Comput Stat (2018) 33:863–885
https://doi.org/10.1007/s00180-017-0786-3

ORIGINAL PAPER

Optimal difference-based estimation for partially linear models

Yuejin Zhou[1,2] · **Yebin Cheng**[3] · **Wenlin Dai**[4] · **Tiejun Tong**[5]

Received: 29 May 2017 / Accepted: 8 December 2017 / Published online: 16 December 2017

Abstract Difference-based methods have attracted increasing attention for analyzing partially linear models in the recent literature. In this paper, we first propose to solve the optimal sequence selection problem in difference-based estimation for the linear component. To achieve the goal, a family of new sequences and a cross-validation method for selecting the adaptive sequence are proposed. We demonstrate that the existing sequences are only extreme cases in the proposed family. Secondly, we propose a new estimator for the residual variance by fitting a linear regression method to some difference-based estimators. Our proposed estimator achieves the asymptotic optimal rate of mean squared error. Simulation studies also demonstrate that our proposed estimator performs better than the existing estimator, especially when the sample size is small and the nonparametric function is rough.

✉ Yuejin Zhou
yjchow@163.com

Yebin Cheng
chengyebin@hotmail.com

Wenlin Dai
wenlin.dai@kaust.edu.sa

Tiejun Tong
tongt@hkbu.edu.hk

1 School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China

2 School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China

3 Glorious Sun School of Business and Management, Donghua University, Shanghai, China

4 CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

5 Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong, China

Keywords Asymptotic normality · Difference-based method · Difference sequence · Least squares estimator · Partially linear model

1 Introduction

Consider the partially linear model

$$Y_i = X_i^T \beta + f(Z_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where Y_i are responses, $\beta = (\beta_1, \ldots, \beta_p)^T$ is an unknown p dimensional vector of parameters, $f(\cdot)$ is an unknown nonparametric function, and ε_i are independent and identically distributed (i.i.d.) random errors with mean zero and variance σ^2. In addition, X_i and Z_i are design points for parametric and nonparametric components, respectively. Model (1) has been extensively studied in the literature with popular methods including, for example, kernel smoothing methods, spline smoothing methods, penalized least squares methods, and profile likelihood methods (Chen and Shiau 1991; Cuzick 1992; Spechman 1988; Severini and Wong 1992; Eubank et al. 1998; Hardle et al. 2000; Fan and Huang 2005).

Recently, Wang et al. (2011) proposed a difference-based method for analyzing model (1). They first applied the classical difference-based method to eliminate the nonparametric component f, and then applied the standard linear regression to estimate the linear component β and the residual variance σ^2. Let $d = \{d_0, \ldots, d_m\}$ be a sequence of real numbers such that

$$\sum_{j=0}^{m} d_j = 0 \quad \text{and} \quad \sum_{j=0}^{m} d_j^2 = 1, \tag{2}$$

where $d_0 d_m \neq 0$ and $m > 0$ is referred to as the order of sequence. To eliminate the nonparametric component, the authors applied the following linear transformation to model (1),

$$\tilde{Y}_i = \tilde{X}_i^T \beta + \delta_i + \tilde{\varepsilon}_i, \quad i = 1, \ldots, n - m, \tag{3}$$

where $\tilde{Y}_i = \sum_{j=0}^{m} d_j Y_{i+j}$, $\tilde{X}_i = \sum_{j=0}^{m} d_j X_{i+j}$, $\delta_i = \sum_{j=0}^{m} d_j f(Z_{i+j})$, and $\tilde{\varepsilon}_i = \sum_{j=0}^{m} d_j \varepsilon_{i+j}$. Under the constraint (2), the term δ_i is asymptotically negligible under some mild conditions so that the transformed model (3) reduces to nearly a classical linear regression model. In view of this, the authors then proposed to estimate the linear component by

$$\hat{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}, \tag{4}$$

where $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_{n-m})^T$ and $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_{n-m})^T$. Further, by the residual sum of squares the residual variance is estimated as

$$\hat{\sigma}_{\mathrm{W}}^2 = \frac{1}{n-m-p}\sum_{i=1}^{n-m}\left(\tilde{Y}_i - \tilde{X}_i^T\hat{\beta}\right)^2 = \frac{1}{n-m-p}\tilde{Y}^T\Delta\tilde{Y}, \tag{5}$$

where $\Delta = I - \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T$ with I being the unit matrix of size $(n-m)\times(n-m)$. We refer to $\hat{\sigma}_{\mathrm{W}}^2$ as a partially difference-based variance estimator. It is noteworthy that other works in difference-based estimation for model (1) are also available, see for example, Akdeniz and Duran (2013), Eubank et al. (1998), He et al. (2014), Hu et al. (2016), Levine (2015), Liu and Zhao (2012), Lokshin (2006), Tabakan (2013), Yatchew (1997), Zhao and You (2011), and the references therein.

To implement (4), one needs an appropriate choice of sequence d under the constraint (2). As pointed out in Dette et al. (1998), the choice of the difference sequence can be rather delicate in practice and requires further attention. In this paper, we consider the following family for the optimal choice of sequence:

$$d_j = \begin{cases} \{m/(m+1)\}^{1/2} & j = k, \\ -\{m(m+1)\}^{-1/2} & j = 0,\ldots,k-1,k+1,\ldots,m, \end{cases} \tag{6}$$

where the integer k is a tuning parameter and represents the location where the spike is taken. All other d_j values are equally distributed so that $\sum_{j=0}^{m} d_j = 0$. Due to the symmetry, it is sufficient to consider the parameter space as $0 \le k \le m/2$ if m is even and $0 \le k \le (m-1)/2$ if m is odd. To estimate β optimally, Wang et al. (2011) suggested to use the spike sequence with the spike at the boundary of the sequence. Note that, though optimal in the asymptotic sense, the spike sequence may not perform well for small sample sizes. This motivates us to develop an adaptive sequence for the difference-based estimation and demonstrate its superiority over the existing one.

Needless to say, an accurate estimate of σ^2 is also important and crucial for partially linear models. It is needed, for instance, in constructing confidence intervals, in checking goodness of fit and outliers, and in many other applications. We note, however, that the partially residual-based variance estimator $\hat{\sigma}_{\mathrm{W}}^2$ in (5) does not achieve the asymptotic optimal rate of mean squared error (MSE). Specifically, we will show in Sect. 3 that

$$\mathrm{MSE}\left(\hat{\sigma}_{\mathrm{W}}^2\right) > \frac{1}{n}\mathrm{Var}(\varepsilon^2) + o\left(\frac{1}{n}\right). \tag{7}$$

This motivates us to also propose a new estimator for σ^2 in partially linear models. The proposed estimator is optimal in the sense that its asymptotic MSE is equal to $\mathrm{Var}(\varepsilon^2)/n$.

The rest of the paper is organized as follows. In Sect. 2, we study the appropriate choice of sequence for the estimator $\hat{\beta}$ in (4). By drawing connections between the existing sequences and the proposed family, we propose a novel adaptive sequence by the cross-validation method. In Sect. 3, we propose a new estimator of the variance σ^2 and show that the MSE of the proposed estimator achieves the asymptotic optimal rate of MSE. In Sect. 4, we conduct two simulation studies to assess the proposed optimal estimators, where the first one is for the adaptive sequence and the other is

for the variance estimation. Simulation studies support our findings that the proposed methods perform better than the existing competitors. We conclude the paper in Sect. 5 and provide the technical proofs in Sect. 6.

2 Optimal choice of sequence

In this section, we investigate the appropriate choice of sequence in difference-based estimation for partially linear models. To achieve this, we first derive the approximate MSE of the estimator $\hat{\beta}$ in (4). Let $\delta = (\delta_1, \ldots, \delta_{n-m})^T$ and $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots, \tilde{\varepsilon}_{n-m})^T$. Then in matrix form, model (3) can be written as

$$\tilde{Y} = \tilde{X}\beta + \delta + \tilde{\varepsilon}. \tag{8}$$

Assume also that X_i are i.i.d. random vectors with mean vector μ and covariance matrix Σ_X, and Z_i are equally spaced design points with $Z_i = i/n$ for $i = 1, \ldots, n$. For ease of notation, let $A = (D^T\delta)(D^T\delta)^T = (a_{ij})_{n\times n}$, $B = \mu\mu^T$, and D is an $(n-m) \times n$ matrix of form

$$D = \begin{pmatrix} d_0 & d_1 & \cdots & d_m & 0 & \cdots & 0 \\ 0 & d_0 & d_1 & \cdots & d_m & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \vdots \\ 0 & \cdots & 0 & d_0 & d_1 & \cdots & d_m \end{pmatrix}.$$

Theorem 1 *Assume that f has a bounded first derivative. The approximate MSE of $\hat{\beta}$ in (4) is given as*

$$\mathrm{MSE}\left(\hat{\beta}\right) \sim \frac{1}{n^2}\left\{\mathrm{tr}(A)\Sigma_X^{-1} + \sum_{i,j=1}^{n} a_{ij}\Sigma_X^{-1}B\Sigma_X^{-1}\right\} + \frac{1}{n}\left(1 + 2\sum_{l=1}^{m} c_l^2\right)\sigma^2\Sigma_X^{-1}, \tag{9}$$

where $c_l = \sum_{j=0}^{m-l} d_j d_{j+l}$ for $l = 1, \ldots, m$ and '$\sim$' is defined the same way as in Stirling's approximation.

The Proof of Theorem 1 is given in Sect. 6.1. It shows that the MSE of $\hat{\beta}$ consists of two distinct components, where one is associated with the nonparametric component and the other one with the random errors. Note that the sequence family (5) solely depends on the location k of the spike. To be specific, we represent the sequence d as a function of k, i.e., $d(k) = (d_0(k), d_1(k) \ldots, d_m(k))$ where

$$d_j(k) = \begin{cases} \{m/(m+1)\}^{1/2} & j = k, \\ -\{m(m+1)\}^{-1/2} & j = 0, \ldots, k-1, k+1, \ldots, m. \end{cases} \tag{10}$$

Then to find the optimal sequence is equivalent to finding the optimal k value.

With the sequence $d(k)$ in (10), the approximate MSE of $\hat{\beta}$ derived in (9), can now be represented as a function of k. Specifically, we have

$$\mathrm{MSE}\left(\hat{\beta}\right) \sim (V_1 + V_2)\Sigma_X^{-1}, \tag{11}$$

where

$$V_1 = \frac{(m+1)(m-2k)^2}{4mn^4}\mathrm{tr}\left(\left(D^T f'\right)\left(D^T f'\right)^T\right),$$
$$V_2 = \frac{m\left(3m^2+5m+1\right)+12k(k+1)}{3nm^2(m+1)}\sigma^2,$$

and $f' = (f'(Z_{1+k}), \ldots, f'(Z_{n-m+k}))^T$ are the first-order derivatives of f.

2.1 The sequence in Wang et al. (2011)

Note that $V_1 = O(n^{-4})$ and $V_2 = O(n^{-1})$ for any fixed m. Therefore, from an asymptotic point of view, the contribution of the nonparametric component to the MSE is negligible compared to that from the random errors. As a consequence, if we ignore the term V_1 in (11), the approximate $\mathrm{MSE}(\hat{\beta})$ reduces to

$$\frac{m(3m^2+5m+1)+12k(k+1)}{3nm^2(m+1)}\sigma^2\Sigma_X^{-1}. \tag{12}$$

Noting that $k \geq 0$, the minimum value of (12) is achieved at $k = 0$. This results in the sequence suggested in Wang et al. (2011), i.e.,

$$d_j(0) = \begin{cases} \{m/(m+1)\}^{1/2} & j = 0, \\ -\{m(m+1)\}^{-1/2} & j = 1, \ldots, m. \end{cases}$$

We refer to it as the WBC sequence. Note that the WBC sequence takes the spike in the boundary and so is an extreme case of the proposed sequence family. In addition, the WBC sequence is optimal in the asymptotic sense and has been applied in the recent literature, e.g., Liu and Zhao (2012).

2.2 The sequence in Hall et al. (1990)

When the sample size is small, however, the contribution of the nonparametric component, i.e., the term V_1, is often non-negligible. In particular, when f is very rough, V_1 may even dominate the MSE of $\hat{\beta}$. Thus, as an alternative option, one may also consider the sequence that minimizes V_1 only. That is, we are to minimize the quantity

$$\frac{(m+1)(m-2k)^2}{4mn^4}\mathrm{tr}\left(\left(D^T f'\right)\left(D^T f'\right)^T\right)\Sigma_X^{-1}. \tag{13}$$

Note that $\mathrm{tr}((D^T f')(D^T f')^T) > 0$ for any non-constant function f. Therefore, the unique minimum MSE of (13) has to be achieved at $k = m/2$ when m is an even number. Specifically, the resulting sequence is

$$d_j\left(\frac{m}{2}\right) = \begin{cases} \{m/(m+1)\}^{1/2} & j = m/2, \\ -\{m(m+1)\}^{-1/2} & 0 \le j \le \frac{m}{2} - 1 \text{ or } \frac{m}{2} + 1 \le j \le m. \end{cases} \tag{14}$$

Similarly, when m is an odd number, it can be also shown that the minimum MSE is achieved at $k = (m-1)/2$ or $k = (m+1)/2$. The resulting sequence is then

$$d_j\left(\frac{m-1}{2}\right) = \begin{cases} \{m/(m+1)\}^{1/2} & j = (m-1)/2, \\ -\{m(m+1)\}^{-1/2} & 0 \le j \le \frac{m-3}{2} \text{ or } \frac{m+1}{2} \le j \le m. \end{cases} \tag{15}$$

We note that the sequences in (14) and (15) are the same as the spike sequence in Hall et al. (1990). Here we refer to them as the HKT sequence. The HKT sequence takes the spike in the middle and so is another extreme case of the proposed sequence family. We also note that the HKT sequence can be transformed to the sequence used by Gasser et al. (1986) when $m = 2$.

2.3 Adaptive sequence

Recall that the WBC sequence is achieved by minimizing the random errors part and it may only perform well in the asymptotic sense. On the other hand, the HKT sequence is achieved by minimizing the nonparametric component and it may only work for small sample size. As a consequence, neither of them may be the optimal in practice. Especially when the sample size is moderate, V_1 and V_2 can be very comparable so that both of them should be taken in account.

To explore the relationship between V_1 and V_2, we consider the following simple example. Let $n = 100$, $m = 16$, $\sigma^2 = 1$, and $f(Z) = \sin(\omega\pi Z)$ with $\omega = 0,\ 1,\ 2$ and 4, corresponding to the different levels of oscillation. We plot V_1, V_2 and $V_1 + V_2$ as a function of k in Fig. 1, respectively. From the plotted curves, we note that V_1 is a decrease function of k on $[0, m/2]$, and V_2 is an increase function of k on $[0, m/2]$. This coincides with the fact that the HKT sequence takes the spike at $k = m/2$ and the WBC sequence at $k = 0$. Note also that $V_1 + V_2$ is no longer a monotonic function of k when $\omega = 1$, 2 or 4, that is, when f is not a constant function. It is evident that the minimum MSE, i.e., the minimum of $V_1 + V_2$, is located in somewhere between 0 and $m/2$. This demonstrates that the HKT sequence and the WBC sequence are both extreme cases and may not be the optimal in practice.

To find the optimal k value, we propose a data-driven method that controls V_1 and V_2 simultaneously. Specifically, the following cross-validation method is considered. For n pair observations $\{(X_i, Z_i, Y_i), i = 1, \ldots, n\}$, let $\hat{\beta}_{-i}$ be a leave-one-out estimator of β without the ith pair (X_i, Z_i, Y_i). We then choose the optimal $k = k_{\mathrm{CV}}$ that minimizes

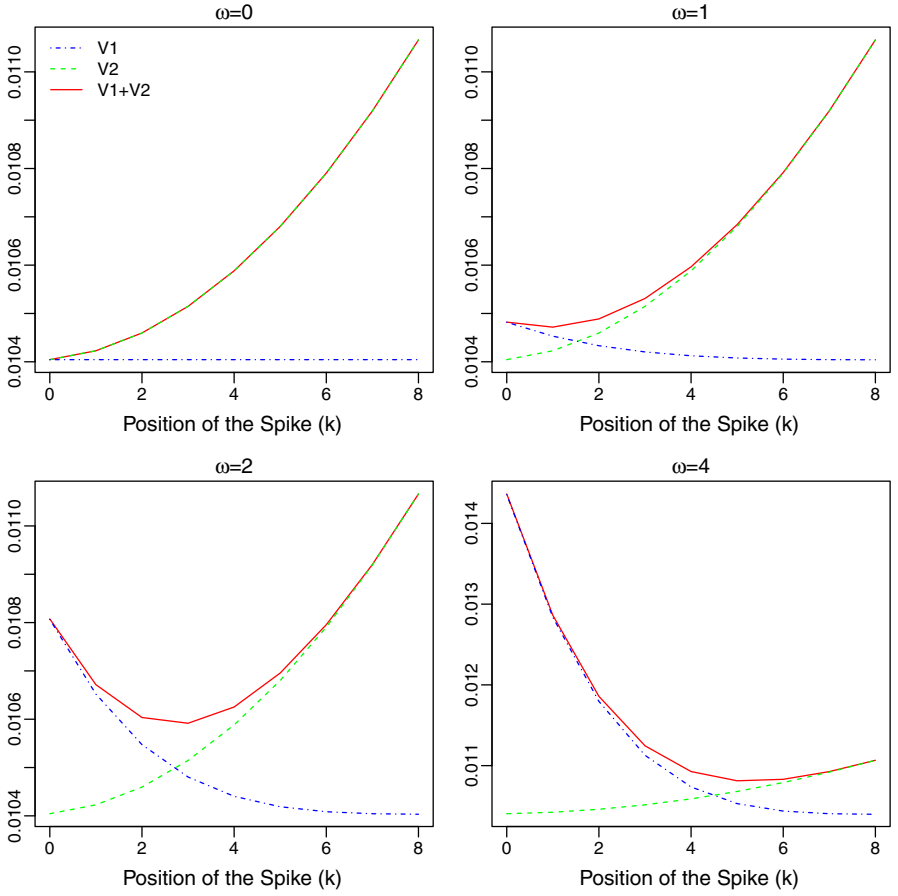

Fig. 1 With $n = 100$, $m = 15$ and $f(z) = \sin(\omega\pi z)$. Solid lines: $V_1 + V_2$; Dot-dashed lines: V_1; Dashed lines: V_2

$$\mathrm{CV}(k) = \sum_{i=1}^{n} \left\| \hat{\beta}(d(k)) - \hat{\beta}_{-i}(d(k)) \right\|^2 \tag{16}$$

within the parameter space of $0 \le k \le m/2$ if m is even or $0 \le k \le (m-1)/2$ if m is odd. Correspondingly, we refer to $d(k_{\mathrm{CV}})$ as the adaptive sequence.

3 Optimal variance estimation

Needless to say, an accurate estimate of σ^2 is also desired in partially linear models. However, as shown in Theorem 2, the estimator $\hat{\sigma}^2_{\mathrm{W}}$ in Wang et al. (2011) does not achieve the asymptotic optimal rate of MSE. In this section, we propose a new estimator of σ^2, derive its theoretical results, and show that it is superior to the existing competitor $\hat{\sigma}^2_{\mathrm{W}}$.

Let $Y_i^* = Y_i - X_i^T\hat{\beta}$ for $i = 1, 2, \ldots, n$. Note that, by Wang et al. (2011), the estimator $\hat{\beta}$ is $\sqrt{n}$-consistent. We have $Y_i^* \approx Y_i - X_i^T\beta$ and consequently

$$Y_i^* \approx f(Z_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{17}$$

In such a way, we have reduced model (1) approximatively to a standard nonparametric regression model. It is noteworthy that, although relatively new in partially linear models, difference-based methods have been extensively studied in nonparametric regression (Rice 1984; Gasser et al. 1986; Hall et al. 1990; Dette et al. 1998; Hall and Keilegom 2003; Tong and Wang 2005; Tong et al. 2013; Dai et al. 2015; Zhou et al. 2015; Wang and Lin 2015; Dai et al. 2016, 2017; Wang et al. 2017). In this paper, we propose to apply the least squares methods in Tong and Wang (2005) and Tong et al. (2013) to optimally estimate σ^2 using model (17). Let

$$s_k = \frac{1}{2(n-k)} \sum_{i=k+1}^{n} \left(Y_i^* - Y_{i-k}^*\right)^2, \quad k = 1, 2, \ldots, \tau.$$

We refer to s_k as the lag-k Rice estimators. For any fixed $\tau = o(n)$ with the equidistant design, it is easy to verify that

$$\begin{aligned} E(s_k) &\approx \frac{1}{2(n-k)} \sum_{i=k+1}^{n} E\left\{(f(Z_i) + \varepsilon_i) - (f(Z_{i-k}) + \varepsilon_{i-k})\right\}^2 \\ &\approx \sigma^2 + d_k J, \end{aligned}$$

where $d_k = k^2/n^2$ and $J = \int_0^1 (f'(x))^2 dx/2$.

Now treating s_k as the response variable and d_k as the independent variable, we can fit the following linear regression model and estimate σ^2 as the fitted intercept,

$$s_k = \alpha + \gamma d_k + \epsilon_k, \quad k = 1, 2, \ldots, \tau, \tag{18}$$

where ϵ_k are random errors. Note that s_k involves $(n-k)$ pairs of difference. We assign the weights $w_k = (n-k)/N$ to the response variable s_k where $N = \sum_{k=1}^{\tau}(n-k) = n\tau - \tau(\tau+1)/2$. We fit model (18) to get the weighted least squares estimate. By minimizing the weighted sums of squares $\sum_{k=1}^{\tau} w_k(s_k - \alpha - \gamma d_k)^2$, we have the estimator of σ^2 as

$$\hat{\sigma}^2_{\text{new}} = \bar{s}_w - \hat{\gamma}\bar{d}_w, \tag{19}$$

where $\hat{\gamma} = \sum_{k=1}^{\tau} w_k s_k(d_k - \bar{d}_w)/\sum_{k=1}^{\tau} w_k(d_k - \bar{d}_w)^2$, $\bar{s}_w = \sum_{k=1}^{\tau} w_k s_k$ and $\bar{d}_w = \sum_{k=1}^{\tau} w_k d_k$. Let $h_0 = 0$ and $h_k = 1 - (d_k - \bar{d}_w)\bar{d}_w/\sum_{k=1}^{\tau} w_k(d_k - \bar{d}_w)^2$ for $k = 1, 2, \ldots, \tau$. The quadratic form of $\hat{\sigma}^2_{\text{new}}$ can be represented as

$$\hat{\sigma}^2_{\text{new}} = \frac{1}{2N} Y^{*T} H Y^*, \tag{20}$$

where $H = (h_{ij})_{n\times n}$ is a symmetric matrix with elements $h_{ij} = \sum_{i=1}^{\tau} h_k + \sum_{k=0}^{\min\{i-1,n-i,\tau\}} h_k$ for $i = j$, $h_{ij} = -h_{|i-j|}$ for $0 < |i-j| \le \tau$ and $h_{ij} = 0$ otherwise, and $Y^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)^T$.

In what follows, we derive the theoretical results of the proposed estimators including the asymptotic MSE and the asymptotic normality. For comparison, the asymptotic MSE of the partially residual-based variance estimator $\hat{\sigma}_{\mathrm{W}}^2$ is also derived.

Theorem 2 *Assume that f has a bounded first derivative and $E(\varepsilon^4) < \infty$. For the equidistant design with $m \to \infty$ and $m/n \to 0$, we have*

$$\mathrm{MSE}\left(\hat{\sigma}_{\mathrm{W}}^2\right) = \frac{1}{n}\left\{\mathrm{var}(\varepsilon^2) + 4\sigma^4 \sum_{k=1}^{m}\sum_{j=1}^{m-k} d_j^2 d_{j+k}^2\right\} + o\left(\frac{1}{n}\right).$$

For the equidistant design with $\tau \to \infty$ and $\tau/n \to 0$, we have

$$\mathrm{MSE}\left(\hat{\sigma}_{\mathrm{new}}^2\right) = \frac{1}{n}\mathrm{var}(\varepsilon^2) + o\left(\frac{1}{n}\right).$$

Theorem 3 *For the equidistant design, $\hat{\sigma}_{\mathrm{new}}^2$ is an unbiased estimator of σ^2 when f is a linear function, regardless of the choice of τ.*

Theorem 4 *Assume that f has a bounded second derivative and $E(\varepsilon^6) < \infty$. For any $\tau = o(n^r)$ with $0 < r < 1/2$, then*

$$\sqrt{n}\left(\hat{\sigma}_{\mathrm{new}}^2 - \sigma^2\right) \xrightarrow{D} N\left(0, (\gamma_4 - 1)\sigma^4\right),$$

where $\gamma_4 = E\varepsilon^4/\sigma^4$ and $\xrightarrow{D}$ denotes in distribution convergence.

The proofs of the theorems are given in Sect. 6. Theorem 2 shows that our proposed estimator achieves the asymptotic optimal rate of MSE, and it is hence a more efficient estimator of σ^2 than $\hat{\sigma}_{\mathrm{W}}^2$. Theorem 4 establishes the asymptotic normality for the proposed estimator $\hat{\sigma}_{\mathrm{new}}^2$. It can be used to construct confidence intervals for σ^2. For instance, if $n > (\gamma_4 - 1)z_{\alpha/2}^2$ where z_α is the upper α-th percentile of the standard normal distribution and if $\hat{\gamma}_4$ is an estimate of γ_4, then an approximate $1-\alpha$ confidence interval for σ^2 can be constructed as $(\hat{\sigma}_{\mathrm{new}}^2/\{1 + z_{\alpha/2}\sqrt{(\hat{\gamma}_4 - 1)/n}\}, \hat{\sigma}_{\mathrm{new}}^2/\{1 - z_{\alpha/2}\sqrt{(\hat{\gamma}_4 - 1)/n}\})$.

4 Simulation studies

In this section, we report two simulation studies. The first study is to assess the impact of the various sequences on the performance of the estimator $\hat{\beta}$, and the other one is to evaluate the finite sample performance of the new estimator and compare it with the existing competitor.

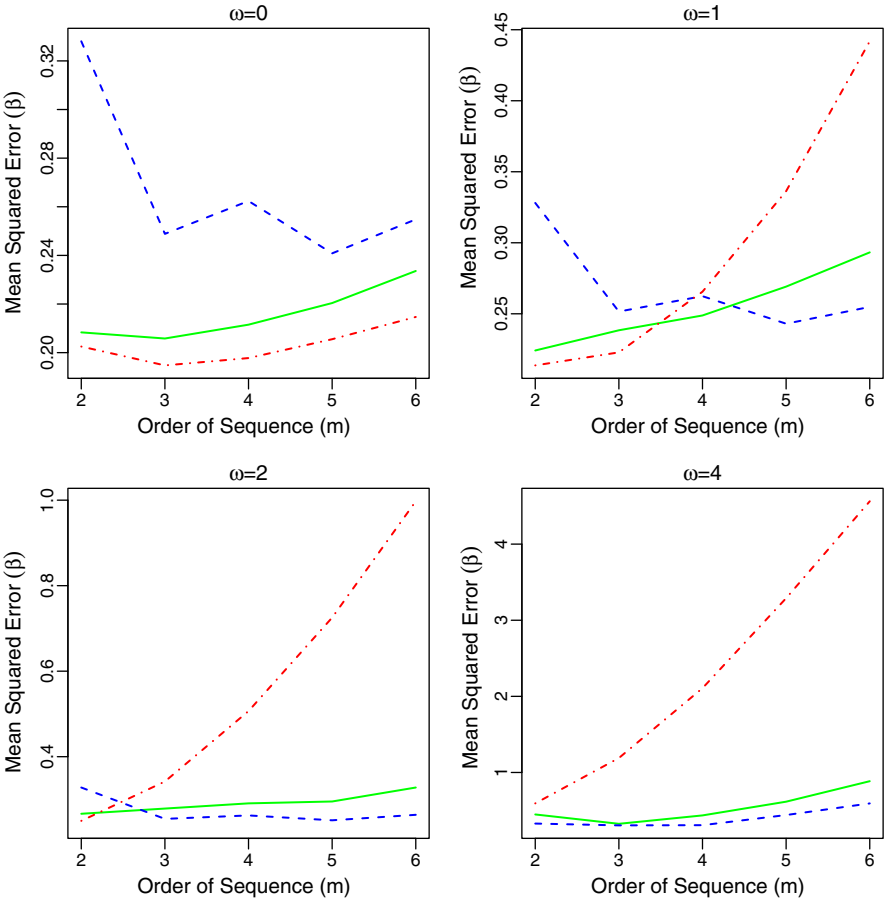

Fig. 2 MSEs for $n = 25$, $\beta = (2, 2, 4)^T$ and $f(z) = 5\sin(\omega\pi z)$, respectively. Solid lines: the adaptive sequence; Dashed lines: the HKT sequence; Dot-dashed lines: the WBC sequence

4.1 Sequence selection

To assess the impact of the sequences on the estimation, we consider the following three estimators for $\hat{\beta}$ in (4): $\hat{\beta}$ with the WBC sequence, $\hat{\beta}$ with the HKT sequence, and $\hat{\beta}$ with the adaptive sequence by the cross-validation method. For the sample size, we consider $n = 25$ and 200. For the linear component of the regression model, we consider $\beta = (2, 2, 4)^T$ and X_i are i.i.d. from $N((1, 2, 3)^T, I_3)$, where I_3 is an identity matrix of size 3×3. For the nonparametric component of the regression model, we consider $f(Z) = 5\sin(\omega\pi Z)$ with $\omega = 0,\ 1,\ 2,\ 4$ for $n = 25$, and $w = 0,\ 2,\ 4,\ 6$ for $n = 200$, respectively. The design points Z_i are equidistant with $Z_i = i/n$ for $i = 1, \ldots, n$. Finally, the random errors ε_i are independently generated from $N(0, 1)$.

For each simulation setting, we repeat the procedure for 1000 times and compute the corresponding MSEs of $\hat{\beta}$ for the distinct sequences. We then plot the MSE($\hat{\beta}$) of the considered estimators in Figs. 2 and 3, respectively. From the simulated results, we

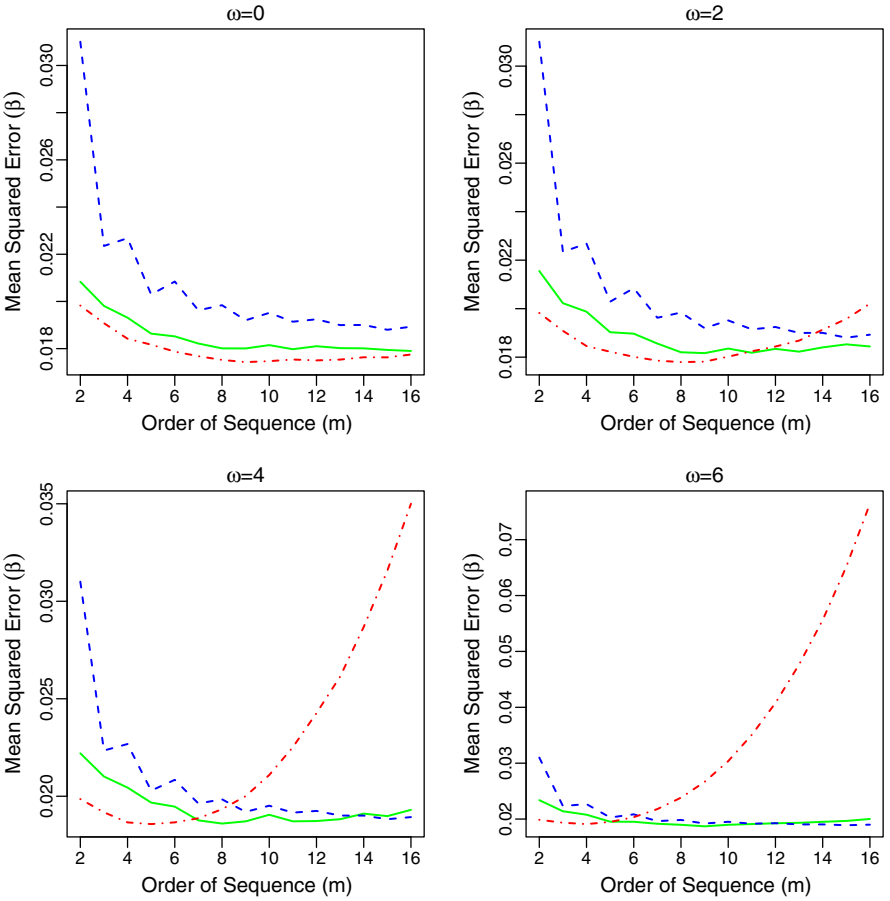

Fig. 3 MSEs for $n = 200$, $\beta = (2, 2, 4)^T$ and $f(z) = 5\sin(\omega\pi z)$, respectively. Solid lines: the adaptive sequence; Dashed lines: the HKT sequence; Dot-dashed lines: the WBC sequence

observe that when f is a constant function, i.e., when $\omega = 0$, the estimator with the WBC sequence provides the smallest mean squared error. In general, MSE($\hat{\beta}$(WBC)) increases as m increases whereas MSE($\hat{\beta}$(HKT)) decreases as m increases. Compared with the WBC sequence and the HKT sequence, the adaptive sequence performs relatively well in most settings, especially when f is a very rough function.

4.2 Variance estimation

We now conduct a simulation study to evaluate the finite sample performance of the proposed estimator $\hat{\sigma}^2_{\text{new}}$ and compared it with the estimator $\hat{\sigma}^2_{\text{W}}$. For the parametric part, we consider $p = 2$, $\beta = (1, 1.5)'$, both x_{i1} and x_{i2} are generated from the uniform distribution $U(0, 1)$. For the nonparametric part, we consider the function $f = 5\sin(\omega\pi Z)$ with $\omega = 0$, 1, 2, 4, and 6, denoted by f_1, f_2, f_3, f_4 and f_5 respectively.

Table 1 The relative MSEs of $\hat{\sigma}^2_{\mathrm{W}}$ and $\hat{\sigma}^2_{\mathrm{new}}$ under various simulation settings

n	σ	Methods	f_1	f_2	f_3	f_4	f_5
30	0.5	$\hat{\sigma}^2_{\mathrm{W}}$	1.53	37.05	524.75	4218.4	8396.3
		$\hat{\sigma}^2_{\mathrm{new}}$	1.54	1.58	2.93	31.16	289.4
	2	$\hat{\sigma}^2_{\mathrm{W}}$	1.54	1.67	4.04	23.44	57.09
		$\hat{\sigma}^2_{\mathrm{new}}$	1.53	1.51	1.53	1.57	2.50
100	0.5	$\hat{\sigma}^2_{\mathrm{W}}$	1.27	2.34	19.63	284.1	1392.8
		$\hat{\sigma}^2_{\mathrm{new}}$	1.13	1.13	1.14	1.45	5.66
	2	$\hat{\sigma}^2_{\mathrm{W}}$	1.24	1.26	1.33	2.39	6.84
		$\hat{\sigma}^2_{\mathrm{new}}$	1.14	1.14	1.13	1.13	1.13
400	0.5	$\hat{\sigma}^2_{\mathrm{W}}$	1.35	1.37	1.51	5.82	25.64
		$\hat{\sigma}^2_{\mathrm{new}}$	1.19	1.19	1.16	1.15	1.15
	2	$\hat{\sigma}^2_{\mathrm{W}}$	1.38	1.38	1.30	1.34	1.41
		$\hat{\sigma}^2_{\mathrm{new}}$	1.19	1.19	1.16	1.14	1.14

The design points Z_i are equally spaced with $Z_i = i/n$ and the random errors ε_i are generated independently from the normal distribution $N(0, \sigma^2)$. In addition, we consider $\sigma^2 = 0.25$ and 4 to represent the small and large variances, and $n = 30$, 100 and 400 to represent the small, moderate and large sample sizes, respectively.

For the difference sequence d, we choose the adaptive sequence as proposed in Sect. 2.3 and the order of sequence $m = 4$. For the bandwidth τ, we choose $\tau = n^{1/3}$. We repeat the simulation 1000 times for each setting and report the relative mean squared errors, $n\mathrm{MSE}/(2\sigma^4)$, of the two estimators in Table 1. From Table 1, it is evident that our proposed estimator $\hat{\sigma}^2_{\mathrm{new}}$ performs better than the existing competitor $\hat{\sigma}^2_{\mathrm{W}}$ in most settings. We also note that the performance of $\hat{\sigma}^2_{\mathrm{W}}$ depends heavily on the smoothness of f, the sample size n, and the signal-to-noise ratio. In particular, when σ^2 is small and f is rough, $\hat{\sigma}^2_{\mathrm{W}}$ fails to provide a reasonable estimate. Together with Theorem 2, we conclude that our proposed estimator improves the existing estimator significantly in both theory and simulations.

5 Conclusion

In the first part of the paper, we investigate the choice of difference sequence for $\hat{\beta}$ in a given sequence family. We derive that the mean squared error of the difference-based estimator consists of two distinct components: one is associated with the nonparametric component and the other one with the random errors. It turns out that the existing sequences are just special cases in the sequence family. Subsequently, we give an adaptive sequence by the cross-validation method. Simulation studies indicate that the criterion is quite effective and the adaptive sequence assesses good performance for most settings. In the second part of the paper, we propose a new estimator for the residual variance σ^2 by fitting a linear regression model to difference-based variance

estimators. We have also derived the theoretical results including the asymptotic MSE and the asymptotic normality of the proposed estimator. In both theory and simulations, we demonstrate that the proposed estimator performs better than the partially residual-based estimator in Wang et al. (2011).

For simplicity, the covariate Z_i are assumed to be equally spaced design points with $Z_i = i/n$ in this paper. In practice, our proposed method can also be readily applied to unequally spaced designs with the design points satisfying $Z_i - Z_{i-1} = 1/n + o(1/n)$. In addition, we note that our proposed method requires the design points to be ordered. Hence as the classical difference-based methods, it may not be easy to extend our method to general models with high-dimensional data or infinite-dimensional data with applications in functional data analysis (Aneiros et al. 2015). Further research is needed in this direction.

6 Proofs

We first present three lemmas. Lemmas 1 and 2 are immediate results from Schott (1997) and Whittle (1964), respectively. Proof of Lemma 3 will be given.

Lemma 1 *Let X be an length m random vector with finite fourth moments so that both $E(XX^T)$ and $E(XX^T \otimes XX^T)$ exist. Let μ and Ω denote the mean vector and covariance matrix of X, respectively. Then for any $m \times m$ symmetric matrix A, we have*

$$\mathrm{Var}(X^T AX) = tr\{(A \otimes A)E(XX^T \otimes XX^T)\} - \{tr(A\Omega) + \mu^T A\mu\}^2.$$

Lemma 2 *Assume that the matrix $A = (a_{ij})_{nn}$ satisfies $a_{ij} = a_{i-j}$ and $\sum_{-\infty}^{\infty} a_{i-j} < \infty$. Furthermore, assume that $E(\varepsilon^6)$ is finite. Then*

$$\frac{1}{n}\varepsilon^T A\varepsilon = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i-j}\varepsilon_i\varepsilon_j \xrightarrow{D} N\left(a_0\sigma^2, \sigma_A^2\right),$$

where $\sigma_A^2 = (\gamma_4 - 3)a_0^2\sigma^4/n + 2\sigma^4\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i-j}^2/n^2$.

Lemma 3 *Assume that $\tau \to \infty$, $\tau/n \to 0$. Then,*

(1) $\sum_{k=1}^{\tau} h_k = O(\tau)$,
(2) $\sum_{k=1}^{\tau} h_k^2 = O(\tau)$,
(3) $\sum_{k=j}^{\tau} h_k = \tau - 9/4j + \frac{5j^3}{4\tau^2} + o(\tau)$,
(4) $f^T Hf = O(\tau^4/n^2)$,
(5) $f^T H^2 f = O(\tau^5/n^2)$.

Proof (1) Note that $N = n\tau - \tau(\tau+1)/2$. It is easy to show that

$$\bar{d}_w = \frac{I_2}{Nn} - \frac{I_3}{Nn^2} = \frac{\tau^2}{3n^2} + o\left(\frac{\tau^2}{n^2}\right),$$
$$\sum_{k=1}^{\tau} w_k (d_k - \bar{d}_w)^2 = \frac{I_4}{Nn^3} - \frac{I_5}{Nn^4} - \left(\frac{I_2}{Nn} - \frac{I_3}{Nn^2}\right)^2 = \frac{4\tau^4}{45n^4} + o\left(\frac{\tau^4}{n^4}\right),$$

where $I_t = \sum_{k=1}^{\tau} k^t$ for $t = 2, 3, 4, 5$. Hence, we obtain

$$\eta = \frac{\bar{d}_w}{\sum_{k=1}^{\tau} w_k \left(d_k - \bar{d}_w\right)^2} = \frac{15n^2}{4\tau^2} + o\left(\frac{n^2}{\tau^2}\right).$$

In addition, we have

$$\begin{aligned}\sum_{k=1}^{\tau} \left(d_k - \bar{d}_w\right) &= \frac{1}{n^2}\sum_{k=1}^{\tau} k^2 - \frac{\tau}{Nn^2}\sum_{k=1}^{\tau}(n-k)k^2 \\ &= \frac{\tau}{Nn^2}\sum_{k=1}^{\tau} k^3 + \frac{1}{n^2}\left(1 - \frac{\tau n}{N}\right)\sum_{k=1}^{\tau} k^2 \\ &= \frac{\tau^4}{12n^3} + o\left(\frac{\tau^4}{n^3}\right).\end{aligned}$$

This leads to $\sum_{k=1}^{\tau} h_k = \tau - \eta\sum_{k=1}^{\tau}(d_k - \bar{d}_w) = \tau - \frac{5\tau^2}{16n} + o\left(\frac{\tau^2}{n}\right) = O(\tau)$.

(2) Note that $\eta = \frac{15n^2}{4\tau^2} + o\left(\frac{n^2}{\tau^2}\right)$ and $\sum_{k=1}^{\tau}(d_k - \bar{d}_w) = \frac{\tau^4}{12n^3} + o\left(\frac{\tau^4}{n^3}\right)$. Furthermore, it is easy to show that $\sum_{k=1}^{\tau}(d_k - \bar{d}_w)^2 = \frac{4\tau^5}{45n^4} + o\left(\frac{\tau^5}{n^4}\right)$. Then,

$$\begin{aligned}\sum_{k=1}^{\tau} h_k^2 &= \tau - 2\eta\sum_{k=1}^{\tau}\left(d_k - \bar{d}_w\right) + \eta^2\sum_{k=1}^{\tau}\left(d_k - \bar{d}_w\right)^2 \\ &= \tau - \left(\frac{5\tau^2}{8n} + o\left(\frac{\tau^2}{n}\right)\right) + \left(\frac{15n^2}{4\tau^2} + o\left(\frac{n^2}{\tau^2}\right)\right)^2\left(\frac{4\tau^5}{45n^4} + o\left(\frac{\tau^5}{n^4}\right)\right) \\ &= \frac{9}{4}\tau + o(\tau) = O(\tau).\end{aligned}$$

(3) For $1 \le j \le \tau$, $\sum_{k=j}^{\tau} h_k = \sum_{k=1}^{\tau} h_k - \sum_{k=1}^{j-1} h_k$. Note that $\eta\bar{d}_w = 5/4 + o(1)$. Then, we have

$$\begin{aligned}\sum_{k=1}^{j-1} h_k &= (j-1)\left(1+\eta\bar{d}_w\right) - \eta\sum_{k=1}^{j-1} d_k \\ &= (j-1)\left(1+5/4+o(1)\right) - \left(\frac{15n^2}{4\tau^2}+o\left(\frac{n^2}{\tau^2}\right)\right)\left(\frac{j^3}{3n^2}+O\left(\frac{j^2}{n^2}\right)\right) \\ &= \frac{9}{4}j - \frac{5j^3}{4\tau^2} + o(j) + O(1).\end{aligned}$$

By $\sum_{k=1}^{\tau} h_k = \tau - \frac{5\tau^2}{16n} + o\left(\frac{\tau^2}{n}\right)$, we get $\sum_{k=j}^{\tau} h_k = \tau - \frac{9}{4}j + \frac{5j^3}{4\tau^2} + o(\tau)$.

(4) We have

$$\begin{aligned}f^T H f &= \sum_{k=1}^{\tau}\left\{h_k \sum_{i=k+1}^{n}(f_i - f_{i-k})^2\right\} \\ &= \sum_{k=1}^{\tau}\left\{h_k \sum_{i=k+1}^{n}\left(f_i'\frac{k}{n}+O\left(\frac{k^2}{n^2}\right)\right)^2\right\} \\ &= \sum_{k=1}^{\tau}\left[h_k\left\{\frac{k^2}{n^2}\sum_{i=k+1}^{n} f_i'^2 + O\left(\frac{(n-k)k^3}{n^3}\right)\right\}\right] \\ &= \frac{J}{n}\sum_{k=1}^{\tau} k^2 h_k + O\left(\frac{1}{n^2}\right)\sum_{k=1}^{\tau} k^3 h_k,\end{aligned}$$

where $J = \int_0^1 f'(x)^2 dx$. Note that $\sum_{k=1}^{\tau} k^2 h_k = (1+\eta\bar{d}_w)\sum_{k=1}^{\tau} k^2 - \eta\sum_{k=1}^{\tau} k^2 = \frac{3}{4}\tau^3 - \frac{3}{4}\tau^3 + o(\tau^3) = o(\tau^3)$ and $\sum_{k=1}^{\tau} k^3 h_k = (1+\eta\bar{d}_w)\sum_{k=1}^{\tau} k^3 - \eta\sum_{k=1}^{\tau} k^3 = O(\tau^4)$. Hence, we get $f^T H f = O(\tau^4/n^2)$.

(5) Since H is symmetric, then $f^T H^2 f = f^T H^T H f = (Hf)^T Hf = p^T p$, where $p = Hf = (p_1, p_2, \ldots, p_n)^T$. For $i \in [\tau+1, n-\tau]$, we have

$$\begin{aligned}p_i &= \sum_{k=1}^{\tau} h_k(f_i - f_{i-k}) - \sum_{k=1}^{\tau} h_k(f_{i+k} - f_i) \\ &= \sum_{k=1}^{\tau} h_k\left(\frac{k}{n}f_i' - \frac{k^2}{2n^2}f_i'' + o\left(\frac{k^2}{2n^2}\right)\right) - \sum_{k=1}^{\tau} h_k\left(\frac{k}{n}f_i' + \frac{k^2}{2n^2}f_i'' + o\left(\frac{k^2}{2n^2}\right)\right) \\ &= -\frac{1}{n^2}f_i''\sum_{k=1}^{\tau} k^2 h_k + o\left(\frac{\tau^3}{n^2}\right) = o\left(\frac{\tau^3}{n^2}\right),\end{aligned}$$

where f_i' and f_i'' denote first and second derivative of $f(Z_i)$, respectively. For $i \in [1, \tau]$, we have

$$p_i = \sum_{k=1}^{i-1} h_k(f_i - f_{i-k}) - \sum_{k=1}^{\tau} h_k(f_{i+k} - f_i)$$

$$= \sum_{k=1}^{i-1} h_k \left(\frac{k}{n} f_i' - \frac{k^2}{2n^2} f_i'' + o\left(\frac{k^2}{2n^2}\right)\right) - \sum_{k=1}^{\tau} h_k \left(\frac{k}{n} f_i' + \frac{k^2}{2n^2} f_i'' + o\left(\frac{k^2}{2n^2}\right)\right)$$
$$= -\frac{f_i'}{n} \sum_{k=i}^{\tau} k h_k - \frac{f_i''}{2n^2} \sum_{k=i}^{\tau} k^2 h_k + o\left(\frac{\tau^3}{n^2}\right) = O\left(\frac{\tau^2}{n}\right).$$

Similarly, for $i \in [n-\tau+1, n]$, we can show that $p_i = O\left(\frac{\tau^2}{n}\right)$. Then, we have

$$f^T H^2 f = p^T p = \sum_{i=1}^{\tau} p_i^2 + \sum_{i=\tau+1}^{n-\tau} p_i^2 + \sum_{i=n-\tau+1}^{n} p_i^2 = O\left(\frac{\tau^5}{n^2}\right).$$

□

6.1 Proof of Theorem 1

By (4) and (8), we have $\hat{\beta} = \beta + (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \delta + (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{\varepsilon}$. Note that δ and $\tilde{\varepsilon}$ are independent of each other and $E\{(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{\varepsilon}\} = 0$. Then

$$\text{MSE}\left(\hat{\beta}\right) = \text{Var}\left((\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{\varepsilon}\right) + E\left((\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \delta\right)\left((\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \delta\right)^T. \quad (21)$$

For the first term, by Remark 5 in Wang et al. (2011), we have

$$\text{Var}\left((\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{\varepsilon}\right) \sim \frac{1}{n}\left(1 + 2\sum_{l=1}^{m} c_l^2\right) \sigma^2 \Sigma_X^{-1}, \quad (22)$$

where $c_l = \sum_{j=0}^{m-l} d_j d_{j+l}$ for $l = 1, \ldots, m$, and '$\sim$' is defined the same way as in Stirling's approximation.

Now we derive the second term in (21). For ease of notation, let $A = (D^T\delta)(D^T\delta)^T = (a_{ij})_{n\times n}$, $B = \mu\mu^T$. Note that $\tilde{X} = DX$ and $\tilde{X}^T\tilde{X}/n \to \Sigma_X$ as $n \to \infty$. We have

$$E\left((\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \delta\right)\left((\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \delta\right)^T = \frac{1}{n^2} E\left\{(\tilde{X}^T\tilde{X}/n)^{-1}(\tilde{X}^T\delta)(\tilde{X}^T\delta)^T(\tilde{X}^T\tilde{X}/n)^{-1}\right\}$$
$$\sim \frac{1}{n^2}\Sigma_X^{-1} E\left\{X^T(D^T\delta)(D^T\delta)^T X\right\}\Sigma_X^{-1}.$$

Further, we have

$$E\left\{X^T(D^T\delta)(D^T\delta)^T X\right\} = E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} X_i X_j^T\right\}$$
$$= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\{(X_i-\mu)+\mu\}\{(X_j-\mu)+\mu\}^T\right]$$

$$= \sum_{i=1}^{n} a_{ii}\,\Sigma_X + \sum_{i,j=1}^{n} a_{ij}\,\mu\mu^T$$
$$= \mathrm{tr}(A)\,\Sigma_X + \sum_{i,j=1}^{n} a_{ij}\,B.$$

Then,

$$E\left((\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\delta\right)\left((\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\delta\right)^T \sim \frac{1}{n^2}\left\{\mathrm{tr}(A)\,\Sigma_X^{-1} + \sum_{i,j=1}^{n} a_{ij}\,\Sigma_X^{-1}B\,\Sigma_X^{-1}\right\}. \tag{23}$$

Finally, by (21), (22) and (23), the approximate MSE of $\hat{\beta}$ is given as

$$\mathrm{MSE}\left(\hat{\beta}\right) \sim \frac{1}{n^2}\left\{\mathrm{tr}(A)\,\Sigma_X^{-1} + \sum_{i,j=1}^{n} a_{ij}\,\Sigma_X^{-1}B\,\Sigma_X^{-1}\right\} + \frac{1}{n}\left(1 + 2\sum_{l=1}^{m} c_l^2\right)\sigma^2\,\Sigma_X^{-1}. \tag{24}$$

6.2 Proof of Theorem 2

By the definition of $\hat{\sigma}_{\mathrm{W}}^2$ and $\sqrt{n}$-consistency of $\hat{\beta}$, we have

$$\hat{\sigma}_{\mathrm{W}}^2 = \frac{1}{n-m-p}\sum_{i=1}^{n-m}\left(\tilde{Y}_i - \tilde{X}_i^T\hat{\beta}\right)^2$$
$$= \frac{1}{n-m-p}\sum_{i=1}^{n-m}\left\{\tilde{X}_i^T(\beta-\hat{\beta}) + \delta_i + \tilde{\varepsilon}_i\right\}^2$$
$$= \frac{1}{n-m-p}\sum_{i=1}^{n-m}\tilde{\varepsilon}_i^2 + O_p\left(\frac{1}{n}\right).$$

Hence, we obtain that

$$\mathrm{E}(\hat{\sigma}_{\mathrm{W}}^2) = \frac{n-m}{n-m-p}\sigma^2 + O\left(\frac{1}{n}\right), \tag{25}$$

$$\mathrm{Var}(\hat{\sigma}_{\mathrm{W}}^2) = \frac{1}{(n-m-p)^2}\mathrm{Var}\left(\sum_{i=1}^{n-m}\tilde{\varepsilon}_i^2\right) + o\left(\frac{1}{n}\right)$$
$$= \frac{1}{(n-m-p)^2}\left\{\sum_{i=1}^{n-m}\mathrm{Var}(\tilde{\varepsilon}_i^2) + 2\sum_{i<j}Cov(\tilde{\varepsilon}_i^2,\tilde{\varepsilon}_j^2)\right\} + o\left(\frac{1}{n}\right). \tag{26}$$

Next, we calculate $\mathrm{Var}(\tilde{\varepsilon}_i^2)$ and $Cov(\tilde{\varepsilon}_i^2, \tilde{\varepsilon}_j^2)$, respectively. We know that

$$\begin{aligned}
\mathrm{Var}\left(\tilde{\varepsilon}_i^2\right) &= E\left(\tilde{\varepsilon}_i^2 - \sigma^2\right)^2 = E\left(\tilde{\varepsilon}_i^4\right) - \sigma^4 \\
&= E\left\{\sum_{j=0}^{m} d_j^2 \varepsilon_{i+j}^2 + 2 \sum_{0\le p<q\le m} d_p d_q \varepsilon_{i+p}\varepsilon_{i+q}\right\}^2 - \sigma^4 \\
&= E\left\{\sum_{j=0}^{m} d_j^2 \varepsilon_{i+j}^2\right\}^2 + 4E\left\{\sum_{0\le p<q\le m} d_p d_q \varepsilon_{i+p}\varepsilon_{i+q}\right\}^2 - \sigma^4 \\
&= \sum_{j=0}^{m} d_j^4 E\varepsilon^4 + 6 \sum_{0\le p<q\le m} d_p^2 d_q^2 \sigma^4 - \sigma^4 \\
&= \sum_{j=0}^{m} d_j^4 \mathrm{Var}(\varepsilon^2) + 4 \sum_{0\le p<q\le m} d_p^2 d_q^2 \sigma^4, \qquad (27)
\end{aligned}$$

and

$$\begin{aligned}
Cov\left(\tilde{\varepsilon}_i^2, \tilde{\varepsilon}_j^2\right) &= E\left(\tilde{\varepsilon}_i^2 \tilde{\varepsilon}_j^2\right) - \sigma^4 \\
&= E\left\{\left(\sum_{s=0}^{m} d_s \varepsilon_{i+s}\right)\left(\sum_{t=0}^{m} d_t \varepsilon_{j+t}\right)\right\}^2 - \sigma^4 \\
&= \sum_{s=0}^{m}\sum_{t=0}^{m} d_s^2 d_t^2 E(\varepsilon_{i+s}^2 \varepsilon_{j+t}^2) - \sigma^4 \\
&= \left\{\sum_{s=0}^{m}\sum_{t=0}^{m}\right\}_{i+s=j+t} d_s^2 d_t^2 E(\varepsilon^4) + \left\{\sum_{s=0}^{m}\sum_{t=0}^{m}\right\}_{i+s\ne j+t} d_s^2 d_t^2 \sigma^4 - \sigma^4 \\
&= \left\{\sum_{s=0}^{m}\sum_{t=0}^{m}\right\}_{i+s=j+t} d_s^2 d_t^2 \mathrm{Var}(\varepsilon^2). \qquad (28)
\end{aligned}$$

Plugging (27) and (28) into (26), we have

$$\mathrm{Var}\left(\hat{\sigma}_{\mathrm{W}}^2\right) = \frac{n-m}{(n-m-p)^2}\left\{\mathrm{Var}(\varepsilon^2) + 4\sigma^4 \sum_{k=1}^{m}\sum_{j=0}^{m-k} d_j^2 d_{j+k}^2\right\} + o\left(\frac{1}{n}\right). \qquad (29)$$

Combining (25) and (29), as $m/n \to 0$ with $m \to \infty$, the asymptotic MSE of $\hat{\sigma}_{\mathrm{W}}^2$ can be denotes as

$$\mathrm{MSE}\left(\hat{\sigma}_{\mathrm{W}}^2\right) = \frac{1}{n}\left\{\mathrm{Var}(\varepsilon^2) + 4\sigma^4 \sum_{k=1}^{m}\sum_{j=0}^{m-k} d_j^2 d_{j+k}^2\right\} + o\left(\frac{1}{n}\right). \qquad (30)$$

In what follows, we consider the asymptotic MSE of $\hat{\sigma}^2_{\text{new}}$. By the definition of $\hat{\sigma}^2_{\text{new}}$ and $\sqrt{n}$-consistency of $\hat{\beta}$, we have

$$\begin{aligned}\hat{\sigma}^2_{\text{new}} &= \frac{1}{2N} Y^{*T} H Y^* = \frac{1}{2N}\left(X(\beta-\hat{\beta})+\delta+\varepsilon\right)^T H \left(X(\beta-\hat{\beta})+\delta+\varepsilon\right)\\ &= \frac{1}{2N}\left\{\varepsilon^T H\varepsilon + O_p\left(\frac{1}{\sqrt{n}}\right)\right\}.\end{aligned}$$

Then, we have

$$E\left(\hat{\sigma}^2_{\text{new}}\right) = \frac{tr(H)}{2N}\sigma^2 + O\left(\frac{1}{mn^{3/2}}\right), \tag{31}$$

$$\text{Var}\left(\hat{\sigma}^2_{\text{new}}\right) = \frac{1}{4N^2}\text{Var}\left(\varepsilon^T H\varepsilon\right) + O\left(\frac{1}{m^2n^{5/2}}\right). \tag{32}$$

By Lemma 1, we have

$$\begin{aligned}\text{Var}\left(\varepsilon^T H\varepsilon\right) &= tr\{(H\otimes H)E\left(\varepsilon\varepsilon^T\otimes\varepsilon\varepsilon^T\right)\} - \{tr(H)\sigma^2\}^2\\ &= tr\left\{(H\otimes H)E\left(\varepsilon\varepsilon^T\otimes\varepsilon\varepsilon^T\right)\right\} - \{tr(H)\}^2\sigma^4.\end{aligned} \tag{33}$$

We know

$$\begin{aligned}tr(H) &= 2\sum_{j=0}^{\tau-1}\left(\sum_{k=1}^{\tau}h_k+\sum_{k=0}^{j}h_k\right)+2(n-2\tau)\sum_{k=1}^{\tau}h_k,\\ tr\left\{(H\otimes H)E\left(\varepsilon\varepsilon^T\otimes\varepsilon\varepsilon^T\right)\right\} &= \left\{2\sum_{j=0}^{\tau-1}\left(\sum_{k=1}^{\tau}h_k+\sum_{k=0}^{j}h_k\right)^2+4(n-2\tau)\left(\sum_{k=1}^{\tau}h_k\right)^2\right\}E(\varepsilon^4)\\ &\quad+\left[2\sum_{j=0}^{\tau-1}\left(\sum_{k=1}^{\tau}h_k+\sum_{k=0}^{j}h_k\right)\left\{2\sum_{j=0}^{\tau-1}\left(\sum_{k=1}^{\tau}h_k+\sum_{k=0}^{j}h_k\right)\right.\right.\\ &\quad\left.+2(n-2\tau)\sum_{k=1}^{\tau}h_k-\left(\sum_{k=1}^{\tau}h_k+\sum_{k=0}^{j}h_k\right)\right\}\\ &\quad+2(n-2\tau)\left(\sum_{k=1}^{\tau}h_k\right)\left\{2\sum_{j=0}^{\tau-1}\left(\sum_{k=1}^{\tau}h_k+\sum_{k=0}^{j}h_k\right)\right.\\ &\quad\left.\left.+\ 2(n-2\tau-1)\sum_{k=1}^{\tau}h_k\right\}\right]\sigma^4,\end{aligned}$$

where $h_0 = 0$. Therefore, we have

$$\mathrm{Var}\left(\varepsilon^T H\varepsilon\right) = \left\{2\sum_{j=0}^{\tau-1}\left(\sum_{k=1}^{\tau} h_k + \sum_{k=0}^{j} h_k\right)^2 + 4(n-2\tau)\left(\sum_{k=1}^{\tau} h_k\right)^2\right\}\mathrm{Var}(\varepsilon^2). \tag{34}$$

Note that $\sum_{k=1}^{\tau} h_k = \tau - \frac{5\tau^2}{16n} + o\left(\frac{\tau^2}{n}\right)$ and $\sum_{k=1}^{j} h_k = \frac{9}{4}j - \frac{5j^3}{4\tau^2} + o(j) + o\left(\frac{j^3}{\tau^2}\right)$ for $1 \le j \le \tau$. By (31), (32) and (34), as $\tau/n \to 0$ with $\tau \to \infty$, we have the asymptotic MSE of $\hat{\sigma}^2_{\mathrm{new}}$ for

$$\mathrm{MSE}\left(\hat{\sigma}^2_{\mathrm{new}}\right) = \frac{1}{n}\mathrm{Var}(\varepsilon^2) + o\left(\frac{1}{n}\right).$$

6.3 Proof of Theorem 3

By the root-n consistency of $\hat{\beta}$ and smoothness of f, we have

$$\begin{aligned} E(s_k) &= \frac{1}{2(n-k)}\sum_{i=k+1}^{n} E\{(X_i - X_{i-k})'(\beta - \hat{\beta}) + (f_i - f_{i-k}) + (\varepsilon_i - \varepsilon_{i-k})\}^2 \\ &= \sigma^2 + o(n^{-1/2}), \end{aligned}$$

where $f_i = f(Z_i)$. Then,

$$\begin{aligned} E(\hat{\sigma}^2_{\mathrm{new}}) &= \sum_{k=1}^{\tau} w_k E(s_k) - \frac{\bar{d}_w}{\sum_{k=1}^{\tau} w_k (d_k - \bar{d}_w)^2}\sum_{k=1}^{\tau} w_k (d_k - \bar{d}_w) E(s_k) \\ &= \sigma^2 + o(n^{-1/2}). \end{aligned}$$

Thus, $\hat{\sigma}^2_{\mathrm{new}}$ is an asymptotically unbiased estimator of σ^2.

6.4 Proof of Theorem 4

Note that $Y^* = Y - X\hat{\beta} = X(\beta - \hat{\beta}) + f + \varepsilon$. We have

$$\hat{\sigma}^2_{\mathrm{new}} = \frac{1}{2N}(X(\beta - \hat{\beta}) + f + \varepsilon)^T H(X(\beta - \hat{\beta}) + f + \varepsilon).$$

Let $A = X(\beta - \hat{\beta})$. Then,

$$\begin{aligned}\hat{\sigma}^2_{\text{new}} &= \frac{1}{2N}(A + f + \varepsilon)^T H (A + f + \varepsilon) \\ &= \frac{1}{2N}\left(A^T HA + 2A^T Hf + f^T Hf + 2A^T H\varepsilon + 2f^T H\varepsilon + \varepsilon^T H\varepsilon\right) \\ &= I_1 + I_2 + I_3 + I_4 + I_5 + I_6.\end{aligned}$$

By the root-n consistency of $\hat{\beta}$ and the smoothness of f, we have $I_1 = o_p(n^{-1/2})$ and $I_2 = o_p(n^{-1/2})$. By $\delta_i = O(\tau/n)$ and $f'Hf = O(\tau^3/n)$. we have $I_3 = O(\tau^2/n^2)$. Further, for any $\tau = n^r$ with $0 < r < 3/4$, we have $I_3 = o(n^{-1/2})$. By the root-n consistency of $\hat{\beta}$ and for any τ, we have $E(A^T H\varepsilon/N)^2 = A^T HH^T A\sigma^2/N^2 = o(1/n)$. This implies that $I_4 = o_p(n^{-1/2})$. Note also that $E(f^T H\varepsilon/N) = f^T H^2 f\sigma^2/N^2$ and $f^T H^2 f = O(\tau^5/n^2)$. Then for any $\tau = o(n)$, we have $I_5 = o_p(n^{-1/2})$.

In what follows, we consider the limiting distribution of $\varepsilon^T H\varepsilon/(2N)$. Let $nH/(2N) = B - C$, where $B = (b_{ij})_{n\times n}$ with elements $b_{ij} = n\sum_{k=1}^{\tau} h_k/N$ for $i = j$, $b_{ij} = -nh_{|i-j|}/(2N)$ for $0 < |i-j| \le \tau$, $b_{ij} = 0$ otherwise; and $C = diag(c_1, c_2, \ldots, c_n)$ with $c_i = n\sum_{min(i,n+1-i,\tau+1)}^{\tau+1} h_k/(2N)$. Then,

$$\frac{1}{2N}\varepsilon^T H\varepsilon = \frac{1}{n}\varepsilon^T B\varepsilon - \frac{1}{n}\varepsilon^T C\varepsilon. \tag{35}$$

Note that the asymmetric matrix B satisfies $b_{ij} = b_{i-j}$ with $b_0 = n\sum_{k=1}^{\tau} h_k/N$, $b_{i-j} = b_{j-i} = -nh_{|i-j|}/(2N)$ for $0 < |i-j| \le \tau$ and $b_{i-j} = b_{j-i} = 0$ otherwise. By Lemma 3, for any $\tau = o(n)$, we have $\sum_{k=-\infty}^{\infty} b_k^2 = b_0^2 + 2\sum_{k=1}^{\tau} b_k^2 < \infty$. We assume that $E(\varepsilon^6) < \infty$. By Lemma 2, we have

$$\sqrt{n}\left(\frac{1}{n}\varepsilon^T B\varepsilon - b_0\sigma^2\right) \xrightarrow{d} N\left(0, \sigma_B^2\right), \tag{36}$$

where

$$\sigma_B^2 = \frac{n^2(\gamma_4 - 1)\sigma^4}{N^2}\left(\sum_{k=1}^{\tau} h_k\right)^2 + \frac{n\sigma^4}{N^2}\sum_{k=1}^{\tau}(n-k)h_k^2.$$

We know $\varepsilon^T C\varepsilon = \sum_{i=1}^{\tau} c_i\varepsilon_i^2 + \sum_{i=n-\tau+1}^{n} c_i\varepsilon_i^2$. Note that

$$\begin{aligned}E\left(\sum_{i=1}^{\tau} c_i\varepsilon_i^2\right)^2 &= (\gamma_4 - 1)\frac{n^2\sigma^4}{4N^2}\sum_{i=1}^{\tau}\left(\sum_{min(i,n+1-i,\tau+1)}^{\tau+1} h_k\right)^2 \\ &\quad + \frac{n^2\sigma^4}{4N^2}\left(\sum_{i=1}^{\tau}\sum_{min(i,n+1-i,\tau+1)}^{\tau+1} h_k\right)^2.\end{aligned}$$

By Lemma 3, we have $E\left(\sum_{i=1}^{\tau} c_i\varepsilon_i^2\right)^2 = O(\tau^2)$. Similarly, we can obtain $E\left(\sum_{i=n-\tau+1}^{n} c_i\varepsilon_i^2\right)^2 = O(\tau^2)$. Then, $E\left(\frac{1}{n}\varepsilon^T C\varepsilon\right)^2 = O(\tau^2/n^2)$. For any $\tau = n^r$ with $0 < r < 1/2$, we have

$$\frac{1}{n}\varepsilon^T C\varepsilon = o_p(n^{-1/2}).$$

By Slutsky's theorem, we can get

$$\sqrt{n}\left(\hat{\sigma}^2_{\text{new}} - b_0\sigma^2\right)/\sigma_B \xrightarrow{d} N(0,1), \quad as \quad n \to \infty.$$

Note also that $b_0 = 1 + O(\tau/n)$, $\hat{\sigma}_B^2 = (\gamma_4 - 1)\sigma^4 + o(1)$. Therefore, for any $\tau = n^r$ with $0 < r < 1/2$, we have $\sqrt{n}(b_0 - 1) = o(1)$. Applying Slutsky's theorem, we have

$$\begin{aligned}\frac{\sqrt{n}\left(\hat{\sigma}^2_{\text{new}} - \sigma^2\right)}{\sqrt{\gamma_4 - 1}\sigma^2} &= \frac{\sigma_B}{\sqrt{\gamma_4 - 1}\sigma^2}\left(\frac{\sqrt{n}\left(\hat{\sigma}^2_{\text{new}} - b_0\sigma^2\right)}{\sigma_B} + \frac{\sqrt{n}(b_0 - 1)\sigma^2}{\sigma_B}\right)\\ &\xrightarrow{d} N(0,1), \quad as \quad n \to \infty.\end{aligned}$$

This completes the proof of theorem.

Acknowledgements Yuejin Zhou's research was supported in part by the Natural Science Foundation of Anhui Grant (No. KJ2017A087), and the National Natural Science Foundation of China Grant (No. 61472003). Yebin Cheng's research was supported in part by the National Natural Science Foundation of China Grant (No. 11271241). Tiejun Tong's research was supported in part by the Hong Kong Baptist University Grants FRG1/16-17/018 and FRG2/16-17/074, and the National Natural Science Foundation of China Grant (No. 11671338).

References

Akdeniz F, Duran E (2013) New difference-based estimator of parameters in semiparametric regression models. J Stat Comput Simul 83:810–824

Aneiros G, Ling N, Vieu P (2015) Error variance estimation in semi-functional partially linear regression models. J Nonparametr Stat 27:316–330

Chen H, Shiau J (1991) A two-stage spline smoothing method for partially linear models. J Stat Plan Inference 27:187–201

Cuzick J (1992) Semiparametric additive regression. J Roy Stat Soc B 54:831–843

Dai W, Ma Y, Tong T, Zhu L (2015) Difference-based variance estimation in nonparametric regression with repeated measurement data. J Stat Plan Inference 163:1–20

Dai W, Tong T, Genton M (2016) Optimal estimation of derivatives in nonparametric regression. J Mach Learn Res 17:1–25

Dai W, Tong T, Zhu L (2017) On the choice of difference sequence in a unified framework for variance estimation in nonparametric regression. Stat Sci 32:455–468

Dette H, Munk A, Wagner T (1998) Estimating the variance in nonparametric regression—what is a reasonable choice? J Roy Stat Soc B 60:751–764

Eubank R, Kambour E, Kim J, Klipple K, Reese C, Schimek M (1998) Estimation in partially linear models. Comput Stat Data Anal 29:27–34

Fan J, Huang T (2005) Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli 11:1031–1057

Gasser T, Sroka L, Jennen-Steinmetz C (1986) Residual variance and residual pattern in nonlinear regression. Biometrika 73:625–633
Hall P, Kay JW, Titterington DM (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika 77:521–528
Hall P, Keilegom IV (2003) Using difference-based methods for inference in nonparametric regression with time series errors. J Roy Stat Soc B 65:443–456
Hardle W, Liang H, Gao J (2000) Partially linear models. Physika Verlag, Heidelberg
He H, Tang W, Zuo G (2014) Statistical inference in the partial inear models with the doubles moothing local linear regression method. J Stat Plan Inference 146:102–112
Hu H, Zhang Y, Pan X (2016) Asymptotic normality of dhd estimators in a partially linear model. Stat Pap 57:567–587
Levine M (2015) Minimax rate of convergence for an estimator of the functional component in a semiparametric multivariate partially linear model. J Multivar Anal 140:283–290
Liu Q, Zhao G (2012) A comparison of estimation methods for partially linear models. Int J Innov Manag Inf Prod 3:38–42
Lokshin M (2006) Difference-based semiparametric estimation of partial linear regression models. Stata J 6:377–383
Rice JA (1984) Bandwidth choice for nonparametric regression. Ann Stat 12:1215–1230
Schott JR (1997) Matrix analysis for statistics. Wiley, New York
Severini T, Wong W (1992) Generalized profile likelihhood and conditional parametric models. Ann Stat 20:1768–1802
Spechman P (1988) Kernel smoothing in partial linear models. J Roy Stat Soc B 50:413–436
Tabakan G (2013) Performance of the difference-based estimators in partially linear models. Statistics 47:329–347
Tong T, Ma Y, Wang Y (2013) Optimal variance estimation without estimating the meam function. Bernoulli 19:1839–1854
Tong T, Wang Y (2005) Estimating residual variance in nonparametric regression using least squares. Biometrika 92:821–830
Wang L, Brown L, Cai T (2011) A difference based approach to the semiparametric partial linear model. Electron J Stat 5:619–641
Wang W, Lin L (2015) Derivative estimation based on difference sequence via locally weighted least squares regession. J Mach Learn Res 16:2617–2641
Wang W, Lin L, Yu L (2017) Optimal variance estimation based on lagged second-order difference in nonparametric regression. Comput Stat 32:1047–1063
Whittle P (1964) On the convergence to normality of quadratic forms in independent variables. Teor. Verojatnost. i Primenen 9:113–118
Yatchew A (1997) An elementary estimator of the partial linear model. Econ Lett 57:135–143
Zhao H, You J (2011) Difference based estimation for partially linear regression models with measurement errors. J Multivar Anal 102:1321–1338
Zhou Y, Cheng Y, Wang L, Tong T (2015) Optimal difference-based variance estimation in heteroscedastic nonparametric regression. Stat Sin 25:1377–1397