Computational Statistics and Data Analysis 105 (2017) 125–143

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Asymptotically optimal differenced estimators of error variance in nonparametric regression

WenWu Wang [a,*], Ping Yu [b]

[a] Zhongtai Institute for Financial Studies & School of Mathematics, Shandong University, Jinan 250100, China

[b] School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received 14 May 2015
Received in revised form 13 July 2016
Accepted 16 July 2016
Available online 3 August 2016

Keywords:
Bias correction
Difference order
Error estimation
Kernel estimation
Optimal difference sequence
Quadratic form
Taylor expansion

ABSTRACT

The existing differenced estimators of error variance in nonparametric regression are interpreted as kernel estimators, and some requirements for a "good" estimator of error variance are specified. A new differenced method is then proposed that estimates the errors as the intercepts in a sequence of simple linear regressions and constructs a variance estimator based on estimated errors. The new estimator satisfies the requirements for a "good" estimator and achieves the asymptotically optimal mean square error. A feasible difference order is also derived, which makes the estimator more applicable. To improve the finite-sample performance, two bias-corrected versions are further proposed. All three estimators are equivalent to some local polynomial estimators and thus can be interpreted as kernel estimators. To determine which of the three estimators to be used in practice, a rule of thumb is provided by analysis of the mean square error, which solves an open problem in error variance estimation which difference sequence to be used in finite samples. Simulation studies and a real data application corroborate the theoretical results and illustrate the advantages of the new method compared with the existing methods.

1. Introduction

Consider the nonparametric regression model

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

where the design points x_i satisfy $0 \le x_1 < x_2 < \cdots < x_n \le 1$, m is an unknown smooth mean function, and ϵ_i, $i = 1, \ldots, n$, are independent and identically distributed random errors with zero mean and variance σ^2. Estimation of σ^2 is an important topic in statistics. It is required in constructing confidence intervals, in checking goodness of fit, outliers, and homoscedasticity, and also in estimating detection limits of immunoassay.

Most estimators of σ^2 proposed in the literature are quadratic forms of the observation vector $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, namely,

$$\hat{\sigma}_W^2 = Y^{\mathrm{T}}\tilde{W}Y/tr(\tilde{W}) \triangleq Y^{\mathrm{T}}WY, \tag{2}$$

for some matrix $\tilde{W}$, where Y^{T} means Y's transpose, $tr(\tilde{W})$ means $\tilde{W}$'s trace, and $W = \tilde{W}/tr(\tilde{W})$. Roughly speaking, there are two methods to obtain these estimators: the residual-based method and the differenced method. In the residual-based

* Corresponding author.
E-mail addresses: wengewsh@sina.com (W. Wang), pingyu@hku.hk (P. Yu).

http://dx.doi.org/10.1016/j.csda.2016.07.012
0167-9473/

method, one usually fits the mean function m first by smoothing spline (Wahba, 1978; Buckley et al., 1988; Carter and Eagleson, 1992; Carter et al., 1992) or by kernel regression (Müller and Stadtmüller, 1987; Hall and Carroll, 1989; Hall and Marron, 1990; Neumann, 1994), and then estimates the variance σ^2 by the residual sum of squares. In general, the fitted values of Y are $\hat{Y} = HY$ for a linear smoother matrix H, and σ^2 is then estimated in the form of (2) with $\tilde{W} = (I-H)^{\mathrm{T}}(I-H)$, where I is the identity matrix. Buckley et al. (1988) showed that estimators based on minimax methods achieve the asymptotically optimal mean square error (MSE) $n^{-1}\mathrm{var}(\epsilon^2)$; Hall and Marron (1990) proposed a kernel-based estimator which achieves the optimal MSE

$$n^{-1}(\mathrm{var}(\epsilon^2) + O(n^{-(4r-1)/(4r+1)})), \tag{3}$$

where r is the order of m's derivative, and the rate with $r = 2$ is that achieved in Buckley et al. (1988). They also pointed out that optimal estimation of σ^2 demands less smoothing than optimal estimation of m. In spite of asymptotic effectiveness of these estimators, they depend critically on some smoothing condition in practical applications (Seifert et al., 1993)

$$\int \{m^{(r)}(x)\}^2 \mathrm{d}x/\sigma^2 \leq c_r,$$

for some smoothness order r and constant c_r, e.g., $r = 2$ in Buckley et al. (1988). Given that our target is σ^2, while such estimators require knowledge about m, it is commonly believed that these estimators are "indirect" for the estimation of σ^2.

The differenced method does not require estimation of the mean function, rather, it uses differencing to remove the trend in the mean function, an idea originating in mean square successive difference (von Neumann, 1941) and time series analysis (Anderson, 1971). Rice (1984) proposed the first-order differenced estimator

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2.$$

Later, Hall et al. (1990) generalized to the higher-order differenced estimator

$$\hat{\sigma}_{HKT}^2 = \frac{1}{n-k_1-k_2} \sum_{i=k_1+1}^{n-k_2} \left(\sum_{j=-k_1}^{k_2} d_j y_{i+j} \right)^2,$$

where $k_1, k_2 \geq 0$, $k_1 + k_2$ is referred to as the difference order, and $d_{-k_1}, \ldots, d_{k_2}$ satisfy $d_{-k_1} d_{k_2} \neq 0$, and

$$\sum_{j=-k_1}^{k_2} d_j^2 = 1, \qquad \sum_{j=-k_1}^{k_2} d_j = 0. \tag{4}$$

The first condition in (4) ensures the asymptotic unbiasedness of the variance estimator, and the second condition removes the constant term of $m(x_i)$ from the viewpoint of Taylor expansion. Obviously, $(d_{-1}, d_0) = \left(-1/\sqrt{2}, 1/\sqrt{2}\right)$ in Rice (1984) satisfies these two conditions. Gasser et al. (1986) proposed the second-order differenced estimator

$$\hat{\sigma}_{GSJ}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} \frac{((x_{i+1}-x_i)\, y_{i-1} - (x_{i+1}-x_{i-1}) y_i + (x_i - x_{i-1}) y_{i+1})^2}{(x_{i+1}-x_i)^2 + (x_{i+1}-x_{i-1})^2 + (x_i - x_{i-1})^2},$$

whose difference sequence satisfies the former two conditions (4), and an implied condition

$$\sum_{j=-1}^{1} d_{i,j} x_{i+j} = 0. \tag{5}$$

Note here that the difference sequence $\left\{d_{i,j}\right\}_{j=-1}^{1}$ depends on i; for equidistant design

$$\hat{\sigma}_{GSJ}^2 = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} \left(\frac{1}{2} y_{i-1} - y_i + \frac{1}{2} y_{i+1} \right)^2, \tag{6}$$

whose difference sequence does not depend on i. The new condition (5) further eliminates the first-order term of $m(x_i)$ besides the constant term, and results in less bias in variance estimation. Seifert et al. (1993) further developed the idea through constraining

$$\sum_{j=-k_1}^{k_2} d_{i,j} r(x_{i+j}) = 0,$$

where $r(\cdot)$ is an "unknown" smooth function for the same purpose of bias-correction. Seifert et al. (1993) showed that Gasser et al. (1986)'s estimator is a better choice than Hall et al. (1990)'s estimator; Dette et al. (1998) compared Hall et al. (1990)'s

optimal differenced estimators and ordinary differenced estimators, and showed that ordinary differenced estimators have a much better overall performance because they control the bias much better.

It seems that the residual-based method and the differenced method are two sharply different methods since the former need estimate $m(\cdot)$ while the latter need not. However, this distinction is only an illusion. Dette et al. (1998) indicated implicitly the connection between these two methods: the differences order has an influence which is comparable with the smoothing parameter. In Section 2, we show explicitly that four most used differenced estimators can be interpreted as kernel estimators. Meanwhile, we point out the weaknesses of each estimator, and specify some requirements that a "good" estimator of σ^2 should satisfy. In Section 3, we put forward our estimator which satisfies the requirements specified in Section 2. Our estimator is a weighted combination of second-order differences with different lags, so it is a differenced estimator. On the other hand, the purpose of combining different second-order differences is to estimate ϵ_i (or equivalently, $m(x_i)$), so it is also a kernel estimator. We show that this estimator achieves the asymptotically optimal MSE $n^{-1}(\mathrm{var}(\epsilon^2) + O(n^{-1/2}))$ as derived in Tong and Wang (2005) and Wang et al. (2016). We also provide a feasible difference order which is based on balancing two higher-order variance terms of our estimator. This makes our estimator more applicable.

As a differenced estimator, the asymptotic MSE of our estimator is determined by its variance and the bias is negligible. In finite samples, however, the bias component may also be important. To improve the finite-sample performance of our estimator, we further consider its two bias-corrected versions in Section 4. Bias correction is also considered in Tong and Wang (2005) but there is a key difference. Their estimator is a weighted combination of the lagged Rice-type estimators

$$\hat{\sigma}^2_{TW} = \sum_{k=1}^{m} w_k \hat{\sigma}^2_R(k), \quad \text{with } \hat{\sigma}^2_R(k) = \frac{1}{2(n-k)} \sum_{i=k+1}^{n} (y_i - y_{i-k})^2,$$

where w_k are weights. It is hard to improve to higher-order bias correction. While our estimators correct the bias in the summands of different second-order sequences (or the bias in estimating ϵ_i) rather than the bias of σ^2 estimators, so can further correct the bias of variance estimate. We show that our two bias-corrected estimators are equivalent to kernel estimators with higher-order kernels, so as in kernel smoothing, smaller biases are attributed to the usage of higher-order kernels. Relatedly, Dette et al. (1998) showed that Hall et al. (1990)'s estimator performs like a residual-based kernel-type variance estimator using a kernel of order 1.

We now have three estimators in hand; which one should be used in practice? In Section 5, we propose a rule-of-thumb to choose among the three estimators, which solves the open problem in Seifert et al. (1993) which difference sequence to be used in finite samples. Our method is straightforward—choose the estimator that minimizes the MSE including the higher-order terms. Simulation studies in Section 6 show that in most cases the performance of our estimator is better than other differenced estimators and our method of choosing the best estimator in Section 5 matches the theoretical prediction quite well. The paper concludes by some further discussions and some possible extensions of our approach. Throughout the paper, our discussion will concentrate on the equidistant design, and the extension to non-equidistant design and random design is only briefly discussed in Section 7. The proofs for all theorems are given in five Appendices. Since the proof idea of the corollaries in the paper is similar to that of the theorems, we neglect these proofs in the paper but they are available upon request.

A word on notation: $\approx$ means that the higher-order terms are omitted; for a vector α, $\|\alpha\|_2 = (\alpha^{\mathrm{T}}\alpha)^{1/2}$ is its Euclidean norm; and $\lfloor k \rfloor$ is the largest integer below k.

2. Comments on the existing differenced estimators

If ϵ_i were known, a natural estimator of σ^2 is $n^{-1}\sum_{i=1}^{n}\epsilon_i^2$. When ϵ_i is unknown, a natural idea is to estimate ϵ_i by $\hat{\epsilon}_i$ and then estimate σ^2 by a normalized $\sum \hat{\epsilon}_i^2$. However, estimating ϵ_i is equivalent to estimate $m(x_i)$ since $m(x_i) = y_i - \epsilon_i$. For a given estimator $\hat{\epsilon}_i$, $m(x_i)$ is implicitly estimated by $\hat{m}(x_i) = y_i - \hat{\epsilon}_i$. Essentially, the differenced method estimates ϵ_i directly, but as argued above, $m(x_i)$ is implicitly estimated. In this section, we show that all differenced methods implicitly estimate $m(\cdot)$ using different kernels and bandwidths. Specifically,

$$\hat{m}(x_i) = \frac{1}{nh} \sum_{j \neq i} K\left(\frac{x_j - x_i}{h}\right) y_j, \tag{7}$$

where $K(\cdot)$ is the kernel function with $\int K(u)du = 1$, and h is the bandwidth. We do not need to divide $1/(nh)\sum_{j\neq i} K(\frac{x_j - x_i}{h})$ because the density of x is known as 1 on its support $[0, 1]$ in the equidistant design. Note that y_i is excluded from the kernel smoothing to keep the most important information in estimating ϵ_i. More specifically, $\hat{\epsilon}_i = y_i - \hat{m}(x_i) = \epsilon_i + \big(m(x_i) - \hat{m}(x_i)\big)$; since $\hat{m}(x_i)$ does not involve y_i, ϵ_i rather than a fraction of ϵ_i appears in $\hat{\epsilon}_i$. The variance σ^2 is then estimated by a normalized $\sum \hat{\epsilon}_i^2$.

Rice (1984)'s first-order differenced estimator estimates $m(x_i)$ by y_{i-1}. This estimator of $m(x_i)$ uses the kernel function $K(u) = \mathbf{1}(-1 \leq u \leq 0)$ and the bandwidth $h = n^{-1}$, where $\mathbf{1}(\cdot)$ is the indicator function. Obviously, this estimator of σ^2 cannot work well because a one-sided kernel is used and no smoothing ($nh = 1$) is employed in estimating $m(x_i)$ so that $\hat{\epsilon}_i$ has a large variance and bias. In Gasser et al. (1986)'s estimator (6), $K(u) = \mathbf{1}(-1 \leq u \leq 1)/2$ and $h = n^{-1}$, which generates the second-order differenced estimator. This estimator should perform better (less bias) than Rice's estimator

because a symmetric kernel is used but the bandwidth is still $O(n^{-1})$. In Hall et al. (1990)'s estimator, $d_0 > 0$, so the kernel function is

$$K(u) = \begin{cases} -\sum_{j=-k_1}^{-1} k\frac{d_j}{d_0}\mathbf{1}\left(\frac{j}{k} \le u < \frac{j+1}{k}\right) - \sum_{j=1}^{k_2} k\frac{d_j}{d_0}\mathbf{1}\left(\frac{j}{k} \le u < \frac{j+1}{k}\right), & \text{if } k \text{ is even,} \\ -\sum_{j=1}^{k} k\frac{d_j}{d_0}\mathbf{1}\left(\frac{j}{k} \le u < \frac{j+1}{k}\right), & \text{if } k \text{ is odd,} \end{cases}$$

where $k = k_1 + k_2$. That is, $K(u)$ is two-sided when k is even and one-sided when k is odd. $K(u)$ is a step function with $\int K(u)du = -\sum_{j\neq 0}\frac{d_j}{d_0} = 1$ because $\sum_{j=-k_1}^{k_2} d_j = 0$ and $d_0 > 0$ imply $-\sum_{j\neq 0}\frac{d_j}{d_0} = 1$, and $h = k/n$. Dette et al. (1998) also noticed that Hall et al.'s estimator performs like a residual-based kernel estimator using a kernel of order 1 (so that the bias cannot be well controlled), and we make this kernel estimator explicit. The values of d_j's in $K(u)$ are motivated by minimizing the asymptotic variance of $\hat{\sigma}^2_{HKT}$ rather than precisely estimating $m(x_i)$, so this kernel cannot be a good choice for estimating $m(x_i)$ (especially in the aspect of bias) and so cannot be a good choice for estimating σ^2. Furthermore, since k is fixed, no smoothing is involved in their estimator. As a result, the variance of $m(x_i)$ (or equivalently ϵ_i) estimate is large. This is why neither of these three estimators reaches the optimal efficiency bound, $\mathrm{var}(\epsilon^2)$, as developed in Tong et al. (2013). In one word, the lesson here is that to achieve the efficiency bound, we must let $k \to \infty$ or $nh \to \infty$.

Tong and Wang (2005)'s estimator indeed involves smoothing in estimating $m(x_i)$ as their $k \to \infty$, so their estimator reaches the efficiency bound. However, like Hall et al. (1990)'s estimator, since their estimator is not motivated by estimating ϵ_i (or equivalently, estimating $m(x_i)$), there is some freedom in their kernel choice. From their Theorem 1, their estimator $\hat{\sigma}^2_{TW}$ can be written in the form of (2), where $\tilde{W}$ is symmetric but need not be positive definite (so $\hat{\sigma}^2_{TW}$ need not be positive). As a result, $\tilde{W}$ can have different decompositions, such as the LU or QR decomposition (but cannot be decomposed as $A^{\mathrm{T}}A$ for some matrix A). For example, in the LU decomposition, with proper row and/or column permutations, $\tilde{W} = L^{\mathrm{T}}U$, where L and U are upper triangular. Given that LY and UY are estimating $\{\epsilon_i\}_{i=1}^n$, Tong and Wang implicitly estimate $\{\epsilon_i\}_{i=1}^n$ in two different ways and then use the inner product of the two $\hat{\epsilon}_i$ sequences to estimate σ^2. There is no reason why such an estimator of σ^2 should perform well.

In summary, a good estimator of σ^2 should satisfy at least two conditions. First, the implied bandwidth h satisfies $nh \to \infty$. Second, the implied estimator of $m(x_i)$ (or ϵ_i) has the interpretation of a kernel smoother. In other words, the estimator should have the form (2) with $\tilde{W} = \tilde{D}^{\mathrm{T}}\tilde{D}$, where

$$\tilde{D} = \begin{pmatrix} \tilde{d}_{-k} & \cdots & \tilde{d}_k & 0 & \cdots & 0 \\ & \ddots & & \ddots & & \\ & & \ddots & & \ddots & \\ 0 & \cdots & 0 & \tilde{d}_{-k} & \cdots & \tilde{d}_k \end{pmatrix}_{(n-2k)\times n}. \tag{8}$$

We normalize $\tilde{d} \triangleq \left(\tilde{d}_{-k}, \ldots, \tilde{d}_0, \ldots, \tilde{d}_k\right)$ as $d \triangleq (d_{-k}, \ldots, d_0, \ldots \cdots, d_k) = \tilde{d}/\|\tilde{d}\|_2$, and denote the counterpart of $\tilde{D}$ as D. The normalized $d_j, j = -k, \ldots, k$, satisfy $\sum_{j=-k}^k d_j^2 = 1$, $d_0 \to 1$ and $d_j \to 0$ for $j = \pm 1, \ldots, \pm k$ as $k \to \infty$. When a symmetric kernel is used, $d_{-j} = d_j, j = 1, \ldots, k$. Our estimators to be developed satisfy these requirements. Our difference sequences $\{d_j\}_{j=-k}^k$ further satisfy some orthogonality conditions that eliminate the higher-order biases in estimating m.

On the other hand, although our estimators can be interpreted as kernel estimators, they indeed explores special structures of the data design, namely, ϵ_i is i.i.d. and x_i is equally spaced. Ignoring these special information may incur practical difficulties in implementation (see the discussion in Section 7). Our estimator combines the estimation ideas in both kinds of estimators, so is different from and does not suffer the drawbacks of any existing estimator.

3. Estimation methodology

As in the differenced estimator, we estimate ϵ_i directly to estimate σ^2. We first use a simple example to illustrate our estimation strategy of ϵ_i. Suppose

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 0, \ldots, n), \tag{9}$$

where x_i's are equidistantly designed (i.e., $x_i = i/n$), and ϵ_i's are i.i.d. random errors with zero mean and variance σ^2. Suppose $n = 100$; our target is to estimate ϵ_{50} which is a random variable rather than a fixed parameter as in the usual case. A common solution is to first estimate $\beta \triangleq (\beta_0, \beta_1)^{\mathrm{T}}$ by the least squares and then estimate ϵ_{50} by $\tilde{\epsilon}_{50} = y_{50} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{50}$, where $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1)^{\mathrm{T}}$ is the least squares estimator (LSE) of β. Our solution is different: we first remove the trend in (9) by symmetric second-order differencing and then estimate ϵ_{50} by the least squares. In spirit, our method is similar to the

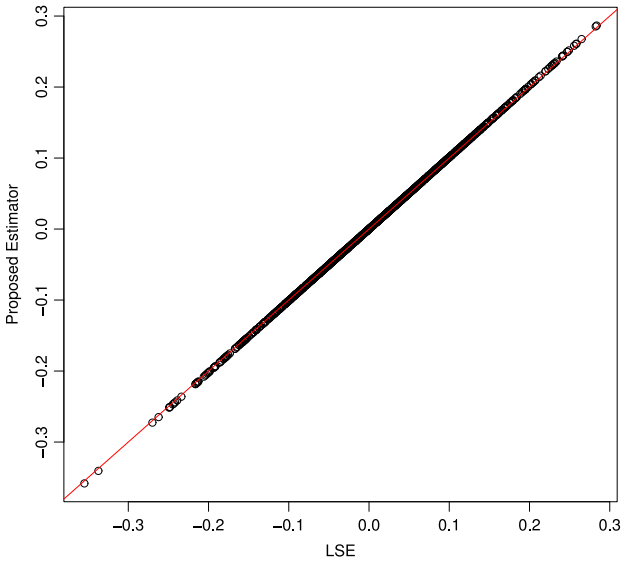

Fig. 1. Scatterplot of the proposed estimator against the LSE of ϵ_{50}: the red line is $y = x$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

derivative estimation method of Wang and Lin (2015). Specifically, define $y_j^{(2)} = (-y_{50-j} + 2y_{50} - y_{50+j})/2$ $(j = 1, \ldots, 50)$. It is easy to see that the trend is removed in $y_j^{(2)}$,

$$y_j^{(2)} = \epsilon_{50} + \varepsilon_j, \tag{10}$$

where $\varepsilon_j = -(\epsilon_{50-j} + \epsilon_{50+j})/2$ is independent of ϵ_{50} and has variance $\sigma^2/2$. Now, ϵ_{50} is like a random intercept in the regression of $y_j^{(2)}$ on a constant, so can be estimated by the least squares. In other words, we estimate ϵ_{50} by $\hat{\epsilon}_{50} = \bar{y}^{(2)}$, the sample mean of $y_j^{(2)}$. The consistency of $\hat{\epsilon}_{50}$ is guaranteed by observing that

$$\mathrm{E}(y_j^{(2)} \mid \epsilon_{50}) = \epsilon_{50}.$$

Fig. 1 shows our estimates $\hat{\epsilon}_{50}$ against the least squares residuals $\tilde{\epsilon}_{50}$ when $\beta_0 = 1$, $\beta_1 = 2$ and $\sigma^2 = 1/100$. It is clear that these two estimators are almost identical in this simple scenario. When $m(\cdot)$ is nonparametric, we "localize" the idea above.

3.1. Error estimation based on the first-order Taylor expansion

Suppose the mean function m is continuously differentiable on $[0, 1]$. Then the first-order Taylor expansions of $m(x_{i\pm j})$ at x_i are

$$m(x_{i+j}) = m(x_i) + m^{(1)}(x_i)\frac{j}{n} + o\left(\frac{j}{n}\right),$$
$$m(x_{i-j}) = m(x_i) - m^{(1)}(x_i)\frac{j}{n} + o\left(\frac{j}{n}\right).$$

That is, $m(\cdot)$ behaves like a linear function in the neighborhood of x_i, just as in the simple setup (9). Consequently, we can eliminate the "local" trend $m(x_i) + m^{(1)}(x_i)j/n$ by the second-order differencing. Specifically, define

$$y_{ij}^{(2)} = \frac{-y_{i+j} + 2y_i - y_{i-j}}{2} \quad (j = 1, \ldots, k).$$

Then

$$y_{ij}^{(2)} = \frac{-m(x_{i+j}) + 2m(x_i) - m(x_{i-j})}{2} + \frac{-\epsilon_{i+j} + 2\epsilon_i - \epsilon_{i-j}}{2}$$
$$= o\left(\frac{j}{n}\right) + \epsilon_i + \varepsilon_{ij},$$

where $\varepsilon_{ij} = -(\epsilon_{i+j}+\epsilon_{i-j})/2$ is independent of ϵ_i. Note here that i is fixed while j changes. Since $\mathrm{E}(y_{ij}^{(2)} \mid \epsilon_i) = \epsilon_i + o(j/n) \approx \epsilon_i$, we can estimate ϵ_i by

$$\hat{\epsilon}_i = \arg\min_{\beta_{i0}} \sum_{j=1}^{k} (y_{ij}^{(2)} - \beta_{i0})^2 = (X_1^{\mathrm{T}} X_1)^{-1} X_1^{\mathrm{T}} Y_i^{(2)},$$

where $k \to \infty$ employs smoothing to improve efficiency, $k = o(n)$ ensures the approximation of Taylor expansion accurate enough, $X_1 = (1, \ldots, 1)^{\mathrm{T}}$ is a vector of ones of length k, $Y_i^{(2)} = (y_{i1}^{(2)}, \ldots, y_{ik}^{(2)})^{\mathrm{T}} = GY_i^{(0)}$ with

$$G = \begin{pmatrix} & & -1/2 & 1 & -1/2 & & \\ & \iddots & & \vdots & & \ddots & \\ -1/2 & & & 1 & & & -1/2 \end{pmatrix}_{k\times(2k+1)} \quad \text{and} \quad Y_i^{(0)} = \begin{pmatrix} y_{i-k} \\ \vdots \\ y_i \\ \vdots \\ y_{i+k} \end{pmatrix}_{(2k+1)\times 1}.$$

How to understand this estimator of ϵ_i? It turns out that it can be treated as either a differenced estimator or a kernel estimator. Since

$$\hat{\epsilon}_i = (X_1^{\mathrm{T}} X_1)^{-1} X_1^{\mathrm{T}} G Y_i^{(0)} = \sum_{j=-k}^{k} \tilde{d}_j y_{i+j},$$

where $\tilde{d} = (\tilde{d}_{-k}, \ldots, \tilde{d}_{-1}, \tilde{d}_0, \tilde{d}_1, \ldots, \tilde{d}_k) = (X_1^{\mathrm{T}} X_1)^{-1} X_1^{\mathrm{T}} G = \left(-\frac{1}{2k}, \ldots, -\frac{1}{2k}, 1, -\frac{1}{2k}, \ldots, -\frac{1}{2k}\right)$ is the difference sequence in (8), $\hat{\epsilon}_i$ is a $2k$th-order differenced estimator. The key difference from the existing differenced estimators is that $k \to \infty$, that is, smoothing is involved. On the other hand, since $\hat{\epsilon}_i = y_i - \hat{m}(x_i)$ with $\hat{m}(x_i) = (2k)^{-1} \sum_{j\neq i} y_j \mathbf{1}(-k \le j - i \le k)$, $\hat{\epsilon}_i$ is a kernel estimator with the uniform kernel $K(u) = \mathbf{1}(-1 \le u \le 1)/2$ and the bandwidth $h = k/n$.

The following two theorems provide asymptotic results on the bias, variance, and MSE of $\hat{\epsilon}_i$, and establish its pointwise consistency and asymptotic normality.

Theorem 1. *Assume that the nonparametric model (1) holds with equidistant design and normally distributed error. Assume further that the unknown function $m(\cdot)$ is twice continuously differentiable on* $[0, 1]$. *Then*

$$\mathrm{var}(\hat{\epsilon}_i \mid \epsilon_i) = \frac{1}{2}\frac{\sigma^2}{k}$$

uniformly over $k + 1 \le i \le n - k$, and

$$\mathrm{bias}(\hat{\epsilon}_i \mid \epsilon_i) \approx -\frac{m^{(2)}(x_i)}{6}\frac{k^2}{n^2}$$

for $k + 1 \le i \le n - k$. The optimal k that minimizes the asymptotic MSE (AMSE) of $\hat{\epsilon}_i$ is

$$k_{opt} = 1.35 \left(\frac{\sigma^2}{(m^{(2)}(x_i))^2}\right)^{1/5} n^{4/5},$$

and the corresponding AMSE is

$$\mathrm{AMSE}(\hat{\epsilon}_i) = 0.46 \left\{\sigma^8 \left(m^{(2)}(x_i)\right)^2\right\}^{1/5} n^{-4/5}.$$

There is no mystery in the form of bias and variance of $\hat{\epsilon}_i$. Since $\hat{\epsilon}_i - \epsilon_i = y_i - \hat{m}(x_i) - (y_i - m(x_i)) = m(x_i) - \hat{m}(x_i)$, $\mathrm{bias}(\hat{\epsilon}_i|\epsilon_i) = -\mathrm{bias}(\hat{m}(x_i))$ and $\mathrm{var}\left(\hat{\epsilon}_i \mid \epsilon_i\right) = \mathrm{var}(\hat{m}(x_i))$. It is well known that for a kernel estimator (7), $\mathrm{bias}\left(\hat{m}(x_i)\right) \approx h^2 m^{(2)}(x_i) \int K(u)u^2 du/2$ and $\mathrm{var}(\hat{m}(x_i)) \approx \sigma^2/(nh) \int K(u)^2 du$. For the uniform kernel, $\int K(u)u^2 du/2 = 1/6$ and $\int K(u)^2 du = 1/2$. From Theorem 1, if $k \to \infty$ and $k/n \to 0$, then $\hat{\epsilon}_i - \epsilon_i \xrightarrow{p} 0$. Also the implied optimal bandwidth $h_{opt} = k_{opt}/n = O\left(n^{-1/5}\right)$, which is the popular rate of optimal bandwidth in estimating $m(x_i)$. Note that k_{opt} is not an integer in general, and we can replace it by $\lfloor k_{opt} \rfloor$ in practice. We further establish its asymptotic normality in the following theorem.

Theorem 2. *Under the assumptions of Theorem 1, if $k \to \infty$ and $k/n \to 0$, then*

$$k^{1/2}\left(\hat{\epsilon}_i - \epsilon_i + \frac{m^{(2)}(x_i)}{6}\frac{k^2}{n^2}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{2}\right)$$

for $k+1 \le i \le n-k$. *Furthermore, if* $k \to \infty$ *and* $k^{5/4}/n \to 0$, *then*

$$k^{1/2}\left(\hat{\epsilon}_i - \epsilon_i\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{2}\right)$$

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

Theorem 2 shows that with suitable choice of k our error estimator $\hat{\epsilon}_i$ is asymptotically normal. This asymptotic approximation can be used for constructing confidence intervals of $\hat{\epsilon}_i$ (or equivalently $\hat{m}(x_i)$).

3.2. Variance estimation based on the estimated errors

Given the estimated errors $\{\hat{\epsilon}_i\}_{i=k+1}^{n-k}$, σ^2 can be naturally estimated by

$$\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\hat{\epsilon}_i^2 = \frac{1}{n-2k}Y^{\mathrm{T}}\tilde{D}_1^{\mathrm{T}}\tilde{D}_1 Y,$$

where $\tilde{D}_1$ takes the form of (8) with $\tilde{d}$ defined in the last subsection. We further normalize our estimator to

$$\hat{\sigma}_{D_1}^2 = Y^{\mathrm{T}}\tilde{D}_1^{\mathrm{T}}\tilde{D}_1 Y / tr\left(\tilde{D}_1^{\mathrm{T}}\tilde{D}_1\right),$$

which takes the form of (2) with $\tilde{W} = \tilde{D}_1^{\mathrm{T}}\tilde{D}_1$. Note that $tr\left(\tilde{W}\right) = (n-2k)\left\|\tilde{d}\right\|_2^2$, so

$$\hat{\sigma}_{D_1}^2 = \frac{1}{n-2k}Y^{\mathrm{T}}D_1^{\mathrm{T}}D_1 Y = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(\sum_{j=-k}^{k} d_j y_{i+j}\right)^2,$$

where $D_1 = \tilde{D}_1/\|\tilde{d}\|_2$ is the normalized $\tilde{D}_1$ with

$$d_j = \begin{cases} \{2k/(2k+1)\}^{1/2}, & j = 0, \\ -\{2k(2k+1)\}^{-1/2}, & -k \le j \le -1,\ 1 \le j \le k, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $d_0 \to 1$ and $d_j \to 0, j \ne 0$, as $k \to \infty$; the former is to reserve ϵ_i and the latter is to remove $m(x_i)$. This 'spike' difference sequence is also intuitively suggested in Hall et al. (1990) but no theoretical justification is provided.

As in our reinterpretation of the differenced estimators in Section 2, the new estimator is motivated from estimating ϵ_i as the building block, so is intuitively interpretable: every difference term in $\tilde{D}_1 Y$ is an error estimator. With normally distributed errors, we have

$$\begin{aligned} \text{bias}\left(\hat{\sigma}_{D_1}^2\right) &= m_n^{\mathrm{T}}\tilde{W}m_n / tr(\tilde{W}), \\ \text{var}\left(\hat{\sigma}_{D_1}^2\right) &= \left(4\sigma^2 m_n^{\mathrm{T}}\tilde{W}^2 m_n + 2\sigma^4 tr(\tilde{W}^2)\right)/tr(\tilde{W})^2, \end{aligned}$$

from Buckley et al. (1988), where $m_n = (m(x_1), \ldots, m(x_n))^{\mathrm{T}}$. We now derive the asymptotic bias, variance, and MSE for the variance estimator $\hat{\sigma}_{D_1}^2$ in the following theorem.

Theorem 3. *Under the assumptions of Theorem 1,*

$$\text{bias}\left(\hat{\sigma}_{D_1}^2\right) \approx \frac{L_2 k^4}{36n^4}, \qquad \text{var}\left(\hat{\sigma}_{D_1}^2\right) \approx 2\sigma^4\left(\frac{1}{n} + \frac{2k}{n^2} + \frac{5}{6nk}\right),$$

where $L_2 = \int_0^1 \left(m^{(2)}(x)\right)^2 \mathrm{d}x$. *So the AMSE of* $\hat{\sigma}_{D_1}^2$ *is*

$$\text{AMSE}\left(\hat{\sigma}_{D_1}^2\right) \approx \frac{2\sigma^4}{n} + \frac{4\sigma^4 k}{n^2} + \frac{5\sigma^4}{3nk} + \frac{L_2^2}{36^2}\frac{k^8}{n^8},$$

and the optimal k *that minimizes the AMSE of* $\hat{\sigma}_{D_1}^2$ *is* $k_1^* = (5n/12)^{1/2}$.

From this theorem, the bias of $\hat{\sigma}_{D_1}^2$ is related to the roughness of $m^{(2)}(\cdot)$ which also appears in the asymptotic mean integrated squared error of the kernel estimation of $m(\cdot)$. This is understandable since we implicitly estimate $m(x_i)$ at all x_i, $k+1 \le i \le n-k$, to estimate σ^2. The first term of $\text{var}(\hat{\sigma}_{D_1}^2)$, $2\sigma^4/n$, is the asymptotic variance of the ideal estimator $\sum_{i=1}^n \epsilon_i^2/n$ or the efficiency bound. The second term $4\sigma^4 k/n^2$ is due to the missing $2k$ error estimators $\hat{\epsilon}_i$ for $1 \le i \le k$ and $n-k+1 \le i \le n$. The third term $5\sigma^4/(3nk)$ is caused by the correlation between the error estimators. The implied optimal

bandwidth $h^* = k^*/n = O(n^{-1/2})$ is smaller than the optimal bandwidth in estimating $m(x_i)$ but is of the same rate as that in Tong and Wang (2005). In other words, we need undersmooth the estimate of $m(x_i)$ such that $\hat{\epsilon}_i$ contains less bias than in its pointwise optimal estimation. This result matches that in Wang et al. (2008) where it is suggested to estimate $m(\cdot)$ with minimal bias in estimating the conditional variance function.

For differenced estimators, we need two inputs—the difference order $2k$ and the difference sequence $\{d_j\}_{j=-k}^{k}$. To our best knowledge, there is no theoretical result on the choice of k. By minimizing the MSE, we obtain the optimal k value. Note that this optimal k value is determined by a trade-off between two higher-order variance terms, rather than between higher-order variance terms and the bias term. This is why the MSE of our estimators is $n^{-1}\mathrm{var}(\epsilon^2)+O(n^{-3/2})$ rather than the optimal rate $n^{-1}\mathrm{var}(\epsilon^2)+O(n^{-8r/(4r+1)})$ developed in Hall and Marron (1990). Roughly speaking, the kernel estimator of Hall and Marron (1990) estimates all ϵ_i's, while we only estimate ϵ_i's for $k+1 \le i \le n-k$. This is why we have two higher-order variance terms, while they have only one. This is also why the optimal higher-order term of MSE is $O(n^{-8r/(4r+1)})$ rather than $O(n^{-3/2})$. Although less efficient in theory, our estimator is more practical because an optimal k is provided. Truncating the estimates of ϵ_i's at the boundary of x's support is a common feature of the differenced estimator. Such a truncation is understandable from the practical point of view—our target is σ^2 instead of ϵ_i, so it is preferable to use only the easily-available ϵ_i estimates rather than all ϵ_i estimates.

Remark 1. For comparison, we repeat the AMSE of Tong and Wang (2005) here,

$$\mathrm{AMSE}(\hat{\sigma}^2_{TW}) \approx \frac{2\sigma^4}{n} + \frac{9\sigma^4 k}{56n^2} + \frac{9\sigma^4}{4nk}.$$

This expression of AMSE has a similar explanation: the first term is the asymptotic variance of the ideal estimator $\sum_{i=1}^{n}\epsilon_i^2/n$; the second term is due to the missing information of $2k$ boundary error estimators; the third term is caused by the correlation between the error estimators. From the formula of $\hat{\sigma}^2_K(k)$ in Tong and Wang (2005), their estimator misses only k boundary errors but has a larger correlation term:

$$\frac{9\sigma^4 k}{56n^2} < \frac{4\sigma^4 k}{n^2}, \qquad \frac{9\sigma^4}{4nk} > \frac{5\sigma^4}{3nk}.$$

Most importantly, their estimator is hard to extend to correct higher-order biases, while such corrections are straightforward to our estimator (see Section 4).

Theorem 3 indicates that $\hat{\sigma}^2_{D_1}$ is a consistent estimator of σ^2. We further establish its asymptotic normality in the following theorem.

Theorem 4. *Under the assumptions of Theorem 1, if $k \to \infty$ and $k/n \to 0$, then*

$$n^{1/2}\left(\hat{\sigma}^2_{D_1} - \sigma^2 - \frac{L_2 k^4}{36n^4}\right) \xrightarrow{d} N\left(0, 2\sigma^4\right).$$

Furthermore, if $k^{8/7}/n \to 0$, then

$$n^{1/2}\left(\hat{\sigma}^2_{D_1} - \sigma^2\right) \xrightarrow{d} N\left(0, 2\sigma^4\right).$$

Different from most previous differenced estimators (except Tong and Wang, 2005), our estimator achieves the efficiency bound due to $k \to \infty$.

Remark 2. The above four theorems assume ϵ_i following the normal distribution to simplify proofs. Actually, we require only that ϵ_i has the fourth-order moment. Without normality, the asymptotic variance in Theorem 4 changes to $\mathrm{var}(\epsilon^2)$.

4. Bias-corrected variance estimation

In Theorem 3, the optimal k is determined solely by minimizing the asymptotic variance since the bias term is dominated at k_1^*. In finite samples, the bias term of $\hat{\sigma}^2_{D_1}$ may not be neglectable especially when $m(\cdot)$ highly oscillates (which will generate a large L_2). To alleviate this problem, we eliminate higher order terms in the bias of $\mathrm{E}(y_{ij}^{(2)}|\epsilon_i)$ by applying higher order Taylor expansions on $m(\cdot)$.

4.1. Variance estimation based on the third-order Taylor expansion

Assume that the mean function $m(\cdot)$ is three times continuously differentiable on $[0, 1]$. We can show by a similar argument as in Section 3.1 that

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

As a result,

$$y_{ij}^{(2)} = \epsilon_i - \frac{m^{(2)}(x_i)}{2}\frac{j^2}{n^2} + o\left(\frac{j^3}{n^3}\right) - \varepsilon_{ij}, \quad j = 1, \ldots, k.$$

Therefore, ϵ_i can be estimated by the minimizer $\hat{\beta}_{i0}$ in the following minimization problem:

$$\min_{\beta_{i0},\beta_{i1}} \sum_{j=1}^{k} \left(y_{ij}^{(2)} - \beta_{i0} - \beta_{i1}\frac{j^2}{n^2}\right)^2. \tag{11}$$

In other words,

$$\hat{\epsilon}_i = e_1^{(2)\mathrm{T}}(X_2^{\mathrm{T}}X_2)^{-1}X_2^{\mathrm{T}}Y_i^{(2)},$$

where

$$e_1^{(q)} = (1, \underbrace{0, \ldots, 0}_{q-1})^{\mathrm{T}} \quad \text{and} \quad X_2 = \begin{pmatrix} 1 & 1^2/n^2 \\ 1 & 2^2/n^2 \\ \vdots & \vdots \\ 1 & k^2/n^2 \end{pmatrix}_{k\times 2}.$$

Corollary 1. *Assume that the nonparametric model (1) holds with equidistant design and normally distributed error. Assume further that the unknown function $m(\cdot)$ is four times continuously differentiable on $[0, 1]$. Then*

$$\mathrm{var}(\hat{\epsilon}_i \mid \epsilon_i) \approx \frac{9}{8}\frac{\sigma^2}{k}$$

uniformly over $k + 1 \le i \le n - k$, and

$$\mathrm{bias}(\hat{\epsilon}_i \mid \epsilon_i) \approx \frac{m^{(4)}(x_i)}{280}\frac{k^4}{n^4}$$

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

Corollary 2. *Under the assumptions of Corollary 1, if $k \to \infty$ and $k^{9/8}/n \to 0$, then*

$$k^{1/2}\left(\hat{\epsilon}_i - \epsilon_i\right) \xrightarrow{d} N\left(0, \frac{9}{8}\sigma^2\right)$$

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

Given the error estimator $\{\hat{\epsilon}_i\}$, we follow the construction of $\hat{\sigma}_{D_1}^2$ and propose the new variance estimator

$$\hat{\sigma}_{D_2}^2 = Y^{\mathrm{T}}\tilde{D}_2^{\mathrm{T}}\tilde{D}_2 Y / tr(\tilde{D}_2^{\mathrm{T}}\tilde{D}_2), \tag{12}$$

where $\tilde{D}_2$ is the same as $\tilde{D}_1$ except that the difference sequence

$$(\tilde{d}_{-k}, \ldots, \tilde{d}_{-1}, \tilde{d}_0, \tilde{d}_1, \ldots, \tilde{d}_k) = e_1^{(2)\mathrm{T}}(X_2^{\mathrm{T}}X_2)^{-1}X_2^{\mathrm{T}}G.$$

Similarly as in $\hat{\sigma}_{D_1}^2$, it is not hard to show that $\tilde{d}_0 = 1$ here. We derive the bias and variance of $\hat{\sigma}_{D_2}^2$, and establish its asymptotic normality in the following two corollaries.

Corollary 3. *Under the assumptions of Corollary 1,*

$$\mathrm{bias}\left(\hat{\sigma}_{D_2}^2\right) \approx \frac{L_4 k^8}{280^2 n^8}, \qquad \mathrm{var}\left(\hat{\sigma}_{D_2}^2\right) \approx 2\sigma^4\left(\frac{1}{n} + \frac{2k}{n^2} + \frac{1.61}{nk}\right),$$

where $L_4 = \int_0^1 (m^{(4)}(x))^2 dx$. *So the AMSE of* $\hat{\sigma}^2_{D_2}$ *is*

$$\text{AMSE}\left(\hat{\sigma}^2_{D_2}\right) \approx \frac{2\sigma^4}{n} + \frac{4\sigma^4 k}{n^2} + \frac{3.22\sigma^4}{nk} + \frac{L_4^2 k^{16}}{280^4 n^{16}},$$

and the optimal k that minimizes AMSE is $k_2^* \approx 0.90 n^{1/2}$.

Corollary 4. *Under the assumptions of Corollary 1, if* $k \to \infty$ *and* $n^{-1}k^{16/15} \to 0$, *then*

$$n^{1/2}(\hat{\sigma}^2_{D_2} - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4).$$

From Corollary 3, $\text{bias}(\hat{\sigma}^2_{D_2})$ is smaller than $\text{bias}(\hat{\sigma}^2_{D_1})$ when the optimal k's are used. When the bias is still suspected to be too large, we can continue our bias correction by applying the fifth-order Taylor expansion on $m(\cdot)$.

4.2. Variance estimation based on the fifth-order Taylor expansion

Assume that the mean function $m(\cdot)$ is five times continuously differentiable on [0, 1]. We can show similarly as in the last subsection that

$$\hat{\epsilon}_i = e_1^{(3)\text{T}}(X_3^{\text{T}}X_3)^{-1}X_3^{\text{T}}Y_i^{(2)}, \tag{13}$$

where

$$X_3 = \begin{pmatrix} 1 & 1^2/n^2 & 1^4/n^4 \\ 1 & 2^2/n^2 & 2^4/n^4 \\ \vdots & \vdots & \vdots \\ 1 & k^2/n^2 & k^4/n^4 \end{pmatrix}_{k\times 3}.$$

Corollary 5. *Assume that the nonparametric model (1) holds with equidistant design and normally distributed error. Assume further that the unknown function* $m(\cdot)$ *is six times continuously differentiable on* [0, 1]. *Then*

$$\text{var}\left(\hat{\epsilon}_i \mid \epsilon_i\right) \approx \frac{225}{128}\frac{\sigma^2}{k}$$

uniformly over $k+1 \le i \le n-k$, *and*

$$\text{bias}(\hat{\epsilon}_i \mid \epsilon_i) \approx -\frac{m^{(6)}(x_i)}{33\,264}\frac{k^6}{n^6}$$

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

Corollary 6. *Under the assumptions of Corollary 5, if* $k \to \infty$ *and* $k^{13/12}/n \to 0$, *then*

$$k^{1/2}\left(\hat{\epsilon}_i - \epsilon_i\right) \xrightarrow{d} N\left(0, \frac{225}{128}\sigma^2\right)$$

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

Similar to $\hat{\sigma}^2_{D_1}$ and $\hat{\sigma}^2_{D_2}$, we construct the new variance estimator as

$$\hat{\sigma}^2_{D_3} = Y^{\text{T}}\tilde{D}_3^{\text{T}}\tilde{D}_3 Y / tr(\tilde{D}_3^{\text{T}}\tilde{D}_3), \tag{14}$$

where $\tilde{D}_3$ is the same as $\tilde{D}_1$ except that difference sequence

$$(\tilde{d}_{-k}, \ldots, \tilde{d}_{-1}, \tilde{d}_0, \tilde{d}_1, \ldots, \tilde{d}_k) = e_1^{(3)\text{T}}(X_3^{\text{T}}X_3)^{-1}X_3^{\text{T}}G.$$

Corollary 7. *Under the assumptions of Corollary 5,*

$$\text{bias}\left(\hat{\sigma}^2_{D_3}\right) \approx \frac{L_6 k^{12}}{33\,264^2 n^{12}}, \qquad \text{var}\left(\hat{\sigma}^2_{D_3}\right) \approx 2\sigma^4\left(\frac{1}{n} + \frac{2k}{n^2} + \frac{2.35}{nk}\right),$$

where $L_6 = \int_0^1 (m^{(6)}(x))^2 \text{dx}$. *So the AMSE of* $\hat{\sigma}^2_{D_3}$ *is*

$$\text{AMSE}\left(\hat{\sigma}^2_{D_3}\right) \approx \frac{2\sigma^4}{n} + \frac{4\sigma^4 k}{n^2} + \frac{4.70\sigma^4}{nk} + \frac{L_6^2 k^{24}}{33\,264^4 n^{24}},$$

and the optimal k that minimizes AMSE is $k_3^* \approx 1.08 n^{1/2}$.

Corollary 8. *Under the assumptions of Corollary 5, if* $k \to \infty$ *and* $n^{-1}k^{24/23} \to 0$, *then*

$$n^{1/2}(\hat{\sigma}^2_{D_3} - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4).$$

Like $\hat{\sigma}^2_{D_1}$, $\hat{\sigma}^2_{D_2}$ and $\hat{\sigma}^2_{D_3}$ can be interpreted as kernel estimators. Actually, the $\hat{\epsilon}_i$ estimators in $\hat{\sigma}^2_{D_2}$ and $\hat{\sigma}^2_{D_3}$ are equivalent to the local polynomial estimator (LPE) of order 3 and 5, respectively. Specifically, the pth order LPE of $m(x_i)$ is defined as the minimizer $\hat{\beta}_{i0}$ in the following minimization problem:

$$\min_{\beta_{i0},\ldots,\beta_{ip}} \sum_{j=-k, j\neq 0}^{k} \left(y_{i+j} - \beta_{i0} - \beta_{i1}\frac{j}{n} - \cdots - \beta_{ip}\left(\frac{j}{n}\right)^p \right)^2. \tag{15}$$

Explicitly,

$$\hat{\beta}^{(p)}_{i0} = e_1^{(p+1)\mathrm{T}} \left(\tilde{X}_p^{\mathrm{T}}\tilde{X}_p\right)^{-1} \tilde{X}_p^{\mathrm{T}} Y_i^{(0)},$$

where

$$\tilde{X}_p = \begin{pmatrix} 1 & -k/n & \cdots & (-k/n)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1/n & \cdots & (-1/n)^p \\ 0 & 0 & \cdots & 0 \\ 1 & 1/n & \cdots & (1/n)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & k/n & \cdots & (k/n)^p \end{pmatrix}_{(2k+1)\times(p+1)}.$$

The following theorem rigorously states this equivalence result.

Theorem 5. $\hat{\epsilon}_i$ *in* $\hat{\sigma}^2_{D_1}$ *equals* $y_i - \hat{\beta}^{(1)}_{i0}$, $\hat{\epsilon}_i$ *in* $\hat{\sigma}^2_{D_2}$ *equals* $y_i - \hat{\beta}^{(3)}_{i0}$, *and* $\hat{\epsilon}_i$ *in* $\hat{\sigma}^2_{D_3}$ *equals* $y_i - \hat{\beta}^{(5)}_{i0}$.

It is well known that the LPE is equivalent to a kernel estimator. Fig. 2 shows the implied higher-order kernels in $\hat{\sigma}^2_{D_2}$ and $\hat{\sigma}^2_{D_3}$. So essentially, the smaller biases in $\hat{\sigma}^2_{D_2}$ and $\hat{\sigma}^2_{D_3}$ are attributed to the usage of higher-order kernels in estimating $m(x_i)$.

Remark 3. All of our three variance estimators $\hat{\sigma}^2_{D_q}$ $(q = 1, 2, 3)$ achieve the asymptotically optimal MSE. The dominating term of their asymptotic variances is $2\sigma^4/n$, and high-order terms are $5.16\sigma^4/n^{3/2}$, $7.18\sigma^4/n^{3/2}$, $8.70\sigma^4/n^{3/2}$ respectively, which increase approximately at a linear rate. The bias orders are $O(k^4/n^4)$, $O(k^8/n^8)$ and $O(k^{12}/n^{12})$ respectively, which decrease at an exponential rate. In other words, as more approximation terms of $m(\cdot)$ are included, variance increases slowly whereas bias decreases rapidly. So in practice we can sacrifice variance a little bit to exchange for a great reduction in bias, especially when the mean function has immense oscillation and the sample size is small. Also note that the optimal k values are $0.65n^{1/2}$, $0.90n^{1/2}$ and $1.08n^{1/2}$ respectively, which are increasing. This is reasonable because more observations are required when a higher-order Taylor expansion is applied.

Remark 4. As to the choice of k, neither of Rice (1984), Gasser et al. (1986) and Hall et al. (1990) conducts theoretical analysis. Tong and Wang (2005) provide a theoretically optimal k value equal to $(14n)^{1/2}$, but this k value is too big to use in practical applications. They alternatively suggest using $k = n^{1/2}$ while this value of k is almost the same as our theoretical results. As for the difference sequence $\{d_j\}$, all estimators except ours neglect the effect of the mean function on σ^2 estimation, although Buckley and Eagleson (1989), Seifert et al. (1993), and Dette et al. (1998) realized the importance of bias-correction; details are relegated to the discussion section.

5. Determining the optimal order of Taylor expansion in finite samples

Under different smoothness assumptions, we obtain three different variance estimators $\hat{\sigma}^2_{D_q}$ $(q = 1, 2, 3)$. Which one should be used in finite samples? It is still an open problem, see Seifert et al. (1993). The order of Taylor expansion mainly determines the bias, so the question can be reformulated as when introducing an additional correction term will reduce the bias more than increase the variance. As a result, a natural criterion to determine the order of Taylor expansion is to use the MSE. We must emphasize here that order selection is a finite-sample problem. In large samples, the bias is always dominated by the variance, while in finite samples the bias may be nonnegligible because the constants L_s, $s = 2q$, $q = 1, 2, 3$, in the bias formula may become significant.

To compare the MSEs of the three estimators, we must estimate the nuisance parameters in the MSE formulas. If k takes the optimal value k^*_q, and σ^2 takes the corresponding estimator $\hat{\sigma}^2_{D_q}$, then the only nuisance parameter is L_s. We

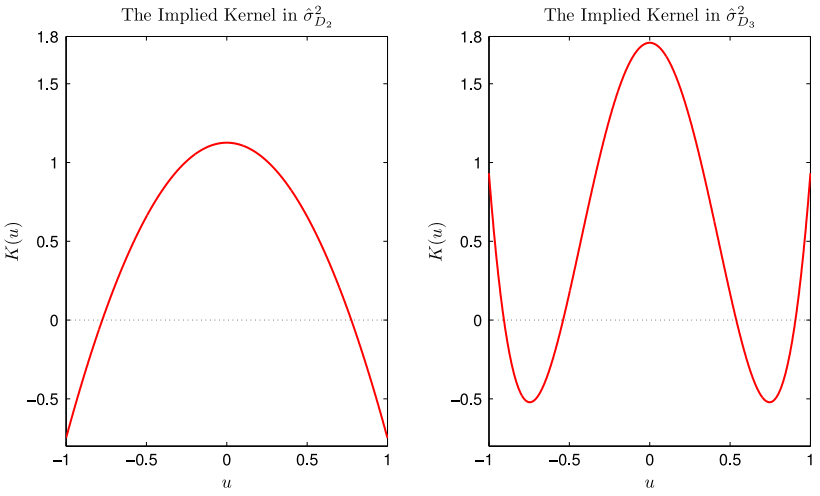

Fig. 2. Equivalent higher-order kernels in the estimation of $\hat{\epsilon}_i$ in $\hat{\sigma}^2_{D_2}$ and $\hat{\sigma}^2_{D_3}$.

describe how to estimate L_s below. First fit a polynomial of order $(s+2)$ globally to $m(\cdot)$, leading to the parametric fit $\tilde{m}_{s+2}(x) = \tilde{\alpha}_0 + \tilde{\alpha}_1 x + \cdots + \tilde{\alpha}_{s+2}x^{s+2}$, and then estimate L_s, by

$$\hat{L}_s = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(s+2)!}{2!}\tilde{\alpha}_{s+2}\left(\frac{i}{n}\right)^2 + (s+1)!\tilde{\alpha}_{s+1}\left(\frac{i}{n}\right) + s!\tilde{\alpha}_s\right]^2.$$

$\tilde{m}^{(s)}_{s+2}(x) = (s+2)!/2!\tilde{\alpha}_{s+2}x^2 + (s+1)!\tilde{\alpha}_{s+1}x + s!\tilde{\alpha}_s$ takes a quadratic form, allowing for a certain flexibility in estimating the curvature. A similar rule-of-thumb procedure is used in the bandwidth selection of local polynomial regression, see Section 4.2 of Fan and Gijbels (1996). Finally, we pick $q^* \in \{1, 2, 3\}$ that minimizes

$$\begin{aligned}
&\frac{2\hat{\sigma}^4_{D_1}}{n} + \frac{4\hat{\sigma}^4_{D_1}k^*_1}{n^2} + \frac{5\hat{\sigma}^4_{D_1}}{3nk^*_1} + \frac{\hat{L}^2_2}{36^2}\frac{k^{*8}_1}{n^8} \quad \text{with } k^*_1 = \lfloor 0.65n^{1/2}\rfloor,\\
&\frac{2\hat{\sigma}^4_{D_2}}{n} + \frac{4\hat{\sigma}^4_{D_2}k^*_2}{n^2} + \frac{3.22\hat{\sigma}^4_{D_2}}{nk^*_2} + \frac{\hat{L}^2_4 k^{*16}_2}{280^4 n^{16}} \quad \text{with } k^*_2 = \lfloor 0.90n^{1/2}\rfloor,\\
&\frac{2\hat{\sigma}^4_{D_3}}{n} + \frac{4\hat{\sigma}^4_{D_3}k^*_3}{n^2} + \frac{4.70\hat{\sigma}^4_{D_3}}{nk^*_3} + \frac{\hat{L}^2_6 k^{*24}_3}{33\,264^4 n^{24}} \quad \text{with } k^*_3 = \lfloor 1.08n^{1/2}\rfloor.
\end{aligned}$$

In practice, we can monitor $\frac{\hat{L}_2}{36}\frac{k^{*4}_1}{n^4}\Big/\left(\frac{2\hat{\sigma}^4_{D_1}}{n}\right)^{1/2}$, $\frac{\hat{L}_4 k^{*8}_2}{280^2 n^8}\Big/\left(\frac{2\hat{\sigma}^4_{D_2}}{n}\right)^{1/2}$, $\frac{\hat{L}_6 k^{*12}_3}{33\,264^2 n^{12}}\Big/\left(\frac{2\hat{\sigma}^4_{D_3}}{n}\right)^{1/2}$ and $\frac{\hat{L}_2}{36}\frac{k^{*4}_1}{n^4}\Big/\left(\frac{8\hat{\sigma}^4_{D_1}k^*_1}{n^2}\right)^{1/2}$, $\frac{\hat{L}_4 k^{*8}_2}{280^2 n^8}\Big/\left(\frac{8\hat{\sigma}^4_{D_2}k^*_2}{n^2}\right)^{1/2}$, $\frac{\hat{L}_6 k^{*12}_3}{33\,264^2 n^{12}}\Big/\left(\frac{8\hat{\sigma}^4_{D_3}k^*_3}{n^2}\right)^{1/2}$ to check the ratio of bias to the dominating variance term and higher order terms, where $8\hat{\sigma}^4_{D_q}k^*_q/n^2$ rather than $4\hat{\sigma}^4_{D_q}k^*_q/n^2$ appears in the denominator because the higher order variance is double of $4\hat{\sigma}^4_{D_q}k^*_q/n^2$ given the optimal k^*_q. Although we can check even higher order expansions in principle, orders higher than 6 seem unattractive in practice (see Table 6).

6. Simulations and application

In this section, we conduct some simulations to compare the performance of the estimators of Rice, Gasser et al., Hall et al., Tong and Wang, and ours. Similar to Seifert et al. (1993), Dette et al. (1998) and Tong and Wang (2005), we specify $m(x)$ as the periodic function $A\sin(2\pi fx)$ and $\epsilon_i \sim N(0, \sigma^2)$. We consider a few specifications of the amplitude A, the frequency f and the error standard deviation σ: $A = 1100, f = 0.5, 1, 2$, and $\sigma = 0.1, 0.5, 2$. As to the sample size n, we consider $n = 50, 200, 1000$, corresponding to small, moderate and large samples, respectively.

The smoothing parameter k is a key input in all estimators. We use k^*_q in $\hat{\sigma}^2_{D_q}$ $(q = 1, 2, 3)$, set $k = 2$ in $\hat{\sigma}^2_{HKT}$, i.e.,

$$\hat{\sigma}^2_{HKT} = \frac{1}{n-2}\sum_{i=2}^{n-1}(0.809Y_{i-1} - 0.5Y_i - 0.309Y_{i+1})^2,$$

Table 1
Relative MSEs of various estimators for $n = 1000$.

A	f	σ	$\hat{\sigma}^2_R$	$\hat{\sigma}^2_{GSJ}$	$\hat{\sigma}^2_{HKT}$	$\hat{\sigma}^2_{TW}$	$\hat{\sigma}^2_{D_1}$	$\hat{\sigma}^2_{D_2}$	$\hat{\sigma}^2_{D_3}$
1	0.5	0.1	1.50	1.95	1.25	1.03	1.07	1.11	1.13
		0.5	1.53	1.99	1.26	1.06	1.11	1.14	1.16
		2	1.51	1.96	1.25	1.05	1.10	1.13	1.17
	1	0.1	1.49	1.94	1.23	1.03	1.08	1.11	1.14
		0.5	1.51	1.96	1.26	1.06	1.09	1.12	1.15
		2	1.53	1.99	1.26	1.05	1.08	1.12	1.15
	2	0.1	1.53	1.98	1.31	1.17	1.14	1.15	1.18
		0.5	1.50	1.94	1.26	1.03	1.07	1.11	1.13
		2	1.48	1.91	1.25	1.04	1.09	1.12	1.15
100	0.5	0.1	3.04E3	1.92	1.89E4	1.72E4	4.85E1	1.10	1.13
		0.5	6.50	1.94	3.18E1	2.89E1	1.16	1.11	1.14
		2	1.50	1.92	1.37	1.15	1.07	1.10	1.12
	1	0.1	4.86E4	1.97	3.03E5	3.27E5	1.21E4	1.13	1.16
		0.5	7.91E1	1.93	4.86E2	5.23E2	2.04E1	1.11	1.14
		2	1.79	1.93	3.13	3.11	1.16	1.13	1.15
	2	0.1	7.78E5	2.00	4.85E6	9.25E6	3.06E6	1.12	1.15
		0.5	1.25E3	1.93	7.76E3	1.48E4	4.90E2	1.10	1.13
		2	6.37	1.92	3.16E1	5.91E1	2.03E1	1.12	1.14

Table 2
Relative MSEs of various estimators for $n = 200$.

A	f	σ	$\hat{\sigma}^2_R$	$\hat{\sigma}^2_{GSJ}$	$\hat{\sigma}^2_{HKT}$	$\hat{\sigma}^2_{TW}$	$\hat{\sigma}^2_{D_1}$	$\hat{\sigma}^2_{D_2}$	$\hat{\sigma}^2_{D_3}$
1	0.5	0.1	1.49	1.93	1.28	1.09	1.18	1.29	1.39
		0.5	1.54	2.01	1.29	1.09	1.21	1.32	1.41
		2	1.521	1.99	1.26	1.09	1.22	1.32	1.41
	1	0.1	1.58	1.95	1.67	1.29	1.23	1.29	1.37
		0.5	1.53	1.98	1.29	1.10	1.23	1.32	1.41
		2	1.50	1.95	1.25	1.10	1.22	1.31	1.40
	2	0.1	2.50	1.96	7.24	7.72	9.38	1.28	1.35
		0.5	1.47	1.91	1.25	1.10	1.22	1.30	1.39
		2	1.53	1.99	1.27	1.09	1.21	1.32	1.41
100	0.5	0.1	3.77E5	1.99	2.33E6	7.91E5	1.41E4	1.31	1.39
		0.5	6.05E2	1.95	3.73E3	1.27E3	2.39E1	1.31	1.39
		2	3.88	1.95	1.59E1	6.29	1.28	1.27	1.36
	1	0.1	6.03E6	2.60	3.72E7	1.70E7	3.52E6	1.31	1.38
		0.5	9.65E3	1.94	5.96E4	2.73E4	5.64E3	1.28	1.36
		2	3.94E1	2.03	2.34E2	1.08E2	2.33E1	1.30	1.39
	2	0.1	9.64E7	1.74E2	5.95E8	6.49E8	8.15E8	3.54E2	1.37
		0.5	1.54E5	2.29	9.52E5	1.04E6	1.30E6	1.85	1.37
		2	6.05E2	1.96	3.72E3	4.06E3	5.10E3	1.30	1.38

and choose $k = n^{1/3}$ for small n and $n^{1/2}$ for moderate and large n in Tong and Wang's estimator as suggested in their simulation studies.

Tables 1–3 summarize our simulation results for $n = 1000, 200, 50$ based on 1000 repetitions, where $Ea = 10^a$, $a = 1, 2, \ldots$, is a display in scientific notation. To compare with the ideal estimator $n^{-1}\sum_{i=1}^{n}\epsilon_i^2$, we report the relative MSE of all estimators. The relative MSE is a ratio between MSE in simulation and the asymptotically optimal MSE $2\sigma^4/n$, that is, $n \cdot MSE/(2\sigma^4)$. The smaller (or closer to 1) the relative MSE is, the better the estimator performs. In most cases, the relative MSE is increasing in A and f, and decreasing in σ and n. In general, our estimators perform better than the existing estimators in most settings. First, when A is small and n is large, $MSE(\hat{\sigma}^2_{TW}) \simeq MSE(\hat{\sigma}^2_{D_1}) < MSE(\hat{\sigma}^2_{D_2}) < MSE(\hat{\sigma}^2_{D_3}) < MSE(\hat{\sigma}^2_{HKT}) < MSE(\hat{\sigma}^2_R) < MSE(\hat{\sigma}^2_{GSJ})$; this ranking is roughly maintained when n gets smaller. Second, when A is large and n is large, $\hat{\sigma}^2_R$, $\hat{\sigma}^2_{HKT}$, $\hat{\sigma}^2_{TW}$, and $\hat{\sigma}^2_{D_1}$ are dominated by $\hat{\sigma}^2_{GSJ}$ which is in turn dominated by $\hat{\sigma}^2_{D_2}$ and $\hat{\sigma}^2_{D_3}$. Third, when A is large, n gets smaller, and f gets larger, all estimators lose effectiveness except $\hat{\sigma}^2_{D_3}$. All these results are intuitively understandable.

We next compare the performance of the smoothness-adaptive estimator with $\hat{\sigma}^2_{D_q}$, $q = 1, 2, 3$. We first determine the ideal best estimator in each specification, that is, the minimizer q_{opt} of $AMSE(\hat{\sigma}^2_{D_q})$ with σ and L_s known. It can be shown

Table 3
Relative MSEs of various estimators for $n = 50$.

A	f	σ	$\hat{\sigma}^2_R$	$\hat{\sigma}^2_{GSJ}$	$\hat{\sigma}^2_{HKT}$	$\hat{\sigma}^2_{TW}$	$\hat{\sigma}^2_{D_1}$	$\hat{\sigma}^2_{D_2}$	$\hat{\sigma}^2_{D_3}$
1	0.5	0.1	1.73	1.95	2.74	1.39	1.48	1.73	1.96
		0.5	1.51	1.99	1.28	1.36	1.44	1.70	1.96
		2	1.52	2.00	1.30	1.36	1.44	1.72	1.96
	1	0.1	5.30	2.02	2.36E1	1.60	5.21	1.71	1.96
		0.5	1.55	2.02	1.37	1.38	1.46	1.73	1.97
		2	1.54	2.03	1.31	1.39	1.48	1.76	2.01
	2	0.1	6.08E1	2.04	3.48E2	7.26	5.81E2	1.90	1.98
		0.5	1.59	1.98	1.84	1.35	2.40	1.69	1.97
		2	1.58	2.06	1.35	1.42	1.50	1.79	2.03
100	0.5	0.1	2.34E7	4.78E1	1.39E8	6.61E5	1.71E6	1.74	1.99
		0.5	3.74E4	2.11	2.23E5	1.07E3	2.73E3	1.74	1.99
		2	1.47E2	2.00	8.73E2	5.98	1.21E1	1.73	1.98
	1	0.1	3.73E7	1.16E4	2.22E9	1.44E7	3.78E8	5.95E2	2.00
		0.5	5.97E5	2.06E1	3.55E6	2.30E4	6.05E5	2.67	1.96
		2	2.33	2.093	1.39E4	9.31E1	2.36E3	1.67	1.91
	2	0.1	5.93E9	2.92E6	3.46E10	5.62E8	5.78E10	1.65E7	2.05E2
		0.5	9.48E6	4.67E3	5.54E7	8.99E5	9.24E7	2.64E4	2.28
		2	3.70E4	2.01E1	2.17E5	3.52E3	3.61E5	1.05E2	1.98

that $L_s = \frac{1}{2}A^2(2\pi f)^{2s}$, so q_{opt} minimizes

$$\frac{4\sigma^4 k_1^*}{n^2} + \frac{5\sigma^4}{3nk_1^*} + \frac{A^4(2\pi f)^8}{4\times 36^2}\frac{k_1^{*8}}{n^8} \quad \text{with } k_1^* = \lfloor 0.65n^{1/2}\rfloor,$$

$$\frac{4\sigma^4 k_2^*}{n^2} + \frac{3.22\sigma^4}{nk_2^*} + \frac{A^4(2\pi f)^{16}}{4\times 280^4}\frac{k_2^{*16}}{n^{16}} \quad \text{with } k_2^* = \lfloor 0.90n^{1/2}\rfloor,$$

$$\frac{4\sigma^4 k_3^*}{n^2} + \frac{4.70\sigma^4}{nk_3^*} + \frac{A^4(2\pi f)^{24}}{4\times 33\,264^4}\frac{k_3^{*24}}{n^{24}} \quad \text{with } k_3^* = \lfloor 1.08n^{1/2}\rfloor.$$

Table 4 reports relative $AMSE(\hat{\sigma}^2_{D_q})$. Compared to Tables 1–3, the AMSE is a reasonable approximation of the MSE in finite samples. Table 5 summarizes the q_{opt}. As expected from Table 4, q_{opt} matches the minimizer of the three finite-sample MSEs in Tables 1–3 perfectly. Table 6 summarizes the coincidence rate of the q^* selected as in Section 5 and the ideal q_{opt}, and Table 7 summarizes the relative MSE of the selected estimator. From Table 6, q^* matches q_{opt} quite well except in the case with $n = 1000$ and $A = 1$, but the selected estimator performs very closely to the optimal one in that case. More optimistically, the MSE of the selected estimator is close to the optimal one in all cases, while neither of the three estimators performs uniformly well in all scenarios.

We next apply our estimation method of σ^2 to a real data set, which is named as "Mandible" in R package "lmtest" (Chitty et al., 1993; Royston and Altman, 1994). There are 167 observations on two variables: gestational age in weeks and mandible length in mm. We treat age and length as independent and response variables, respectively. For this data set, the estimators of σ^2 are $\hat{\sigma}^2_R = 5.49$, $\hat{\sigma}^2_{GSJ} = 5.18$, $\hat{\sigma}^2_{HKT} = 5.77$, $\hat{\sigma}^2_{TW} = 5.68$, $\hat{\sigma}^2_{D_1} = 5.37$, $\hat{\sigma}^2_{D_2} = 5.46$, and $\hat{\sigma}^2_{D_3} = 5.20$, respectively. The smoothness-adaptive estimator is $\hat{\sigma}^2_{D_3}$ which performs similarly as $\hat{\sigma}^2_{GSJ}$ because both have better bias control.

Combining all the simulation and application results in this section, we can conclude that our estimators, especially the smoothness-adaptive estimator, is preferable to other estimators in practice.

7. Discussion

We first discuss the difference sequence $\{d_j\}$. One main role of $\{d_j\}$ is to eliminate the bias in estimating ϵ_i. Rice's first-order differenced estimator eliminates the constant term in the Taylor series of the mean function $m(\cdot)$; Gasser et al.'s second-order differenced estimator and Tong and Wang's estimator eliminates both the constant term and the first-order term. On the other hand, Hall et al.'s kth-order differenced estimator does not eliminate higher-order terms, rather, it eliminates the constant term only. Our estimators (especially $\hat{\sigma}^2_{D_2}$ and $\hat{\sigma}^2_{D_3}$) eliminate the terms of even higher orders, so have less biases. As a cost, more restrictive conditions are imposed on the difference sequence $\{d_j\}$ in our estimators. For the previous differenced estimators, the difference sequence $\{d_j\}$ satisfies two conditions in (4). As mentioned in the introduction, the first condition ensures the unbiasedness of the variance estimators, and the second condition removes the constant term. $\{d_j\}$ in our estimators satisfies $p + 2$ other conditions besides the former two,

$$d_j = d_{-j},\ k_1 = k_2 = k, \quad \sum_{j=-k}^{k} d_j \frac{j^\ell}{n^\ell} = 0\ (\ell = 1, \ldots, p),$$

Table 4
Relative AMSEs of $\hat{\sigma}^2_{D_q}$, $q = 1, 2, 3$.

A	f	σ	1000			200			50		
			$\hat{\sigma}^2_{D_1}$	$\hat{\sigma}^2_{D_2}$	$\hat{\sigma}^2_{D_3}$	$\hat{\sigma}^2_{D_1}$	$\hat{\sigma}^2_{D_2}$	$\hat{\sigma}^2_{D_3}$	$\hat{\sigma}^2_{D_1}$	$\hat{\sigma}^2_{D_2}$	$\hat{\sigma}^2_{D_3}$
1	0.5	0.1	1.08	1.11	1.14	1.18	1.25	1.31	1.37	1.51	1.61
		0.5	1.08	1.11	1.14	1.18	1.25	1.31	1.37	1.51	1.61
		2	1.08	1.11	1.14	1.18	1.25	1.31	1.37	1.51	1.61
	1	0.1	1.08	1.11	1.14	1.20	1.25	1.31	2.54	1.51	1.61
		0.5	1.08	1.11	1.14	1.18	1.25	1.31	1.37	1.51	1.61
		2	1.08	1.11	1.14	1.18	1.25	1.31	1.37	1.51	1.61
	2	0.1	1.08	1.11	1.14	5.87	1.25	1.31	3.01E2	1.60	1.61
		0.5	1.08	1.11	1.14	1.19	1.25	1.31	1.85	1.51	1.61
		2	1.08	1.11	1.14	1.18	1.25	1.31	1.37	1.51	1.61
100	0.5	0.1	3.57E1	1.11	1.14	7.15E3	1.25	1.31	4.58E5	1.51	1.61
		0.5	1.14	1.11	1.14	1.26E1	1.25	1.31	7.33E2	1.51	1.61
		2	1.08	1.11	1.14	1.23	1.25	1.31	4.23	1.51	1.61
	1	0.1	8.86E3	1.11	1.14	1.83E6	1.25	1.31	1.17E8	1.32E2	1.61
		0.5	1.53E1	1.11	1.14	2.93E3	1.25	1.31	1.87E5	1.72	1.61
		2	1.14	1.11	1.14	1.26E1	1.25	1.31	7.33E2	1.51	1.61
	2	0.1	2.27E6	1.11	1.14	4.69E8	1.61E2	1.31	3.00E10	8.56E6	9.89E1
		0.5	3.63E3	1.11	1.14	7.50E5	1.51	1.31	4.80E7	1.37E4	1.77
		2	1.53E1	1.11	1.14	2.93E3	1.25	1.31	1.87E5	5.50E1	1.61

Table 5
Ideal best variance estimator.

n	f	0.5			1			2		
	$A \mid \sigma$	0.1	0.5	2	0.1	0.5	2	0.1	0.5	2
50	1	1	1	1	2	1	1	2	2	1
	100	2	2	2	3	2	2	3	3	3
200	1	1	1	1	1	1	1	2	1	1
	100	2	2	1	2	2	2	3	3	2
1000	1	1	1	1	1	1	1	1	1	1
	100	2	2	1	2	2	2	2	2	2

Table 6
Coincidence rate of q^* and q_{opt}.

n	f	0.5			1			2		
	$A \mid \sigma$	0.1	0.5	2	0.1	0.5	2	0.1	0.5	2
50	1	633	712	721	682	680	733	369	581	689
	100	622	703	705	1000	916	649	1000	1000	1000
200	1	780	786	781	568	731	780	989	680	765
	100	986	984	559	977	985	982	1000	443	980
1000	1	0	0	0	0	0	0	0	0	0
	100	1000	1000	0	1000	1000	1000	1000	1000	1000

Table 7
Relative MSE of the smoothness-adaptive estimator.

n	f	0.5			1			2		
	$A \mid \sigma$	0.1	0.5	2	0.1	0.5	2	0.1	0.5	2
50	1	1.55	1.47	1.55	1.68	1.50	1.55	1.76	1.79	1.71
	100	1.82	1.66	1.70	1.99	1.96	1.91	2.05E2	2.52	1.99
200	1	1.28	1.23	1.17	1.34	1.27	1.24	1.26	1.29	1.23
	100	1.34	1.39	1.21	1.28	1.37	1.25	1.49	1.74	1.30
1000	1	1.12	1.09	1.11	1.09	1.12	1.08	1.19	1.13	1.10
	100	1.14	1.18	1.12	1.10	1.18	1.16	1.15	1.12	1.10

where k is a positive integer, and p is some positive odd integer which depends on the order of Taylor expansion. The first two conditions rely on equidistant designing, and the latter p conditions remove the high-order terms of Taylor series of m. From Theorem 5, our estimators can be interpreted as the LPEs. The p conditions are actually the discrete orthogonality conditions in the local polynomial estimation. Due to $d_j = d_{-j}$, this condition automatically holds when ℓ is odd. This

is also why we need only regress on j^ℓ/n^ℓ with ℓ even in our estimation — odd terms are automatically eliminated by the second-order differencing because the design is symmetric around x_i, $k+1 \le i \le n-k$. Based on our simulation experience, when the bias is small, all our three estimators perform similarly. However, when the bias is large, their performances are very different, and bias correction is critical for our estimator to perform satisfactorily in such cases.

In this paper, we correct the bias in the second-order differences to estimate ϵ_i more precisely. A natural question is whether a higher-order difference can be employed. For example, we can use $y_{ij}^{(3)} = (y_{i-2j} - 4y_{i-j} + 6y_i - 4y_{i+j} + y_{i+2j})/6$ to further eliminate $m^{(2)}$ in Taylor series of m before correcting the bias. However, such higher-order differences have at least two disadvantages. First, we will lose $4k$ instead of $2k$ ϵ_i estimates now. Second, there is some overlapping in the data usage of $y_{ij}^{(3)}$, e.g., $y_{ij}^{(3)}$ and $y_{i,2j}^{(3)}$ have common elements y_{i-2j} and y_{i+2j}, which makes the MSE analysis of variance estimators more troublesome.

Our method focuses on the case with equidistant design. For the non-equidistant design or the conditioned random design, the analysis is more messy. To see why, consider the parallel of $\hat\sigma_{D_2}^2$ in the non-equidistant design. Note that

$$m(x_{i+j}) = m(x_i) + m^{(1)}(x_i)(x_{i+j} - x_i) + o\left(x_{i+j} - x_i\right),$$
$$m(x_{i-j}) = m(x_i) - m^{(1)}(x_i)\left(x_i - x_{i-j}\right) + o\left(x_i - x_{i+j}\right).$$

So the parallel of $y_{ij}^{(2)}$ is

$$\begin{aligned} y_{ij}^{(2)} &= \frac{\left(y_{i+j} - y_i\right)\left(x_i - x_{i-j}\right) + \left(y_{i-j} - y_i\right)\left(x_{i+j} - x_i\right)}{x_{i-j} - x_{i+j}} \\ &= \frac{x_{i+j} - x_i}{x_{i-j} - x_{i+j}} y_{i-j} + y_i + \frac{x_i - x_{i-j}}{x_{i-j} - x_{i+j}} y_{i+j} \\ &= -m^{(2)}(x_i)\left[\frac{(x_{i+j} - x_i)^2}{2}\frac{x_i - x_{i-j}}{x_{i+j} - x_{i-j}} + \frac{(x_i - x_{i-j})^2}{2}\frac{x_{i+j} - x_i}{x_{i+j} - x_{i-j}}\right] + o\left(\frac{j^3}{n^3}\right) + \epsilon_i + \varepsilon_{ij}, \end{aligned}$$

where we assume $x_{i+j} - x_i = O(j/n)$ and $\varepsilon_{ij} = \frac{x_{i+j} - x_i}{x_{i-j} - x_{i+j}}\epsilon_{i-j} + \frac{x_i - x_{i-j}}{x_{i-j} - x_{i+j}}\epsilon_{i+j}$. $y_{i1}^{(2)}$ appears in Gasser et al. (1986) and eliminates the bias of order $1/n$. If we want to eliminate the bias of order $1/n^2$, we need to regress $y_{ij}^{(2)}, j = 1, \ldots, k$, on a constant and $\frac{(x_{i+j} - x_i)^2}{2}\frac{x_i - x_{i-j}}{x_{i+j} - x_{i-j}} + \frac{(x_i - x_{i-j})^2}{2}\frac{x_{i+j} - x_i}{x_{i+j} - x_{i-j}}$, and the resulting estimator of ϵ_i is not neat (e.g., the difference sequence depends on i and is not symmetric) although mathematically straightforward. As a result, the MSE of $\hat\sigma_{D_2}^2$ depends on the design of x such that the optimal choice of k is hard to derive. We hope our choice of k in the equidistant design can serve as a benchmark for this more general case. There is also some literature on variance estimation in the random design, e.g., Müller et al. (2003) proposed a weighted differenced estimator, and Du and Schick (2009) extended Müller et al.'s idea to Gasser et al.'s estimator.

To further extend to the non-equidistant design, suppose for a general differenced estimator, say,

$$\hat\sigma^2 = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(\sum_{j=-k}^{k} d_{i,j} y_{i+j}\right)^2,$$

its MSE is $MSE(d)$, where $d = \left\{(d_{i,-k}, \ldots, d_{i,k})\right\}_{i=k+1}^{n-k}$. For Hall et al. (1990)'s estimators, we choose d by

$$\min_d MSE(d) \text{ s.t. } \sum_{j=-k}^{k} d_{i,j} = 0, \qquad \sum_{j=-k}^{k} d_{i,j}^2 = 1,$$

and the resulting d does not depend on i, while to eliminate higher-order biases, the constraints are strengthened to

$$\sum_{j=-k}^{k} d_{i,j} x_{i+j}^\ell = 0, \quad \ell = 0, 1, \ldots, p, \qquad \sum_{j=-k}^{k} d_{i,j}^2 = 1,$$

where p is a positive integer. It is hard to derive the explicit formula for d in terms of x_i (Hall et al., 1990 and Yatchew, unpublished derived the optimal difference sequence by the unit-root method). Also, the resulting d depends on i, i.e., for each i, the difference sequence, $d_{i,j}$, $j = -k, \ldots, k$, is different, which makes the asymptotic analysis extremely messy, while for the equidistant design, $d_{i,j}$ does not depend on i at all. Although the theoretical results are hard to derive, the idea of optimization with respect to difference sequence can be used in practice, and deserves further research.

The main purpose of this paper is to estimate σ^2, so we assume ϵ_i is homoskedastic. The heteroskedastic case is well studied in the literature, see, e.g., Hall and Carroll (1989) and Wang et al. (2008). In this case, the literature concentrates on the estimation of the variance function rather than a constant variance. So the optimal convergence rate of the variance

function and how the mean function estimation affects the variance function estimation are of main concerns. An interesting question is what our estimators are estimating in the case of heterogeneous variance. It turns out that

$$\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\hat{\epsilon}_i^2 \xrightarrow{p} \int \sigma^2(x)dx,$$

i.e., the average variance. We are not sure whether this parameter is an interesting parameter in practice.

This paper concentrates on the univariate x case; some extensions to multivariate x have been developed in the literature. Hall et al. (1991) generalized Hall et al. (1990)'s estimator to bivariate lattice designs and discussed the problem of different configurations in details. Kulasekera and Gallagher (2002) introduced an algorithm to order the design points in $x \in [0, 1]^d$. Spokoiny (2002) and Munk et al. (2005) proposed a residual-based estimator and a differenced estimator respectively for multivariate regression. Fan et al. (2012) used refitted cross-validation method in ultrahigh dimensional linear regression. Dicker (2014) proposed a method of moments in high-dimensional linear regression. Extension of our method to multivariate x is a big challenge and left to further research.

Acknowledgments

The authors are grateful to the associate editor and two anonymous referees for helpful comments which significantly improve the paper.

Appendix A. Proof of Theorem 1

The variance of $\hat{\epsilon}_i$ (Fan and Gijbels, 1996) is $\mathrm{var}(\hat{\epsilon}_i \mid \epsilon_i) = \sigma^2(X_1^{\mathrm{T}}X_1)^{-1}/2$. Owing to $X_1^{\mathrm{T}}X_1 = k$, we have

$$\mathrm{var}(\hat{\epsilon}_i \mid \epsilon_i) = \frac{\sigma^2}{2k}.$$

The bias of $\hat{\epsilon}_i$ is $\mathrm{bias}(\hat{\epsilon}_i \mid \epsilon_i) = (X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}\mathrm{E}(\delta_i)$, where $\delta_i = (y_{i1}^{(2)} - \epsilon_i, \ldots, y_{ik}^{(2)} - \epsilon_i)^{\mathrm{T}} = (\delta_{i1}, \ldots, \delta_{ik})^{\mathrm{T}}$. Visiting the second-order Taylor expansion of $m(x_{i\pm j})$ at x_i, we have

$$y_{ij}^{(2)} = \epsilon_i - \frac{m^{(2)}(x_i)}{2!}\frac{j^2}{n^2} + o\left(\frac{j^2}{n^2}\right) - \frac{\epsilon_{i-j}+\epsilon_{i+j}}{2}.$$

Now $\delta_{ij} = -m^{(2)}(x_i)j^2/(2!n^2) + o(j^2/n^2) - (\epsilon_{i-j}+\epsilon_{i+j})/2$, so we have

$$\mathrm{bias}(\hat{\epsilon}_i \mid \epsilon_i) = -\frac{m^{(2)}(x_i)}{6}\frac{k^2}{n^2} + o\left(\frac{k^2}{n^2}\right).$$

The AMSE is

$$\mathrm{AMSE}(\hat{\epsilon}_i \mid \epsilon_i) = \frac{(m^{(2)}(x_i))^2}{36}\frac{k^4}{n^4} + \frac{1}{2}\frac{\sigma^2}{k}.$$

To minimize the AMSE with respect to k, let

$$\frac{\mathrm{d}[AMSE(\hat{\epsilon}_i \mid \epsilon_i)]}{\mathrm{d}k} = \frac{(m^{(2)}(x_i))^2}{9}\frac{k^3}{n^4} - \frac{1}{2}\frac{\sigma^2}{k^2} = 0.$$

Then the optimal k value is $k_{opt} \approx 1.35(\sigma^2/(m^{(2)}(x_i))^2)^{1/5}n^{4/5}$, and the AMSE of $\hat{\epsilon}_i$ is

$$\mathrm{AMSE}(\hat{\epsilon}_i) \approx 0.46\left(\sigma^8(m^{(2)}(x_i))^2\right)^{1/5}n^{-4/5}.$$

Appendix B. Proof of Theorem 2

$\{y_{ij}^{(2)}\}_{j=1}^k$ are independent given ϵ_i and have the same conditional variance $\sigma^2/2$. By the Lindeberg–Feller Theorem (Serfling, 1980), the asymptotic normality can be obtained if and only if the Lindeberg condition is satisfied.

Let $B_k^2 = k\sigma^2/2$. For any $\delta > 0$, we verify the Lindeberg condition:

$$\frac{\sum_{j=1}^{k}\int_{|x|>\delta B_k} x^2 \frac{1}{\sqrt{\pi}\sigma}\exp^{-\frac{x^2}{\sigma^2}}\mathrm{d}x}{B_k^2} = \frac{\int_{|x|>\delta B_k} x^2 \frac{1}{\sqrt{\pi}\sigma}\exp^{-\frac{x^2}{\sigma^2}}\mathrm{d}x}{\sigma^2/2} \to 0, \quad k \to \infty.$$

So we have the asymptotic normality

$$k^{1/2}\left(\hat{\epsilon}_i - \epsilon_i + \frac{m^{(2)}(x_i)}{6}\frac{k^2}{n^2}\right) \xrightarrow{d} N\left(0, \frac{1}{2}\sigma^2\right)$$

for $k+1 \le i \le n-k$. If $n^{-1}k^{5/4} \to 0$, then the bias disappears. We obtain the asymptotic normality

$$k^{1/2}(\hat{\epsilon}_i - \epsilon_i) \xrightarrow{d} N\left(0, \frac{1}{2}\sigma^2\right)$$

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

Appendix C. Proof of Theorem 3

The bias of $\hat{\sigma}^2_{D_1}$ is $m_n^{\mathrm{T}} D_1 m_n / tr(D_1)$. By Theorem 1, we have

$$m_n^{\mathrm{T}} D_1 m_n = (\tilde{D}_1 m_n)^{\mathrm{T}} \tilde{D}_1 m_n = \sum_{i=k+1}^{n-k} \frac{(m^{(2)}(x_i))^2}{6^2}\frac{k^4}{n^4}.$$

Owing to $tr(D_1) = n - 2k$, the bias of $\hat{\sigma}^2_{D_1}$ is

$$\mathrm{bias}(\hat{\sigma}^2_{D_1}) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} \frac{(m^{(2)}(x_i))^2}{6^2}\frac{k^4}{n^4} = \frac{L_2}{36}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right),$$

where $L_2 = \int_0^1 (m^{(2)}(x))^2 \mathrm{d}x$.

The variance of $\hat{\sigma}^2_{D_1}$ is $\{4\sigma^2 m_n^{\mathrm{T}} D_1^2 m_n + 2\sigma^4 tr(D_1^2)\}/tr(D_1)^2$. Since

$$m_n^{\mathrm{T}} D_1^2 m_n = o\left(\frac{n}{k}\right), \qquad tr(D_1^2) = (n-2k)\left\{1 + \frac{5}{6k} + o\left(\frac{1}{k}\right)\right\},$$

the variance of the $\hat{\sigma}^2_{D_1}$ is

$$\mathrm{var}\left(\hat{\sigma}^2_{D_1}\right) = 2\sigma^4\left(\frac{1}{n} + \frac{2k}{n^2} + \frac{5}{6nk}\right) + o\left(\frac{k}{n^2}\right) + o\left(\frac{1}{nk}\right),$$

and the AMSE of $\hat{\sigma}^2_{D_1}$ is

$$\mathrm{AMSE}\left(\hat{\sigma}^2_{D_1}\right) = \frac{2\sigma^4}{n} + \frac{4\sigma^4 k}{n^2} + \frac{5\sigma^4}{3nk} + \frac{L_2^2}{36^2}\frac{k^8}{n^8}.$$

Minimizing the AMSE with respect to k, we have the asymptotically optimal bandwidth $k_1^* = (5n/12)^{1/2}$.

Appendix D. Proof of Theorem 4

Lemma 1. *Consider the form*

$$\Lambda = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{j-k} z_j z_k$$

in which the z_j's are identically distributed with zero mean. This will tend to normality in distribution as $n \to \infty$ if the following conditions are fulfilled:

(1) $\mathrm{E} \mid z_j \mid^{4+2\delta}$ *is finite for some* δ *in* $(0, 1)$;
(2) $\sum_{j=-\infty}^{\infty} a_j^2$ *is finite.*

Lemma 1 is borrowed from Whittle (1964). Note that $\hat{\sigma}^2_{D_1} = m_n^{\mathrm{T}} D_1 m_n / tr(D_1) + 2m_n^{\mathrm{T}} D_1 \epsilon / tr(D_1) + \epsilon^{\mathrm{T}} D_1 \epsilon / tr(D_1)$, where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$. Let $z_j = \epsilon_j$ for $j = 1, \ldots, n$; then the z_j's are independent normally distributed with zero mean and variance σ^2. So condition (1) is satisfied since $E \mid z_j \mid^5$ is finite for $\delta = 1/2$. Condition (2) is satisfied due to $\sum_{j=-\infty}^{\infty} a_j^2 \to 1 + 5/(6k) < 2$. By Lemma 1,

$$n^{1/2}\left(\epsilon^{\mathrm{T}} D_1 \epsilon / tr(D_1) - \sigma^2\right) \xrightarrow{d} N(0, 2\sigma^4).$$

By the proof of Theorem 3, $m_n^{\mathrm{T}} D_1 m_n / tr(D_1) = \frac{L_2}{36}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right)$, and $m_n^{\mathrm{T}} D_1 \epsilon / tr(D_1) = o\left(n^{1/2}/(k^{1/2}(n-2k))\right)$. So if $k \to \infty$ as $n \to \infty$ such that $k/n \to 0$, then

$$n^{1/2}\left(\hat{\sigma}_{D_1}^2 - \sigma^2 - \frac{L_2 k^4}{36 n^4}\right) \xrightarrow{d} N(0, 2\sigma^4).$$

If $k \to \infty$ as $n \to \infty$ such that $n^{-1}k^{8/7} \to 0$, then

$$n^{1/2}\left(\hat{\sigma}_{D_1}^2 - \sigma^2\right) \xrightarrow{d} N(0, 2\sigma^4).$$

Appendix E. Proof of Theorem 5

The proof involves some tedious but straightforward calculations. Specifically, we need to show that

$$e_1^{(2)}\left(\tilde{X}_1^{\mathrm{T}}\tilde{X}_1\right)^{-1}\tilde{X}_1^{\mathrm{T}} = (\underbrace{0,\ldots,0}_{k}, 1, \underbrace{0,\ldots,0}_{k}) - (X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}G,$$

$$e_1^{(4)}\left(\tilde{X}_3^{\mathrm{T}}\tilde{X}_3\right)^{-1}\tilde{X}_3^{\mathrm{T}} = (\underbrace{0,\ldots,0}_{k}, 1, \underbrace{0,\ldots,0}_{k}) - e_1^{(2)}(X_2^{\mathrm{T}}X_2)^{-1}X_2^{\mathrm{T}}G,$$

$$e_1^{(6)}\left(\tilde{X}_5^{\mathrm{T}}\tilde{X}_5\right)^{-1}\tilde{X}_5^{\mathrm{T}} = (\underbrace{0,\ldots,0}_{k}, 1, \underbrace{0,\ldots,0}_{k}) - e_1^{(3)}(X_3^{\mathrm{T}}X_3)^{-1}X_3^{\mathrm{T}}G.$$

References

Anderson, T.W., 1971. The Statistical Analysis of Time Series. Wiley, New York.
Buckley, M.J., Eagleson, G.K., 1989. A graphical method for estimating the residual variance in nonparametric regression. Biometrika 76, 203–210.
Buckley, M.J., Eagleson, G.K., Silverman, B.W., 1988. The estimation of residual variance in nonparametric regression. Biometrika 75, 189–199.
Carter, C.K., Eagleson, G.K., 1992. A comparison of variance estimators in nonparametric regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 54, 773–780.
Carter, C.K., Eagleson, G.K., Silverman, B.W., 1992. A comparison of the reinsch and speckman splines. Biometrika 71, 81–91.
Chitty, L.S., Campbell, S., Altman, D.G., 1993. Measurement of the fetal mandible-feasibility and construction of a centile chart. Prenat. Diagn. 13, 749–756.
Dette, H., Munk, A., Wagner, T., 1998. Estimating the variance in nonparametric regression–what is a reasonable choice? J. Amer. Statist. Assoc. 60, 751–764.
Dicker, L.H., 2014. Variance estimation in high-dimensional linear models. Biometrika 101 (1), 1–16.
Du, J., Schick, A., 2009. A covariate-matched estimator of the error variance in nonparametric regression. J. Nonparametr. Stat. 21 (3), 263–285.
Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Cahpman & Hall, London.
Fan, J., Guo, S., Hao, N., 2012. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 74, 37–65.
Gasser, T., Sroka, L., Jennen-Sternmetz, C., 1986. Residual variance and residual pattern in nonlinear regression. Biometrika 73, 625–633.
Hall, P., Carroll, R.J., 1989. Variance function estimation in regression: the effect of estimating the mean. J. R. Stat. Soc. Ser. B Stat. Methodol. 51, 3–14.
Hall, P., Kay, J.W., Titterington, D.M., 1990. Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika 77, 521–528.
Hall, P., Kay, J.W., Titterington, D.M., 1991. On estimation of noise variance in two-dimensional signal processing. Adv. Appl. Probab. 23, 476–495.
Hall, P., Marron, J.S., 1990. On variance estimation in nonparametric regression. Biometrika 77, 415–419.
Kulasekera, K.B., Gallagher, D.M., 2002. Variance estimation in nonparametric multiple regression. Comm. Statist. A 31, 1373–1383.
Müller, U.U., Schick, A., Wefelmeyer, W., 2003. Estimating the error variance in nonparametri regression by a covariate-matched u-statistic. Statistics 37, 179–188.
Müller, H.G., Stadtmüller, U., 1987. Estimation of heteroscedasticity in regression analysis. Ann. Statist. 15, 610–635.
Munk, A., Bissantz, N., Wagner, T., Freitag, G., 2005. On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. Aust. N. Z. J. Stat. 44, 479–488.
Neumann, M.H., 1994. Fully data-driven nonparametric variance estimators. Statistics 25, 189–212.
Rice, J.A., 1984. Bandwidth choice for nonparametric regression. Ann. Statist. 12, 1215–1230.
Royston, P., Altman, D.G., 1994. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Appl. Stat. 43, 429–453.
Seifert, B., Gasser, T., Wolf, A., 1993. Nonparametric estimation of residual variance revisited. Biometrika 80, 373–383.
Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. John Wiley & Sons, New York.
Spokoiny, V., 2002. Variance estimation for high-dimensional regression models. J. Multivariate Anal. 82, 111–133.
Tong, T., Ma, Y., Wang, Y., 2013. Optimal variance estimation without estimating the mean function. Bernoulli 19 (5A), 1839–1854.
Tong, T., Wang, Y., 2005. Estimating residual variance in nonparametric regression using least squares. Biometrika 92 (4), 821–830.
von Neumann, J., 1941. Distribution of the ratio of the mean square successive difference to the variance. Ann. Math. Statist. 12 (4), 367–395.
Wahba, G., 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 40, 364–372.
Wang, L., Brown, L.D., Cai, T., Levine, M., 2008. Effect of mean on variance function estimation in nonparametric regression. Ann. Statist. 36, 646–664.
Wang, W.W., Lin, L., 2015. Derivative estimation based on difference sequence via locally weighted least squares regression. J. Mach. Learn. Res. 16, 2617–2641.
Wang, W.W., Lin, L., Yu, L., 2016. Optimal variance estimation based on lagged second-order difference in nonparametric regression. Comput. Statist. 18, http://dx.doi.org/10.1007/s00180-016-0666-2.
Whittle, P., 1964. On the convergence to normality of quadratic forms in independent variables. Theory Probab. Appl. 9, 103–108.
Yatchew, A., 1999. Differencing method in nonparametric regression: Simple techniques for the applied econometrician. Unpublished.