springer, New York, 2009.

L. Lin and F. Li. Stable and bias-corrected estimation for nonparametric regression models. *Journal of Nonparametric Statistics*, 20(4):283–303, 2008.

H.-G. Müller. *Nonparametric Regression Analysis of Longitudinal Data.* Springer, New York, 1988.

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

J. Newell and J. Einbeck. A comparative study of nonparametric derivative estimators. In *Proceedings of the 22nd International Workshop on Statistical Modelling*, pages 449–452, Barcelona, 2007.

J. Newell, J. Einbeck, N. Madden, and K. McMillan. Model free endurance markers based on the second derivative of blood lactate curves. In *Proceedings of the 20th International Workshop on Statistical Modelling*, pages 357–364, Sydney, 2005.

C. Park and K.-H Kang. SiZer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis*, 52(8):3954–3970, 2008.

J. Ramsay. Derivative estimation. StatLib-S news, 1998. URL `http://www.math.yorku.ca/Who/Faculty/Monette/S-news/0556.html`.

J. Ramsay and B. Ripley. pspline: Penalized smoothing splines. R packages version 1.0-16, 2013. URL `http://mirrors.ustc.edu.cn/CRAN/web/packages/pspline/index.html`.

J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York, 2002.

D. Ruppert and M.P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3):1346–1370, 1994.

D. Ruppert, S.J. Sheather, and M.P. Wand. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270, 1995.

B. Seifert, T. Gasser, and A. Wolf. Nonparametric estimation of residual variance revisited. *Biometrika*, 80(2):373–383, 1993.

C.J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 13 (2):689–705, 1985.

T. Tong and Y. Wang. Estimating residual variance in nonparametric regression using least squares. *Biometrika*, 92(4):821–830, 2005.

S. Zhou and D.A. Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, 10 (1):93–108, 2000.