Difference-based variance estimation in nonparametric regression with repeated measurement data

Dai, Wenlin; Ma, Yanyuan; Tong, Tiejun; Zhu, Lixing

Published in:
Journal of Statistical Planning and Inference

DOI:
10.1016/j.jspi.2015.02.010

Published: 01/08/2015

Document Version:
Peer reviewed version

Difference-based Variance Estimation in Nonparametric Regression with Repeated Measurement Data

Wenlin Dai[1], Yanyuan Ma[2], Tiejun Tong[1,*] and Lixing Zhu[1]

[1]Department of Mathematics, Hong Kong Baptist University, Hong Kong

[2]Department of Statistics, Texas A&M University, College Station, TX 77843, USA

March 2, 2015

Abstract

Over the past three decades, interest in cheap yet competitive variance estimators in nonparametric regression has grown tremendously. One family of estimators which has risen to meet the task is the difference-based estimators. Unlike their residual-based counterparts, difference-based estimators do not require estimating the mean function and are therefore popular in practice. This work further develops the difference-based estimators in the repeated measurement setting for nonparametric regression models. Three difference-based methods are proposed for the variance estimation under both balanced and unbalanced repeated measurement settings: the sample variance method, the partitioning method, and the sequencing method. Both their asymptotic properties and finite sample performance are explored. The sequencing method is shown to be the most adaptive while the sample variance method and the partitioning method are shown to outperform in certain cases.

KEY WORDS: Asymptotic normality; Difference-based estimator; Least squares; Nonparametric regression; Repeated measurements; Residual variance.

*Corresponding author. E-mail: tongt@hkbu.edu.hk

1 Introduction

Consider the nonparametric regression model with repeated measurement data,

$$Y_{ij} = f(x_i) + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m, \tag{1}$$

where Y_{ij} are observations, x_i are design points, f is an unknown mean function, and ε_{ij} are independent and identically distributed (i.i.d.) random errors with mean zero and variance σ^2. In this paper we are interested in estimating the residual variance σ^2. Needless to say, an accurate estimate of σ^2 is desired in many situations, e.g., in testing the goodness of fit and in deciding the amount of smoothing (Carroll 1987, Carroll & Ruppert 1988, Eubank & Spiegelman 1990, Gasser, Kneip & Kohler 1991). Over the past three decades, interest in cheap yet competitive variance estimates in the nonparametric setting has grown tremendously. One family of estimators which has generated great interest and has become an important tool for this purpose is the difference-based estimators. Unlike their residual-based counterparts, difference-based estimators do not require the estimation of the mean function, which involves nonparametric estimation procedures, and have therefore become quite popular in practice.

In the simple situation when $m = 1$, there already exist a large body of difference-based estimators in the literature (Dette, Munk & Wagner 1998). In this case, model (1) reduces to

$$Y_i = f(x_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2}$$

where Y_i are observations, and ε_i are i.i.d. random errors with mean zero and variance σ^2. Assume that $0 \le x_1 \le \cdots \le x_n \le 1$, and define the order of a difference-based estimator to be the number of observations involved in calculating a local residual. von Neumann (1941) and Rice (1984) proposed the following first-order estimator,

$$\hat{\sigma}^2_{\mathrm{R}} = \frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_i - Y_{i-1})^2. \tag{3}$$

Gasser, Sroka & Jennen-Steinmetz (1986) and Hall, Kay & Titterington (1990) extended the idea behind the first-order estimator and proposed some higher order difference-based estimators. Dette et al. (1998) pointed out that none of the fixed order difference-based estimators can achieve the same asymptotically optimal rate as that is achieved by the residual-based estimators (Hall & Marron 1990). Müller, Schick & Wefelmeyer (2003), Tong, Liu & Wang (2008) and Du & Schick (2009) proposed covariate-matched U-statistic estimators for the residual variance.

Recently, Tong & Wang (2005) and Tong, Ma & Wang (2013) proposed some least squares methods for estimating the residual variance, motivated by the fact that the Rice estimator (3) is always positively biased. For the equally-spaced design, let

$$\hat{\sigma}^2_{\mathrm{R}}(r) = \frac{1}{2(n-r)} \sum_{i=r+1}^{n} (Y_i - Y_{i-r})^2, \qquad r = 1, 2, \ldots.$$

Assuming that f has a bounded first derivative, they showed that $E\{\hat{\sigma}^2_{\mathrm{R}}(r)\} \doteq \sigma^2 + Jd_r + o(d_r)$, where $d_r = r^2/n^2$ and $J = \int_0^1 \{f'(x)\}^2 dx/2$. To reduce the positive bias Jd_r, they constructed a linear regression model

$$\hat{\sigma}^2_{\mathrm{R}}(r) = \sigma^2 + Jd_r + \xi_r, \quad r = 1, 2, \ldots, r_0, \tag{4}$$

where ξ_r are random errors and $r_0 = o(n)$ is the chosen bandwidth. Let $N = nr_0 - r_0(r_0+1)/2$ as the total number of difference pairs involved in (4). They assigned $w_r = (n-r)/N$ as the weight of $\hat{\sigma}^2_{\mathrm{R}}(r)$, and estimated the residual variance as the intercept through the weighted least squares regression. They further showed that the asymptotic optimal bandwidth is $h_{opt} = \{28n\sigma^4/\mathrm{Var}(\varepsilon^2)\}^{1/2}$ with the corresponding mean squared error (MSE) as

$$\mathrm{MSE}(h_{opt}) = \frac{1}{n}\mathrm{Var}(\varepsilon^2) + \frac{9\sqrt{7}}{28n^{3/2}}\sigma^2\{\mathrm{Var}(\varepsilon^2)\}^{1/2} + o(\frac{1}{n^{3/2}}).$$

When $m > 1$, we have repeated measurements. Repeated measurement data are commonly available in many statistical problems. How to take advantage of the repeated measurements and develop a variance estimator that has the same advantage of not requiring a mean estimation is of great importance. Despite the rich literature on difference-based variance estimation for model (2), very little attention has been paid to model (1) with $m \geq 2$. Gasser et al. (1986) encountered the multiple measurements issue, but they decided to order the data sequentially and treat them as if they came from different design points. Thus, the multiple measurements feature is ignored. This is quite a pity, since intuitively the repeated measurement data contain different type of information, and this new information should be taken into account in constructing estimators. We suspect that one reason very few work is available for treating multiple observations in difference based variance estimation literature is that it is not easy to combine the between-design-point difference and the within-design-point difference properly. In addition, even if a certain new treatment is proposed, it is not straightforward to analyze how effective this treatment is in theory. For example, it is difficult to know if the treatment has optimal large sample property, in other words, it is difficult

to know if a better method can be found in treating the multiple measurements, either within the difference based method family or overall. In this work, we will fill this literature in both aspects. Specifically, we will propose three new difference based methods to utilize the multiple measurements, respectively the sample variance method, the partitioning method and the sequencing method. We analyze these methods and illustrate the practical advantages of each method under different data structures and/or model assumptions. In addition, we will show that one of our proposals, the sequencing method is indeed optimal in that it is root-n consistent and it reaches the minimum asymptotic estimation variability among all possible consistent estimators.

The rest of the paper is organized as follows. In Section 2, we propose three difference-based methods for estimating σ^2 in nonparametric regression with repeated measurement data: the sample variance method, the partitioning method, and the sequencing method. We also explore their asymptotic properties, especially for the proposed sequencing estimator, where we derive its MSE, its optimal bandwidth and its asymptotic normality. In Section 3, we derive the optimal efficiency bound of any estimation procedure and show that the proposed sequencing estimator reaches this universal optimal efficiency bound. Extensive simulation studies are conducted in Section 4 to evaluate and compare the finite sample performance of the proposed estimators to the residual-based estimator. We then extend the methods to the nonparametric regression models with unbalanced repeated measurement data in Section 5. Also, we demonstrate the practical application of proposed methods with one real data example in Section 6. Finally, we conclude the paper in Section 7 with a brief discussion and provide all the technical proofs in the Appendices.

2 Main Results

To estimate σ^2 in model (1), a naive approach is to evade the issue of repeated measurements by taking average of the observations at each design point. Assume that $x_i = i/n$ for all i. Let the averaged observations be

$$\bar{Y_i} = \frac{1}{m}\sum_{j=1}^{m} Y_{ij} = f(x_i) + \frac{1}{m}\sum_{j=1}^{m} \varepsilon_{ij}, \qquad i = 1, \ldots, n.$$

Given that ε_{ij} are i.i.d. random errors with variance σ^2, we have $\mathrm{Var}(\bar{Y_i}) = \sigma^2/m$. Then to estimate σ^2, we multiply the sequence $\bar{Y_i}$ by $\sqrt{m}$ and then apply Tong & Wang (2005)'s method to the new sequence to get the estimation. We name this estimator the averaging estimator, written as $\hat{\sigma}^2_{\text{naive}}$. At the asymptotically optimal

bandwidth $h_{opt} = \{28n\sigma^4/\mathrm{Var}(\varepsilon^2)\}^{1/2}$ in Tong & Wang (2005), the MSE of $\hat{\sigma}^2_{\text{naive}}$ is

$$\mathrm{MSE}\{\hat{\sigma}^2_{\text{naive}}(h_{opt})\} = \frac{1}{n}\mathrm{Var}(\varepsilon^2) + O(n^{-3/2}). \tag{5}$$

The number of repeats m does not appear in (5), hence the naive method clearly does not take advantage of the repeated measurement data. Specifically, by taking averages, this method sacrifices the information contained in the repeated measurement data for simplicity. Further, multiplying the average sequence by $\sqrt{m}$ enlarged the mean function. As a consequence, the trend in the mean function is less negligible in finite sample settings. The analysis on the naive method above indicates that in nonparametric regression with repeated measurement data, there are two types of information we can use and should probably treat differently: (i) the variation within design points, and (ii) the variation between design points.

In what follows, we propose three new methods for estimating σ^2 in nonparametric regression with repeated measurement data. The first method is the sample variance method where only the variation within design points is used. The second method proposed is the partitioning method where only the variation between design points is used. Whereas our third method, the sequencing method, uses both types of variations. The statistical properties of all three methods will be investigated.

2.1 Sample Variance Method

Our first method aims to intelligently use the existence of repeated measurements for the variance estimation. Let $s_i^2 = \sum_{j=1}^{m}(Y_{ij} - \bar{Y}_i)^2/(m-1)$ be the sample variance of the repeated measurements at the ith design point, $i = 1, \ldots, n$. Given that $Y_{i1}, \ldots, Y_{im}$ are i.i.d. random variables, we have $E(s_i^2) = \sigma^2$. Note also that $s_1^2, \ldots, s_n^2$ are independent of each other. We define the sample variance estimator of σ^2 as

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2. \tag{6}$$

It is clear that $\hat{\sigma}_1^2$ is an unbiased estimator of σ^2. By Rose & Smith (2002), we have

$$\mathrm{Var}\left(s_i^2\right) = \frac{1}{m}\mathrm{Var}(\varepsilon^2) + \frac{2}{m(m-1)}\sigma^4.$$

This leads to

$$\mathrm{MSE}(\hat{\sigma}_1^2) = \mathrm{Var}(\hat{\sigma}_1^2) + \mathrm{Bias}^2(\hat{\sigma}_1^2) = \frac{1}{mn}\mathrm{Var}(\varepsilon^2) + \frac{2}{m(m-1)n}\sigma^4.$$

Note that Hall & Marron (1990) showed that the residual-based estimators can achieve an optimal estimation variance $\mathrm{Var}(\varepsilon^2)/(nm)$, hence $\hat{\sigma}_1^2$ is not optimal. When m is large, the discrepancy can be very small. Specifically, when $m \to \infty$, the second term in the above display is negligible and $\hat{\sigma}_1^2$ is asymptotically the best unbiased estimator of σ^2. However, when m is small, the sample variance estimator is clearly suboptimal. In this paper, we are particularly interested in the scenario where m is fixed but n is large.

An especially nice feature of $\hat{\sigma}_1^2$ is that it is completely free from any assumptions on the mean function f. This makes the sample variance estimator robust in the most general nonparametric settings, especially when the mean function is nonsmooth, non-continuous or highly oscillating so that other difference-based methods fail to perform well. Finally, the sample variance estimator is extremely easy to implement in practice.

2.2 Partitioning Method

Our second method is a partitioning method. We first partition the observations Y_{ij} into m groups according to the following sampling-based algorithm.

(i) Sample one observation from the set $\{Y_{i1}, \ldots, Y_{im}\}$ for each i to form the first response group $G(1) = \{Y_{1g_1}, \ldots, Y_{ng_1}\}$.

(ii) Sample one observation from the remaining set $\{Y_{i1}, \ldots, Y_{im}\} \setminus \{Y_{ig_1}\}$ for each i to form the second response group $G(2) = \{Y_{1g_2}, \ldots, Y_{ng_2}\}$.

(iii) Repeat Step (ii), until we obtain the last response group $G(m) = \{Y_{1g_m}, \ldots, Y_{ng_m}\}$.

We then apply Tong & Wang (2005)'s method to each group $G(j)$ to get the estimates $\hat{\sigma}_{(j)}^2$, $j = 1, \ldots, m$. The final estimator is defined as

$$\hat{\sigma}_2^2 = \frac{1}{m} \sum_{j=1}^{m} \hat{\sigma}_{(j)}^2. \tag{7}$$

We refer to it as the partitioning estimator of variance.

Under the model assumption, the groups $G(1), \ldots, G(m)$ are independent of each other. Therefore, the estimators $\sigma_{(j)}^2$ are also independent of each other. Then with the optimal bandwidth $h_{opt} = \{28n\sigma^4/\mathrm{Var}(\varepsilon^2)\}^{1/2}$, the MSE of $\hat{\sigma}_2^2$ is given as

$$\mathrm{MSE}\{\hat{\sigma}_2^2(h_{opt})\} = \frac{1}{nm}\mathrm{Var}(\varepsilon^2) + \frac{9\sqrt{7}}{28mn^{3/2}}\sigma^2\{\mathrm{Var}(\varepsilon^2)\}^{1/2} + o(\frac{1}{n^{3/2}}). \tag{8}$$

This shows that the proposed partitioning estimator achieves the same asymptotically optimal estimation variance as that for the residual-based estimators.

2.3 Sequencing Method

Our third method is proposed to combine two kinds of variation properly. To achieve this, we treat all the repeated observations as if they are generated from different design points. We then build a linear regression model to estimate σ^2 systematically. Specifically, we order the observations as $\{Y_{11}, \ldots, Y_{1m}, \ldots, Y_{n1}, \ldots, Y_{nm}\}$ and also relabel the indices as $l = 1, 2, \ldots, nm$. With this notation, model (1) can be written as

$$Z_l = f(t_l) + \epsilon_l, \qquad l = 1, \ldots, nm, \tag{9}$$

where $\{Z_1, Z_2, \ldots, Z_{nm}\} = \{Y_{11}, \ldots, Y_{1m}, \ldots, Y_{n1}, \ldots, Y_{nm}\}$, $\{t_1, t_2, \ldots, t_{nm}\} = \{x_1, \ldots, x_1, \ldots, x_n, \ldots, x_n\}$, and $\{\epsilon_l, \epsilon_2, \ldots, \epsilon_{nm}\} = \{\varepsilon_{11}, \ldots, \varepsilon_{1m}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{nm}\}$.

For model (9), we define the lag-p Rice estimator

$$\hat{\sigma}^2_{\mathrm{R}}(p) = \frac{1}{2(nm-p)} \sum_{l=p+1}^{nm} (Z_l - Z_{l-p})^2, \qquad \text{for } p = 1, \ldots, nm-1.$$

Note that the first m lag-p Rice estimators only use differences of the identical or consecutive design points, i.e., none of the $f(x_i) - f(x_{i-r})$ terms with $r \geq 2$ are involved in the first m lag-p Rice estimators. We thus combine them and define a new Rice-type estimator using the weighted average of the first m lag-p Rice estimators,

$$\begin{aligned}
\hat{\sigma}^2_{\mathrm{Rt}} &= \frac{1}{m^2 n - m(m+1)/2} \sum_{p=1}^{m} (nm-p) \hat{\sigma}^2_{\mathrm{R}}(p) \\
&= \frac{1}{2m^2 n - m(m+1)} \left\{ \sum_{k=1}^{m-1} \sum_{i=1}^{n} \sum_{j=k+1}^{m} (Y_{ij} - Y_{i,j-k})^2 + \sum_{k=1}^{m} \sum_{i=2}^{n} \sum_{j=1}^{k} (Y_{ij} - Y_{i-1,m-k+j})^2 \right\},
\end{aligned}$$

where the weight for $\hat{\sigma}^2_{\mathrm{R}}(p)$ is assigned because the lag-p Rice estimator uses $(nm-p)$ pairs of data.

Some algebra yields

$$\begin{aligned}
E(\hat{\sigma}^2_{\mathrm{Rt}}) &= \sigma^2 + \frac{1}{2m^2 n - m(m+1)} \sum_{k=1}^{m} \sum_{i=2}^{n} \sum_{j=1}^{k} \{f(x_i) - f(x_{i-1})\}^2 \\
&= \sigma^2 + \frac{m(m+1)/2}{2m^2 n - m(m+1)} \sum_{i=2}^{n} \{f(x_i) - f(x_{i-1})\}^2 .
\end{aligned}$$

This reveals that the Rice-type estimator $\hat{\sigma}^2_{\mathrm{Rt}}$ is always positively biased, unless f is a constant function. Suppose that f has a bounded first derivative. By the Taylor expansion we have

$$E(\hat{\sigma}^2_{\mathrm{Rt}}) = \sigma^2 + \frac{(n-1)m(m+1)}{n^2\{2m^2 n - m(m+1)\}} J + o(\frac{1}{n^2}), \tag{10}$$

where $J = \int_0^1 \{f'(x)\}^2 dx/2$. To eliminate the bias term in (10), we further define the lag-r Rice-type estimators

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{Rt}}(r) &= \frac{1}{c_r}\sum_{p=(r-1)m+1}^{rm}(nm-p)\hat{\sigma}^2_{\mathrm{R}}(p) \qquad (11)\\ &= \frac{1}{2c_r}\left\{\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(Y_{ij}-Y_{i-r+1,j-k})^2+\sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(Y_{ij}-Y_{i-r,m-k+j})^2\right\},\end{aligned}$$

where $r = 1, 2, n-1$, and $c_r = \sum_{p=(r-1)m+1}^{rm}(nm-p) = m^2n - rm^2 + m(m-1)/2$. By definition, $\hat{\sigma}^2_{\mathrm{Rt}} = \hat{\sigma}^2_{\mathrm{Rt}}(1)$. Similar calculation at any fixed $r = o(n)$ yields

$$\begin{aligned}&E\left\{\hat{\sigma}^2_{\mathrm{Rt}}(r)\right\}\\ =\ &\sigma^2+\frac{1}{2c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}\{f(x_i)-f(x_{i-r+1})\}^2+\sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}\{f(x_i)-f(x_{i-r})\}^2\right]\\ =\ &\sigma^2+Jd_r+o(r^2/n^2), \qquad (12)\end{aligned}$$

where

$$d_r = \frac{m\{(m-1)(n-r+1)(r-1)^2+(m+1)(n-r)r^2\}}{2c_rn^2}. \qquad (13)$$

The relation in (12) indicates that the lag-r Rice-type estimator $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ has a linear relationship with the quantity d_r. Taking advantage of this relation, we fit a linear regression model by treating $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ as the response variable and d_r as the covariate, and estimate σ^2 as the intercept of the linear model.

We choose the first b pairs of $\{d_r, \hat{\sigma}^2_{\mathrm{Rt}}(r)\}$ to perform the regression, where $b = o(n)$. The choice of b will be discussed in Sections 2.3.2 and 2.3.3. In performing the linear regression estimation, because $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ involves c_r pairs of data, we assign weight $w_r = c_r/s_b$ to the rth observation, where $s_b = \sum_{r=1}^{b} c_r = m^2nb - m^2b(b+1)/2 + m(m-1)b/2$. The advantage of such weight assignment will be investigated in Section 2.3.5. We then minimize the weighted sum of squares $\sum_{r=1}^{b} w_r\{\hat{\sigma}^2_{\mathrm{Rt}}(r) - \alpha - \beta d_r\}^2$ to fit the linear model

$$\hat{\sigma}^2_{\mathrm{Rt}}(r) = \alpha + \beta d_r + e_r, \qquad r = 1, \ldots, b. \qquad (14)$$

For ease of notation, let $\bar{\sigma}^2_w = \sum_{r=1}^{b} w_r\hat{\sigma}^2_{\mathrm{Rt}}(r)$ and $\bar{d}_w = \sum_{r=1}^{b} w_r d_r$. Then the sequencing estimator of σ^2 is given as

$$\hat{\sigma}^2_3 = \hat{\alpha} = \bar{\sigma}^2_w - \hat{\beta}\bar{d}_w, \qquad (15)$$

where $\hat{\beta} = \sum_{r=1}^{b} w_r\hat{\sigma}^2_{\mathrm{Rt}}(r)(d_r - \bar{d}_w)/\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)^2$ is the fitted slope. In Appendix 1 we prove that

Theorem 1. *For the equally spaced design, $\hat{\sigma}_3^2$ is an unbiased estimator of σ^2 when f is a linear function, regardless of the choice of b.*

In what follows we establish further statistical properties of the sequencing estimator $\hat{\sigma}_3^2$. For notational convenience, we let $\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1m}, \ldots, Y_{n1}, \ldots, Y_{nm})^T$, $\boldsymbol{f} = \{f(x_1), \ldots, f(x_1), \ldots, f(x_n), \ldots, f(x_n)\}^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{1m}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{nm})^T$. Then $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$. Also let $\boldsymbol{u} = (1, \ldots, 1)^T$, $\gamma_i = E(\varepsilon^i/\sigma^i)$ for $i = 3, 4$, and assume that $\gamma_4 > 1$.

2.3.1 Quadratic Form Representation

Let $\tau_0 = 0$ and $\tau_r = 1 - \bar{d}_w(d_r - \bar{d}_w)/\sum_{r=1}^b w_r(d_r - \bar{d}_w)^2$, $r = 1, \ldots, b$. By (15),

$$\hat{\sigma}_3^2 = \sum_{r=1}^{b} \tau_r w_r \hat{\sigma}_{\mathrm{Rt}}^2(r) = \frac{1}{2s_b} \sum_{r=1}^{b} \left\{ \tau_r \sum_{p=(r-1)m+1}^{rm} \sum_{l=p+1}^{nm} (Z_l - Z_{l-p})^2 \right\}.$$

With some algebra, we can write $\hat{\sigma}_3^2$ as

$$\hat{\sigma}_3^2 = \frac{1}{2s_b} \boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y},$$

where $\boldsymbol{D}$ is an $(nm) \times (nm)$ symmetric matrix with elements

$$\boldsymbol{D}_{ij} = \begin{cases} d_{ii}(a), & (a-1)m < i = j \le am \text{ with } a = 1, \ldots, n, \\ -\tau_a, & (a-1)m < |i-j| \le am \text{ with } a = 1, \ldots, b, \\ 0, & \text{otherwise,} \end{cases}$$

where $d_{ii}(a) = m\sum_{r=1}^b \tau_r + m\sum_{r=0}^{a-1} \tau_r + \{i - 1 - (a-1)m\}\tau_a$ for $a = 1, \ldots, b$; $d_{ii}(a) = 2m\sum_{r=1}^b \tau_r$ for $a = b+1, \ldots, n-b$; and $d_{ii}(a) = m\sum_{r=1}^b \tau_r + m\sum_{r=0}^{n-a} \tau_r + (am - i)\tau_{n+1-a}$ for $a = n - b + 1, \ldots, n$.

Note that $\boldsymbol{D}$ depends on the design points only. By letting $f = 0$, we have

$$E(\hat{\sigma}_3^2) = \frac{1}{2s_b} E(\boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y}) = \frac{1}{2s_b} E(\boldsymbol{\varepsilon}^T \boldsymbol{D} \boldsymbol{\varepsilon}) = \frac{\sigma^2}{2s_b} \mathrm{tr}(\boldsymbol{D}),$$

Now because of Theorem 1, $\hat{\sigma}_3^2$ is unbiased when $f = 0$, we have $\mathrm{tr}(\boldsymbol{D}) = 2s_b$. This shows that the proposed sequencing estimator possesses a quadratic form,

$$\hat{\sigma}_3^2 = \boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y} / \mathrm{tr}(\boldsymbol{D}). \tag{16}$$

2.3.2 Asymptotic MSE and Optimal Bandwidth

The quadratic form representation (16) of σ_3^2 enables us to take advantage of the existing results in Dette et al. (1998) and directly obtain

$$\begin{aligned} \mathrm{MSE}(\hat{\sigma}_3^2) &= \left[(\boldsymbol{f}^T \boldsymbol{D} \boldsymbol{f})^2 + 4\sigma^2 \boldsymbol{f}^T \boldsymbol{D}^2 \boldsymbol{f} + 4\boldsymbol{f}^T \{\boldsymbol{D} \cdot \mathrm{diag}(\boldsymbol{D}) \boldsymbol{u}\} \sigma^3 \gamma_3 \right. \\ &\quad \left. + \sigma^4 (\gamma_4 - 3) \mathrm{tr}[\mathrm{diag}(\boldsymbol{D})^2] + 2\sigma^4 \mathrm{tr}(\boldsymbol{D}^2) \right] / \{\mathrm{tr}(\boldsymbol{D})\}^2, \end{aligned} \tag{17}$$

where diag($\boldsymbol{D}$) denotes the diagonal matrix of the diagonal elements of $\boldsymbol{D}$. The first term in (17) represents the squared bias, and the last four terms represent the variance term of the estimator. In the case when the random errors are normally distributed, $\gamma_3 = 0$ and $\gamma_4 = 3$ so that the third and fourth terms vanish.

Theorem 2. *Assume that f has a bounded second derivative. For the equally spaced design with $b \to \infty$ and $b/n \to 0$, we have*

$$\text{Bias}(\hat{\sigma}_3^2) = O(b^3 n^{-3}), \tag{18}$$

$$\text{Var}(\hat{\sigma}_3^2) = \frac{\text{Var}(\varepsilon^2)}{mn} + \frac{9\sigma^4}{4m^2nb} + \frac{9b\text{Var}(\varepsilon^2)}{112mn^2} + o\{(nb)^{-1} + bn^{-2}\}, \tag{19}$$

$$\text{MSE}(\hat{\sigma}_3^2) = \frac{\text{Var}(\varepsilon^2)}{mn} + \frac{9\sigma^4}{4m^2nb} + \frac{9b\text{Var}(\varepsilon^2)}{112mn^2} + o\{(nb)^{-1} + bn^{-2}\} + O(b^6 n^{-6}). \tag{20}$$

Theorem 2 indicates that $\hat{\sigma}_3^2$ is a consistent estimator of σ^2, and its MSE reaches the asymptotically optimal rate (Dette et al. 1998). By (20), the asymptotically optimal bandwidth in terms of minimizing the MSE is given as

$$b_{opt} = \left\{ \frac{28n\sigma^4}{m\text{Var}(\varepsilon^2)} \right\}^{1/2}. \tag{21}$$

It is interesting to point out that b_{opt} does not depend on the mean function f. We also note that b_{opt} is a decreasing function of m. Substituting (21) into (20) leads to

$$\text{MSE}\{\hat{\sigma}_3^2(b_{opt})\} = \frac{1}{nm}\text{Var}(\varepsilon^2) + \frac{9\sqrt{7}}{28m^{3/2}n^{3/2}}\sigma^2\{\text{Var}(\varepsilon^2)\}^{1/2} + o(1/n^{3/2}). \tag{22}$$

Comparing (8) and (22), we have $\text{MSE}\{\hat{\sigma}_3^2(b_{opt})\} < \text{MSE}\{\hat{\sigma}_2^2(h_{opt})\}$ for any $m \geq 2$. This implies that the sequencing estimator behaves asymptotically better than the partitioning estimator in the presence of repeated measurement data. Note also that $b_{opt} = h_{opt}/m^{1/2}$. When $m = 1$, $b_{opt} = h_{opt}$ and the two estimators are identical.

2.3.3 Adaptive Choice of Bandwidth

For simplicity, we use normal random errors to illustrate the choice of bandwidth in the finite sample situation. When the errors are not normal, the only additional complexity is to estimate the ratio $\gamma_4 = \text{Var}(\varepsilon^2)/\sigma^4$; all other aspects of the bandwidth selection procedure remain the same as in the normal error case.

For normal random errors, $\text{Var}(\varepsilon^2) = 2\sigma^4$ so that b_{opt} is simplified as $(14n/m)^{1/2}$, which does not depend on the smoothness of the mean function and the magnitude of residual variance. We caution here that the above b_{opt} applies for large n only. When

n is small or when f is rough, the performance of b_{opt} is sometimes not satisfactory in practice. This was also observed in Tong & Wang (2005) for $m = 1$. This is because some higher order terms ignored in the calculation of the asymptotic MSE for the estimator (20) indeed depend on the smoothness of the function. Consequently, we need a smaller bandwidth to diminish the impact of the mean function in the finite sample case. Simulation studies (not shown) indicate that the bandwidth choices for $m = 1$ in Tong & Wang (2005) often work well for $m \geq 2$, as long as m is not too large (say, $m \leq 20$). In summary, we suggest to use (i) $b_s = n^{1/2}$ for large n, and (ii) $b_t = n^{1/3}$ for small n or for rough f. In the remainder of this article, we take the integer part of b_s and b_t whenever necessary.

A cross validation (CV) strategy can also be applied to select the bandwidth. Specifically, we first split the whole data set into V disjoint subsamples $\{S_1, \ldots, S_V\}$, and then select $b = b_{\mathrm{CV}}$ that minimizes $\mathrm{CV}(b) = \sum_{v=1}^{V} \{\hat{\sigma}_3^2(b) - \hat{\sigma}_{3,v}^2(b)\}^2$, where $\hat{\sigma}_3^2(b)$ and $\hat{\sigma}_{3,v}^2(b)$ are the estimates of σ^2 based on the whole sample $\cup_{i=1}^{V} S_i$ and the subsample $\cup_{i \neq v} S_i$ with bandwidth b, respectively. Note that the design points in $\cup_{i \neq v} S_i$ are not equally spaced on $[0, 1]$. Thus to compute $\hat{\sigma}_{3,v}^2(b)$, we need to use the formula developed in the general design; see more details in Section 2.3.6. Finally, the CV method requires much more expensive computation compared to b_s and b_t.

2.3.4 Asymptotic Normality

We have the following asymptotic normality for the Rice-type estimators $\hat{\sigma}_{\mathrm{Rt}}^2(r)$ in (11) and for the sequencing estimator $\hat{\sigma}_3^2$ in (16). Let $\xrightarrow{\mathcal{D}}$ denote convergence in distribution.

Theorem 3. *Assume that f has a bounded second derivative and $E(\varepsilon^4)$ is finite. Then for any $r = n^{\vartheta}$ with $0 \leq \vartheta < 1/2$, the lag-r Rice-type estimator satisfies*

$$\sqrt{n}\{\hat{\sigma}_{Rt}^2(r) - \sigma^2\} \xrightarrow{\mathcal{D}} N\{0, (\gamma_4 - 1 + 1/m)\sigma^4/m\} \qquad \text{as } n \to \infty.$$

Theorem 4. *Assume that f has a bounded second derivative and $E(\varepsilon^{4+2\delta})$ is finite for some δ in $(0, 1)$. Then for any $b = n^{\vartheta}$ with $0 < \vartheta < 1/2$, the sequencing estimator $\hat{\sigma}_3^2$ satisfies*

$$\sqrt{n}(\hat{\sigma}_3^2 - \sigma^2) \xrightarrow{\mathcal{D}} N\{0, (\gamma_4 - 1)\sigma^4/m\} \qquad \text{as } n \to \infty.$$

Proofs of Theorems 3 and 4 are given in Appendices 3 and 4, respectively. Theorem 3 indicates that $\hat{\sigma}_{\mathrm{Rt}}^2(r)$ has the same asymptotic property as "the m-order optimal difference-based estimator" proposed in Hall et al. (1990). Given that $\mathrm{E}(\varepsilon^{4+2\delta})$ is finite

for some δ in $(0,1)$, Theorems 3 and 4 show that the sequencing estimator is more efficient than the Rice-type estimators for any fixed m. Specifically, the efficiency of $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ relative to $\hat{\sigma}^2_3$ is given as $(\gamma_4-1)/(\gamma_4-1+1/m)$, which is an increase function of m. When the random errors are normally distributed, the relative efficiency reduces to $2m/(2m+1)$. As illustration, the relative efficiency is 66.7% when $m=1$, 80% when $m=2$, and 90.9% when $m=5$. Finally, $\hat{\sigma}^2_{\mathrm{Rt}}(r)$ and $\hat{\sigma}^2_3$ become asymptotically equivalent as $m\to\infty$.

Theorem 4 can be easily used to construct confidence intervals for σ^2. For example, when $mn > (\gamma_4-1)z^2_{\alpha/2}$, an approximate $1-\alpha$ confidence interval for σ^2 is

$$[\hat{\sigma}^2_3/\{1+z_{\alpha/2}\sqrt{(\gamma_4-1)/mn}\},\ \hat{\sigma}^2_3/\{1-z_{\alpha/2}\sqrt{(\gamma_4-1)/mn}\}],$$

where z_α is the upper α-th percentile of the standard normal distribution. For normal data, the parameter $\gamma_4=3$ so the confidence interval is fully specified. In general, γ_4 needs to be replaced by an estimate.

2.3.5 Generalized Least Squares Estimator

In constructing the Rice-type estimators at different lags, we have used the same observations to form different pairs. Thus, our linear regression model (14) concerns correlated data. When the responses are correlated, the proper way of performing linear regression is the generalized least squares (GLS) method, where the optimal weighting matrix is the inverse variance-covariance matrix of the observations. In our problem, the variance-covariance matrix is found to have very special property. Specifically, in Appendix 5 we prove the following results.

Lemma 1. *Assume that f has a bounded second derivative and $E(\varepsilon^4)$ is finite. Then for any $b=n^\vartheta$ with $0\le\vartheta<1/2$, the variance-covariance matrix of $\{\hat{\sigma}^2_{Rt}(1),\ldots,\hat{\sigma}^2_{Rt}(b)\}$ has leading order $\Sigma=(\sigma_{pr})_{b\times b}$, where $\sigma_{pp}=(\gamma_4-1+1/m)\sigma^4/(mn)$ for any $1\le p\le b$ and $\sigma_{rp}=\sigma_{pr}=(\gamma_4-1)\sigma^4/(mn)$ for any $1\le r<p\le b$.*

Lemma 1 states that the leading order of the variance-covariance matrix has the same value on the diagonal, even though each diagonal element corresponds to a different lag. In addition, the off-diagonal elements are also identical, hence the matrix Σ is compound symmetric. These properties yield great simplification of GLS. Specifically, let $\boldsymbol{z}=\{\hat{\sigma}^2_{\mathrm{Rt}}(1),\ldots,\hat{\sigma}^2_{\mathrm{Rt}}(r)\}^T$, $\boldsymbol{\beta}=(\alpha,\beta)^T$, $\boldsymbol{e}=(e_1,\ldots,e_b)^T$, $\boldsymbol{d}=(d_1,\ldots,d_b)^T$, and $X=(\boldsymbol{u},\boldsymbol{d})$ be the design matrix. With these notations, the linear model (14) is

equivalent to $\boldsymbol{z} = X\boldsymbol{\beta} + \boldsymbol{e}$, and to the first order, the optimal GLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\boldsymbol{z}.$$

From Lemma 1, we have $\Sigma = (\gamma_4 - 1 + 1/m)\sigma^4\{(1-\rho)I + \rho\boldsymbol{u}\boldsymbol{u}^T\}/(mn)$, where $\rho = (\gamma_4 - 1)/(\gamma_4 - 1 + 1/m)$ and I is the identity matrix. Due to the compound symmetry structure of Σ and the fact that the first column of X is $\boldsymbol{u}$, it is not difficult to show that $(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\boldsymbol{z} = (X^TX)^{-1}X^T\boldsymbol{z}$ (McElroy (1967) and Kariya & Kurata (2004)). This implies that the optimal GLS estimator $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is in fact the same as the ordinary least squares estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ to the first order. In other words, the simplest OLS is already the most efficient way of perform the linear regression.

However, the sequencing method we proposed in Section 2.3 is not the optimal GLS or OLS. In fact, the estimator $\hat{\sigma}_3^2$ is a GLS with a special weighting strategy. Specifically, let $W = \text{diag}(w_1, \ldots, w_b)$ be the weight matrix and write the weighted least squares (WLS) estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\hat{\alpha}_{\text{WLS}}, \hat{\beta}_{\text{WLS}})^T = (X^TW^{-1}X)^{-1}X^TW^{-1}\boldsymbol{z}.$$

Then the sequencing estimator corresponds to the intercept estimation of $\hat{\boldsymbol{\beta}}_{\text{WLS}}$, i.e., $\hat{\sigma}_3^2 = \hat{\alpha}_{\text{WLS}}$. It is not difficult to see that when $n \to \infty$, the weights w_i converge to a constant uniformly. Thus, $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ is asymptotically the same as $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ and hence is also optimal. The reason we propose WLS instead of the simplest OLS to form the sequencing estimator is based on small sample consideration. When n is not too large, WLS takes into account the higher order difference of the variabilities at different lags hence it adapts better to the data and tends to have more stable numerical performance.

3 Optimal Bound for Estimating σ^2

We now consider the optimal bound on the variance in estimating σ^2 regardless how the estimation is carried out. We only assume that the regression errors $\epsilon_{ij} = Y_{ij} - f(X_i), i = 1, \ldots, n, j = 1, \ldots, m$ are independent and identically distributed with mean 0 and variance σ^2, and are independent of X_i's. Denote the probability density function of ϵ_{ij} as $\eta(\epsilon_{ij})$.

The probability density function (pdf) of $(x_i, y_{i1}, \ldots, y_{im})$ is $f_X(x_i)\prod_{j=1}^m \eta(\epsilon_{ij})$, where $f_X(\cdot)$ is the marginal pdf of X_i and η is a pdf that ensures $E(\epsilon_{ij}) = 0$. Because $\sigma^2 = E(\epsilon_{ij}^2)$ is our parameter of interest, f_X, η, f are pure nuisance parameters. This leads us to a semiparametric problem and the semiparametric tools developed in Bickel et. al (1994) can be readily applied to derive the efficient influence function

through projecting an arbitrary influence function onto the tangent space associated with f_X, η and f.

Following standard calculation, the tangent space is

$$\begin{aligned}\Lambda_{\mathcal{T}} &= \left\{ h(x_i) + \sum_{j=1}^{m} g(\epsilon_{ij}) + \sum_{j=1}^{m} \frac{\eta_0'(\epsilon_{ij})}{\eta_0(\epsilon_{ij})} a(x_i) \right. \\ &\quad \left. : \forall h, g \text{ such that } E(h) = 0, E(g) = E(\epsilon_{ij} g) = 0, \forall a \right\}, \end{aligned} \tag{23}$$

where $\eta_0(\cdot)$ is the true probability density function of ϵ_{ij}. We identify $m^{-1}\sum_{j=1}^{m}\epsilon_{ij}^2 - \sigma^2$ as one valid influence function. To see this, consider an arbitrary parametric submodel, denoted $\eta(\epsilon_{ij}, \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is a finite dimensional vector of parameters such that there exists $\boldsymbol{\mu}_0$, so that $\eta(\epsilon_{ij}, \boldsymbol{\mu}_0) = \eta_0(\epsilon_{ij})$. In addition, $\eta(\epsilon_{ij}, \boldsymbol{\mu})$ is a valid probability density function and $\int \epsilon_{ij}\eta(\epsilon_{ij};\boldsymbol{\mu})d\epsilon_{ij} = 0$ for all $\boldsymbol{\mu}$ in a local neighborhood of $\boldsymbol{\mu}_0$. We have $\partial \int \epsilon_{ij}^2 \eta(\epsilon_{ij}, \boldsymbol{\mu})d\epsilon_{ij}/\partial\boldsymbol{\mu}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} = \int \epsilon_{ij}^2 \frac{\partial \log \eta(\epsilon_{ij},\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \eta(\epsilon_{ij}, \boldsymbol{\mu})d\epsilon_{ij}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0}$. Consequently,

$$\begin{aligned}\frac{\partial\sigma^2}{\partial\boldsymbol{\mu}}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} &= m^{-1}\frac{\partial}{\partial\boldsymbol{\mu}}\int (\sum_{j=1}^{m}\epsilon_{ij}^2)\prod_{j=1}^{m}\eta(\epsilon_{ij},\boldsymbol{\mu})d\epsilon_{i1}\ldots d\epsilon_{im}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} \\ &= m^{-1}\sum_{j=1}^{m}\frac{\partial}{\partial\boldsymbol{\mu}}\int \epsilon_{ij}^2\eta(\epsilon_{ij},\boldsymbol{\mu})d\epsilon_{ij}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} \\ &= m^{-1}\sum_{j=1}^{m}\int \epsilon_{ij}^2\frac{\partial\log\eta(\epsilon_{ij},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}\eta(\epsilon_{ij},\boldsymbol{\mu})d\epsilon_{ij}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} \\ &= m^{-1}\sum_{j=1}^{m}\int \epsilon_{ij}^2\frac{\partial\log\prod_{j=1}^{m}\eta(\epsilon_{ij},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}\prod_{j=1}^{m}\eta(\epsilon_{ij},\boldsymbol{\mu})d\epsilon_{i1}\ldots d\epsilon_{im}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} \\ &= E(m^{-1}\sum_{j=1}^{m}\epsilon_{ij}^2 S_{\boldsymbol{\mu}}),\end{aligned}$$

where $S_{\boldsymbol{\mu}} \equiv \partial\log\prod_{j=1}^{m}\eta(\epsilon_{ij},\boldsymbol{\mu})/\partial\boldsymbol{\mu}$ is the score with respect to $\boldsymbol{\mu}$. From Section 4 of Tsiatis (2006), $m^{-1}\sum_{j=1}^{m}\epsilon_{ij}^2 - \sigma^2$ is thus a valid influence function. Writing

$$m^{-1}\sum_{j=1}^{m}\epsilon_{ij}^2 - \sigma^2 = m^{-1}\sum_{j=1}^{m}\{\epsilon_{ij}^2 - \sigma^2 + \gamma_3\sigma^3\eta_0'(\epsilon_{ij})/\eta_0(\epsilon_{ij})\} - m^{-1}\sum_{j=1}^{m}\gamma_3\sigma^3\eta_0'(\epsilon_{ij})/\eta_0(\epsilon_{ij}),$$

we can easily verify that $\epsilon_{ij}^2 - \sigma^2 + \gamma_3\sigma^3\eta_0'(\epsilon_{ij})/\eta_0(\epsilon_{ij})$ satisfies the requirement on g in (23). Letting $h = 0$ and $a(x_i) = -\gamma_3\sigma^3$, we can see that $m^{-1}\sum_{j=1}^{m}\epsilon_{ij}^2 - \sigma^2$ is an element of the tangent space, hence it is in fact the efficient influence function. The corresponding efficient estimation variance is $n^{-1}E\{(m^{-1}\sum_{j=1}^{m}\epsilon_{ij}^2 - \sigma^2)^2\} = (nm)^{-1}E\{(\epsilon_{ij}^2 - \sigma^2)^2\} = (nm)^{-1}(\gamma_4 - 1)\sigma^4$, which agrees with the asymptotic estimation

variance established in Theorem 4. This shows that the proposed sequencing estimator is indeed optimal in terms of its first order estimation variance among the class of all consistent estimators.

In the above derivation, we did not take into consideration that X_i's are actually equally spaced instead of being random. However, assuming $f_X(x)$ to be uniform or more generally assuming $f_X(x)$ to have any particular form does not change the efficiency result. This is because the efficiency bound calculation is conducted conditional on X, and is decided by the property of ϵ only, which is assumed to be independent of x_i. The property of f_X is thus masked out.

4 Simulation Studies

We now conduct simulation studies to evaluate the finite sample performance of the aforementioned estimators: the naive estimator $\hat{\sigma}^2_{\text{naive}}$, the sample variance estimator $\hat{\sigma}^2_1$, the partitioning estimator $\hat{\sigma}^2_2$, and the sequencing estimator $\hat{\sigma}^2_3$. For comparison, we also include a residual-based estimator, where we use the cubic smoothing spline to estimate the mean function and then use the squared residuals to estimate the variance. During the procedure, the smoothing parameter is selected via the generalized cross validation, and the resulting variance estimator is written as $\hat{\sigma}^2_{\text{SS}}$.

We consider the following two mean functions:

$$f_1(x) = 10x(1-x),$$
$$f_2(x) = 3x\sin(4\pi x),$$

where f_1 is a low-frequency function and f_2 is an irregular high-frequency function (see Figure 1). The coefficients 10 in f_1 and 3 in f_2 are chosen so that the two mean functions have similar amplitudes. We set the design points $x_i = i/n$ and simulate ε_{ij} independently from $N(0, \sigma^2)$. For each mean function, we consider $n = 30$ and 200 corresponding to small and large sample sizes respectively, and $\sigma^2 = 0.25$ and 4 corresponding to small and large variances respectively. Further, we choose $m = 2, 3, 4, 5$ and 10 to represent different levels of repeated measurements. In total, we have 40 combinations of simulation settings.

We choose the bandwidths $b_s = n^{1/2}$ and $b_t = n^{1/3}$ for both $\hat{\sigma}^2_2$ and $\hat{\sigma}^2_3$. The corresponding estimators are referred to as $\hat{\sigma}^2_2(b_t)$, $\hat{\sigma}^2_2(b_s)$, $\hat{\sigma}^2_3(b_t)$ and $\hat{\sigma}^2_3(b_s)$ respectively. The CV method can also be used for estimating σ^2, and we find it generally performs as well as b_t and b_s. However, because CV is computationally more expensive, we do not recommend it and hence do not present its corresponding results in the remainder

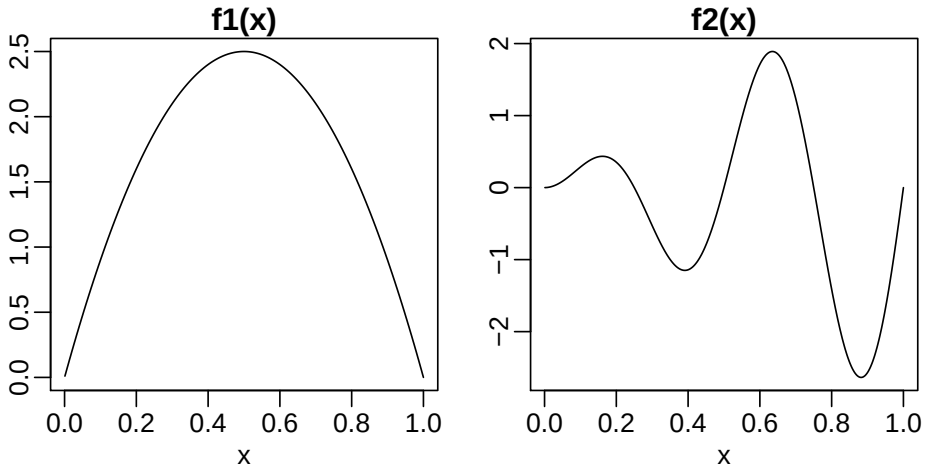

Figure 1: The mean functions $f_1(x)$ and $f_2(x)$, where $0 \leq x \leq 1$.

of the paper. For $\hat{\sigma}^2_{\text{naive}}$, we use the bandwidth $b_t = n^{1/3}$ throughout the simulations. Also note that the quadratic matrix $\boldsymbol{D}$ is not guaranteed to be positive definite. This means that $\hat{\sigma}^2_{\text{naive}}$, $\hat{\sigma}^2_2(b_t)$, $\hat{\sigma}^2_2(b_s)$, $\hat{\sigma}^2_3(b_t)$ and $\hat{\sigma}^2_3(b_s)$ may take negative estimates, though it happens very rarely in our simulations. We replace negative estimates by zero in the calculation of the relative mean squared errors.

We repeat the simulation $10,000$ times for each setting. The relative mean squared errors, $(mn)\text{MSE}/(2\sigma^4)$, are reported in Table 1 for $n = 30$ and in Table 2 for $n = 200$. Based on the simulation results, we summarize the findings below. (i) The sequencing estimator $\hat{\sigma}^2_3(b_s)$ or $\hat{\sigma}^2_3(b_t)$ exhibits the best performance in all but one setting; it even outperforms the residual-based estimator $\hat{\sigma}^2_{\text{SS}}$ when an appropriate bandwidth is used. (ii) The relative performance of $\hat{\sigma}^2_3(b_s)$ and $\hat{\sigma}^2_3(b_t)$ depends on the smoothness of f, the sample size n and the signal-to-noise ratio. In general, $\hat{\sigma}^2_3(b_s)$ performs slightly better than $\hat{\sigma}^2_3(b_t)$ for most settings; whereas for small n and rough f, $\hat{\sigma}^2_3(b_t)$ is much better than $\hat{\sigma}^2_3(b_s)$. (iii) The sequencing estimator always performs better than the partitioning estimator. Specifically, $\hat{\sigma}^2_3(b_s)$ always outperforms $\hat{\sigma}^2_2(b_s)$ and $\hat{\sigma}^2_3(b_t)$ always outperforms $\hat{\sigma}^2_2(b_t)$. (iv) The sample variance estimator $\hat{\sigma}^2_1$ does not suffer from the bias term caused by the lack of smoothness of f and the large signal-to-noise ratio. As a consequence, it outperforms all other methods when n is small (30), σ^2 is small (0.25) and f is rough (f_2). (v) The naive estimator $\hat{\sigma}^2_{\text{naive}}$ is always the worst among all the estimators. (vi) When m increases, all the proposed estimators, except the naive estimator, have a decreased relative MSE. In particular, the MSE of $\hat{\sigma}^2_1$ decreases dramatically as m increases. When $m = 10$, $\hat{\sigma}^2_1$ always performs well and is among the best of all the estimators. This demonstrates again the importance of extracting the

information contained in the repeated measurement data.

f	σ^2	m	$\hat{\sigma}^2_{\text{naive}}$	$\hat{\sigma}^2_1$	$\hat{\sigma}^2_2(b_t)$	$\hat{\sigma}^2_2(b_s)$	$\hat{\sigma}^2_3(b_t)$	$\hat{\sigma}^2_3(b_s)$	$\hat{\sigma}^2_{\text{SS}}$
f_1	0.25	2	3.17	2.07	1.57	1.45	1.31	**1.28**	1.80
		3	4.92	1.55	1.61	1.50	**1.24**	1.25	1.58
		4	6.84	1.37	1.64	1.57	**1.19**	1.24	1.42
		5	8.67	1.25	1.61	1.58	**1.15**	1.22	1.31
		10	21.15	1.14	1.64	1.79	**1.12**	1.28	1.17
	4	2	3.09	2.07	1.55	1.33	1.28	**1.18**	1.53
		3	4.65	1.55	1.58	1.33	1.21	**1.14**	1.37
		4	6.29	1.37	1.61	1.35	1.16	**1.11**	1.26
		5	7.70	1.25	1.57	1.32	1.12	**1.07**	1.19
		10	15.83	1.14	1.59	1.34	1.08	**1.07**	1.14
f_2	0.25	2	8.79	**2.07**	2.98	16.42	2.18	11.08	2.69
		3	22.84	**1.55**	3.63	23.71	2.21	13.63	2.17
		4	49.36	**1.37**	4.30	31.03	2.29	16.28	1.90
		5	90.14	**1.25**	4.90	38.27	2.37	18.97	1.71
		10	658.93	**1.14**	8.02	74.46	2.98	33.04	1.42
	4	2	3.17	2.07	1.57	1.43	1.30	**1.26**	2.16
		3	4.78	1.55	1.60	1.46	1.23	**1.22**	1.77
		4	6.62	1.37	1.63	1.49	**1.18**	1.19	1.55
		5	8.21	1.25	1.60	1.51	**1.14**	1.17	1.43
		10	19.23	1.14	1.62	1.66	**1.10**	1.21	1.27

Table 1: Relative mean squared errors for the seven estimators under various settings with $n = 30$.

5 Nonparametric Regression with Unbalanced Repeated Measurements

In this section, we consider the nonparametric regression model with unbalanced repeated measurements,

$$Y_{ij} = f(x_i) + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m_i,$$

where Y_{ij}, x_i, f and ε_{ij} are defined as before. We assume that m_i are not all the same, where $m_i = 1$ represents a single observation at the ith design point.

f	σ^2	m	$\hat{\sigma}^2_{\text{naive}}$	$\hat{\sigma}^2_1$	$\hat{\sigma}^2_2(b_t)$	$\hat{\sigma}^2_2(b_s)$	$\hat{\sigma}^2_3(b_t)$	$\hat{\sigma}^2_3(b_s)$	$\hat{\sigma}^2_{\text{SS}}$
f_1	0.25	2	2.55	2.02	1.29	1.12	1.15	**1.07**	1.13
		3	3.89	1.48	1.27	1.08	1.07	**1.03**	1.06
		4	5.08	1.31	1.27	1.11	1.06	**1.03**	1.06
		5	6.24	1.25	1.27	1.11	1.05	**1.03**	1.05
		10	12.86	1.10	1.24	1.08	**1.01**	**1.01**	1.02
	4	2	2.55	2.02	1.29	1.11	1.15	**1.07**	1.10
		3	3.89	1.48	1.27	1.08	1.07	**1.02**	1.05
		4	5.08	1.31	1.27	1.10	1.06	**1.03**	1.05
		5	6.23	1.25	1.27	1.10	1.05	**1.03**	1.04
		10	12.83	1.10	1.24	1.07	1.01	**1.00**	1.01
f_2	0.25	2	2.56	2.02	1.29	1.36	**1.15**	1.28	1.40
		3	3.90	1.48	1.27	1.45	**1.08**	1.31	1.25
		4	5.13	1.31	1.28	1.59	**1.07**	1.41	1.21
		5	6.30	1.25	1.27	1.70	**1.06**	1.48	1.17
		10	13.24	1.10	1.24	2.23	**1.01**	1.83	1.09
	4	2	2.55	2.02	1.28	1.11	1.15	**1.07**	1.23
		3	3.89	1.48	1.27	1.08	1.08	**1.02**	1.14
		4	5.08	1.31	1.27	1.11	1.06	**1.03**	1.12
		5	6.23	1.25	1.27	1.10	1.05	**1.03**	1.10
		10	12.83	1.10	1.24	1.08	1.01	**1.00**	1.05

Table 2: Relative mean squared errors for the seven estimators under various settings with $n = 200$.

5.1 Methodology

We first point out that when m_i's are not identical, the averaged observations $\bar{Y}_i$ no longer have homogeneous variances. Instead, $\mathrm{Var}(\bar{Y}_i) = \sigma^2/m_i$. Consequently, the naive method is no longer applicable for unbalanced repeated measurements, simply because Tong & Wang (2005) does not apply to heterogeneous variances. For the other three proposed methods, we derive the following corresponding results. Their numerical performance will be studied in Section 5.2.

The sample variance method still yields a valid estimator. For the ith design point, let $s_i^2 = \sum_{j=1}^{m_i}(Y_{ij} - \bar{Y}_i)^2/(m_i - 1)$ when $m_i \geq 2$ and $s_i^2 = 0$ when $m_i = 1$. The sample variance estimator is then

$$\tilde{\sigma}_1^2 = \frac{1}{M-n}\sum_{i=1}^{n}(m_i - 1)s_i^2,$$

where $M = \sum_{i=1}^{n} m_i$. We note that $\tilde{\sigma}_1^2$ is an unbiased estimator for σ^2. In the special case when $m_i \geq 2$ and are identical, $\tilde{\sigma}_1^2$ reduces to $\hat{\sigma}_1^2$.

The partitioning method continues to work for unbalanced repeated measurements with slight modification. First, we sample one observation Y_{1g_1} from the set $\{Y_{11}, \ldots, Y_{1m_1}\}$, one observation Y_{2g_1} from the set $\{Y_{21}, \ldots, Y_{2m_2}\}$, $\ldots$, and one observation Y_{ng_1} from the set $\{Y_{n1}, \ldots, Y_{nm_n}\}$. Second, we apply Tong & Wang (2005)'s method on the selected group $G(1) = \{Y_{1g_1}, \ldots, Y_{ng_1}\}$ to get one estimate $\hat{\sigma}_{(1)}^2$. We then repeat the process B times and estimate σ^2 by

$$\tilde{\sigma}_2^2 = \frac{1}{B}\sum_{j=1}^{B}\hat{\sigma}_{(j)}^2.$$

Unlike the partitions in Section 2.2, the groups $G(1), \ldots, G(B)$ are not fully separated so that the estimators $\sigma_{(1)}^2, \ldots, \sigma_{(B)}^2$ may not be independent of each other. In the special case when m_i are all the same, it can be shown that $\tilde{\sigma}_2^2$ and $\hat{\sigma}_2^2$ are asymptotically equivalent as $B \to \infty$. In general, the larger the B value, the closer performance between $\hat{\sigma}_2^2$ and $\tilde{\sigma}_2^2$. We suggest to choose a B value that is at least larger than $\max\{m_1, \ldots, m_n\}$ in practice.

The sequencing method can also be adjusted to apply for unbalanced repeated measurements. Let $d_{i_1 i_2} = (x_{i_2} - x_{i_1})^2$ be the squared distances between design points x_{i_2} and x_{i_1}, and $S_{i_1 i_2} = \{s_{i_1(j_1)i_2(j_2)} = (Y_{i_2 j_2} - Y_{i_1 j_1})^2/2 : j_1 = 1, \ldots, m_{i_1},\ j_2 = 1, \ldots, m_{i_2}\}$ be the set of size $m_{i_1}m_{i_2}$ for the half squared differences associated with $d_{i_1 i_2}$. We collect all $d_{i_1 i_2}$ values so that $d_{i_1 i_2} \leq (b/n)^2$, and let $A = \{(i_1, i_2) : d_{i_1 i_2} \leq (b/n)^2,\ 1 \leq i_1 < i_2 \leq n\}$. Correspondingly, we collect all the $s_{i_1(j_1)i_2(j_2)}$ values for each $(i_1, i_2) \in A$. Now

for each paired data $\{(d_{i_1i_2}, s_{i_1(j_1)i_2(j_2)}) : (i_1, i_2) \in A,\ j_1 = 1, \ldots, m_{i_1},\ j_2 = 1, \ldots, m_{i_2}\}$, we fit a simple regression model $s_{i_1(j_1)i_2(j_2)} = \alpha + d_{i_1i_2}\beta + \eta_{i_1i_2}$ by least squares and then estimate σ^2 by the fitted intercept,

$$\tilde{\sigma}_3^2 = \tilde{\alpha} = \frac{1}{NT_2 - T_1^2} \sum_{(i_1,i_2)\in A} (T_2 - T_1 d_{i_1i_2}) \sum_{j_1=1}^{m_{i_1}} \sum_{j_2=1}^{m_{i_2}} s_{i_1(j_1)i_2(j_2)}.$$

where $N = \sum_A m_{i_1} m_{i_2}$, $T_1 = \sum_A m_{i_1} m_{i_2} d_{i_1i_2}$ and $T_2 = \sum_A m_{i_1} m_{i_2} d_{i_1i_2}^2$. When m_i are all the same, it is easy to verify that $\tilde{\sigma}_3^2$ is equivalent to a least squares estimator with

$$\tilde{s}_r = \frac{1}{2m^2(n-r)} \sum_{i=r+1}^{n} \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} (Y_{i,j_2} - Y_{i-r,j_1})^2$$

as the dependent variable and $\tilde{d}_r = r^2/n^2$ as the independent variable. Our analytical and simulation studies (not shown) indicate that under equally spaced and balanced design, $\tilde{\sigma}_3^2$ and $\hat{\sigma}_3^2$ are equivalent asymptotically and similar in finite sample performance. Note that $\tilde{\sigma}_3^2$ also works for unequally spaced designs. In view of this, we claim that $\tilde{\sigma}_3^2$ generalizes the sequencing estimator $\hat{\sigma}_3^2$ not only from balanced repeated measurements to unbalanced repeated measurements, but also from equally spaced designs to unequally spaced designs.

5.2 A Simulation Study

We now study the finite sample performance of the proposed estimators under the unbalanced repeated measurement setting. The estimators considered for comparison are $\tilde{\sigma}_1^2$, $\tilde{\sigma}_2^2(b_t)$, $\tilde{\sigma}_2^2(b_s)$, $\tilde{\sigma}_3^2(b_t)$, $\tilde{\sigma}_3^2(b_s)$, and $\hat{\sigma}_{\rm SS}^2$, where B is set to be 50 for the estimator $\tilde{\sigma}_2^2$. We consider the mean functions $f_1(x)$ and $f_2(x)$, the sample sizes $n = 30$ and $n = 200$, and the residual variances $\sigma^2 = 0.25$ and 4 as in Section 4. The design points are $x_i = i/n$ and ε_{ij} are simulated independently from $N(0, \sigma^2)$. For the different measurements repetitions, we set $m_i = r$ if $i = 5k + r$, where k is a non-negative integer and r is an integer in [1, 5]. In total, there are a total of $3n$ observations.

We repeat the simulation 10,000 times for each setting, and report in Table 3 the relative mean squared errors, i.e., $(3n)\text{MSE}/(2\sigma^4)$. Based on the simulation results, we summarize the following findings. First, $\tilde{\sigma}_3^2(b_s)$ or $\tilde{\sigma}_3^2(b_t)$ performs the best in all but one settings, where f is rough (f_2), σ^2 is small (0.25) and n is small. In this case, $\tilde{\sigma}_1^2$ works the best. The comparative performance of $\tilde{\sigma}_3^2(b_s)$ and $\tilde{\sigma}_3^2(b_t)$ is similar to that of $\hat{\sigma}_3^2(b_s)$ and $\hat{\sigma}_3^2(b_t)$ in the balanced repeated measurement setting. Second, the sequencing estimator always outperforms the partitioning estimator, regardless what

bandwidth is used. Third, the finite sample performance of the sequencing estimator is superior to that of the residual-based estimator in all settings except for the case $(n, f, \sigma^2) = (30, f_2, 0.25)$.

Finally, it is interesting to compare the simulation results in Table 3 with those for the setup $m = 3$ in Tables 1 and 2, where $m = 3$ reflects the average number of repeated measurements. (i) For the sample variance method, we note that $\tilde{\sigma}_1^2$ works as well as or even better than $\hat{\sigma}_1^2$ in most settings. This indicates that $\tilde{\sigma}_1^2$ successfully adapts to the unbalanced measurements by putting more weights (i.e., $m_i - 1$) on the more accurate sample variances obtained from large m_i values. (ii) The partitioning method is less efficient when the repeated measurement data are unbalanced. We believe that this is caused by the finite choice of B. (iii) For the sequencing method, we note that $\tilde{\sigma}_3^2$ and $\hat{\sigma}_3^2$ are comparable in most settings. This indicates that the pairwise adjustment of the sequencing method successfully generalizes the methodology and works effectively for the unbalanced repeated measurement settings.

n	f	σ^2	$\tilde{\sigma}_1^2$	$\tilde{\sigma}_2^2(b_t)$	$\tilde{\sigma}_2^2(b_s)$	$\tilde{\sigma}_3^2(b_t)$	$\tilde{\sigma}_3^2(b_s)$	$\hat{\sigma}_{\text{SS}}^2$
30	f_1	0.25	1.46	1.79	1.81	1.25	**1.21**	1.49
		4	1.46	1.75	1.62	1.16	**1.10**	1.31
	f_2	0.25	**1.46**	3.78	24.22	4.98	16.75	2.06
		4	1.46	1.77	1.75	1.22	**1.20**	1.68
200	f_1	0.25	1.49	1.60	1.48	1.08	**1.05**	1.08
		4	1.49	1.58	1.46	1.09	**1.04**	1.07
	f_2	0.25	1.48	1.59	1.82	**1.09**	1.44	1.26
		4	1.48	1.58	1.46	1.09	**1.04**	1.15

Table 3: Relative mean squared errors for the six estimators under various settings with unbalanced repeated measurements.

6 Real Data Examples

The data set was reported by University of Oxford via the department of statistics consulting service (Venables & Ripley 2002). The data were collected on the concentration of a chemical GAG in the urine, and the aim of the study was to produce a chart to help a pediatrician to assess if a child's GAG concentration is normal or not. The data set is in the data frame "GAGurine" and it can be downloaded in the R package "MASS". The following two variables are included: *age* as the child age in years and *GAG* as the concentration of GAG. To estimate the residual variance, we use all the

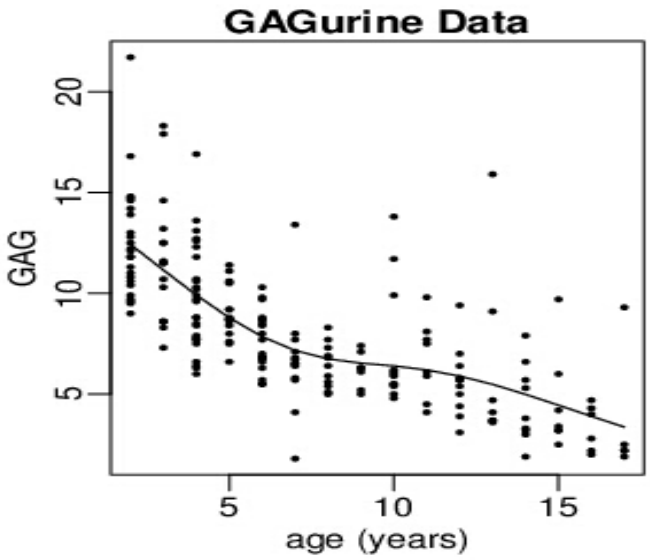

Figure 2: The GAGurine data together with the fitted curve by smoothing spline.

167 children aged from 2 to 17 years. From the scatter plot and its fitted curve, we observe a slightly nonlinear pattern between the two variables. In addition, a constant variance assumption seems not unreasonable and therefore we adopt this assumption in this study. For the proposed methods and the residual-based method, the estimated variances are: $\tilde{\sigma}_1^2 = 5.89$, $\tilde{\sigma}_2^2(b_t) = 5.10$, $\tilde{\sigma}_2^2(b_s) = 5.41$, $\tilde{\sigma}_3^2(b_t) = 5.87$, $\tilde{\sigma}_3^2(b_s) = 5.97$, and $\hat{\sigma}_{\rm SS}^2 = 5.57$, where $\tilde{\sigma}_2^2(b_t)$ and $\tilde{\sigma}_2^2(b_s)$ are computed using $B = 50$. Overall, we note that there is not large discrepancy among these estimates. Since for small n and large m, our simulation indicates that the sample variance estimator usually performs the best, we can evaluate the performance of other estimators by inspecting the difference from $\tilde{\sigma}_1^2$. To this end, the two sequencing estimators are again the winner against other methods including the residual-based estimator.

7 Conclusion

We have proposed three difference-based methods for estimating the residual variance in nonparametric regression with repeated measurement data: the sample variance method by using only the variation within design points, the partitioning method by using only the variation between design points, and the sequencing method by using both between and within variations. We have investigated the statistical properties of the proposed estimators for fixed m and large n and have established the optimality of the sequencing estimator. When n is fixed while m is large, it is seen that the sample

variance estimator is an efficient estimator and is recommended in practice. We further conducted extensive simulation studies to assess the finite sample performance. In terms of implementation specifics, for large n, we recommend the sequencing estimator with the bandwidth $b_t = n^{1/3}$ when f is rough; otherwise, we recommend the sequencing estimator with bandwidth $b_s = n^{1/2}$. For small n, we recommend the sample variance estimator when f is rough or when m is large; otherwise we recommend the sequencing estimator with the bandwidth $b_t = n^{1/3}$. We have also extended the proposed difference-based methods to handle unbalanced repeated measurement settings and found them work well in practice. Further work might be needed to investigate the statistical properties under the unbalanced design.

Finally, we note that the difference-based methods have been extended to more general settings, e.g., to multivariate covariates models (Hall, Kay & Titterington 1991, Kulasekera & Gallagher 2002, Munk, Bissantz, Wagner & Freitag 2005, Bock, Bowman & Ismail 2007, Liitiäinen, Corona & Lendasse 2010) and to semiparametric regression models (Xu & You 2007, Wang, Brown & Cai 2011). Note also that a constant variance assumption may not be realistic in practice and the difference-based methods have been applied to the variance function estimation in the literature (Müller & Stadtmüller 1993, Levine 2006, Brown & Levine 2007, Cai, Levine & Wang 2009). Further research is warranted in these directions when repeated measurement data are presented.

Acknowledgment

Yanyuan Ma's research was supported by NSF grant DMS1000354, DMS1206693 and NINDS grant R01-NS073671. Tiejun Tong's research was supported by Hong Kong RGC grant HKBU202711, and Hong Kong Baptist University grants FRG2/11-12/110, FRG1/13-14/018, and FRG2/13-14/062. Lixing Zhu's research was supported by Hong Kong RGC grant HKBU202810. The authors thank the editor, the associate editor, and two reviewers for their constructive comments that led to a substantial improvement of the paper.

References

Bock, M., Bowman, A. W. and Ismail, B. (2007). Estimation and inference for error variance in bivariate nonparametric regression, *Statistics and Computing* **17**: 39–47.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Springer.

Brown, L. D. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method, *Annals of Statistics* **35**: 2219–2232.

Cai, T., Levine, M. and Wang, L. (2009). Variance function estimation in multivariate nonparametric regression, *Journal of Multivariate Analysis* **100**: 126–136.

Carroll, R. J. (1987). The effects of variance function estimation on prediction and calibration: an example, *Statistical decision theory and related topics, IV* **2**: 273–280.

Carroll, R. J. and Ruppert, D. (1988). *Transforming and Weighting in Regression*, London: Chapman and Hall.

Dette, H., Munk, A. and Wagner, T. (1998). Estimating the variance in nonparametric regression - what is a reasonable choice?, *Journal of the Royal Statistical Society, Series B* **60**: 751–764.

Du, J. and Schick, A. (2009). A covariate-matched estimator of the error variance in nonparametric regression, *Journal of Nonparametric Statistics* **21**: 263–285.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques, *Journal of the American Statistical Association* **85**: 387–392.

Gasser, T., Kneip, A. and Kohler, W. (1991). A flexible and fast method for automatic smoothing, *Journal of the American Statistical Association* **86**: 643–52.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression, *Biometrika* **73**: 625–633.

Hall, P., Kay, J. and Titterington, D. (1991). On estimation of noise variance in two-dimensional signal processing, *Advances in Applied Probability* **23**: 476–495.

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

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression, *Biometrika* **77**: 415–419.

Kariya, T. and Kurata, H. (2004). *Generalized Least Squares*, Wiley.

Kulasekera, K. B. and Gallagher, C. (2002). Variance estimation in nonparametric multiple regression, *Communications in Statistics, Part A–Theory and Methods* **31**: 1373–1383.

Levine, M. (2006). Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach, *Computational Statistics & Data Analysis* **50**: 3405–3431.

Liitiäinen, E., Corona, F. and Lendasse, A. (2010). Residual variance estimation using a nearest neighbor statistic, *Journal of Multivariate Analysis* **101**: 811–823.

McElroy, F. W. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased, *Journal of the American Statistical Association* **62**: 1302–1304.

Müller, H. G. and Stadtmüller, U. (1993). On variance function estimation with quadratic forms, *Journal of Statistical Planning and Inference* **35**: 213–231.

Müller, U., Schick, A. and Wefelmeyer, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, *Statistics* **37**: 179–188.

Munk, A., Bissantz, N., Wagner, T. and Freitag, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional, *Journal of the Royal Statistical Society, Series B* **67**: 19–41.

Rice, J. A. (1984). Bandwidth choice for nonparametric regression, *Annals of Statistics* **12**: 1215–1230.

Rose, C. and Smith, M. D. (2002). *Mathematical Statistics with Mathematica*, New York: Springer-Verlag.

Tong, T., Liu, A. and Wang, Y. (2008). Relative errors of difference-based variance estimators in nonparametric regression, *Communications in Statistics: Theory and Methods* **37**: 2890–2902.

Tong, T., Ma, Y. and Wang, Y. (2013). Optimal variance estimation without estimating the mean function, *Bernoulli* **19**: 1839–1854.

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

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, Springer.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics With S, Fourth Edition*, Springer, New York.

von Neumann, J. (1941). Distribution of the ratio of the mean squared successive difference to the variance, *Annals of Mathematical Statistics* **12**: 367–395.

Wang, L., Brown, L. D. and Cai, T. (2011). A difference based approach to the semiparametric partial linear model, *Electronic Journal of Statistics* **5**: 619–641.

Whittle, P. (1964). On the convergence to normality of quadratic forms in independent variables, *Theory of Probability and Its Applications* **9**: 103–108.

Xu, Q. and You, J. (2007). Difference-based estimation for error variances in repeated measurement regression models, *Statistics & Probability letter* **77**: 811–816.

Appendix 1: Proof of Theorem 1

Assume that the linear function is $f(x) = \mu + \delta x$. For ease of notation, denote $f_i = f(x_i)$, $i = 1, \ldots, n$. Then

$$\begin{aligned} E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] &= \sigma^2 + \frac{\delta^2}{2c_r}\left\{\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}\frac{(r-1)^2}{n^2} + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}\frac{r^2}{n^2}\right\} \\ &= \sigma^2 + \frac{\delta^2}{2c_r n^2}\left[(n-r+1)(r-1)^2\sum_{k=1}^{m-1}(m-k) + (n-r)r^2\sum_{k=1}^{m}k\right] \\ &= \sigma^2 + \frac{1}{2}d_r\delta^2. \end{aligned}$$

Note that $\sum_{r=1}^{b} w_r = 1$ and $\bar{d}_w = \sum_{r=1}^{b} w_r d_r$. We have

$$E(\bar{\sigma}^2_w) = \sum_{r=1}^{b} w_r E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] = \sigma^2 + \frac{1}{2}\delta^2\bar{d}_w. \tag{24}$$

Further, we have

$$E(\hat{\beta}) = \frac{\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right]}{\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)^2} = \frac{(\delta^2/2)\left(\sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2\right)}{\sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2} = \frac{1}{2}\delta^2, \tag{25}$$

where $\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)^2 = \sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2$ and

$$\begin{aligned}
\sum_{r=1}^{b} w_r(d_r - \bar{d}_w) E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] &= \sum_{r=1}^{b} w_r d_r E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] - \bar{d}_w E(\bar{\sigma}^2_w) \\
&= \sigma^2 \bar{d}_w + \frac{1}{2}\delta^2 \sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w \left(\sigma^2 + \frac{1}{2}\delta^2 \bar{d}_w\right) \\
&= \frac{1}{2}\delta^2 \left(\sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2\right).
\end{aligned}$$

Finally, by (24) and (25), we have $E(\hat{\sigma}_3^2) = E(\bar{\sigma}_w^2) - E(\hat{\beta})\bar{d}_w = \sigma^2$. This finishes the proof.

Appendix 2: Proof of Theorem 2

Proof of (18): Instead of using the formula $\mathrm{Bias}(\hat{\sigma}_3^2) = \boldsymbol{f}^T D \boldsymbol{f}/\mathrm{tr}(D)$, we calculate this quantity directly from (15) which gives a more accurate approximation. For ease of notation, denote $f_i = f(x_i)$, $f_i' = f'(x_i)$, and $f_i'' = f''(x_i)$, $i = 1, \ldots, n$. Similarly as Appendix 1, we have

$$\begin{aligned}
E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] &= \sigma^2 + \frac{m}{2c_r}\left[(m-1)\sum_{i=r}^{n}(f_i - f_{i-r+1})^2 + (m+1)\sum_{i=r+1}^{n}(f_i - f_{i-r})^2\right] \\
&= \sigma^2 + \frac{m}{2c_r}\left\{(m-1)\sum_{i=r}^{n}\left[\frac{(r-1)^2}{n^2}(f_i')^2 + O(\frac{(r-1)^3}{n^3})\right]\right. \\
&\quad \left. +(m+1)\sum_{i=r+1}^{n}\left[\frac{r^2}{n^2}(f_i')^2 + O(\frac{r^3}{n^3})\right]\right\} \\
&= \sigma^2 + J d_r + O(\frac{r^3}{n^3}).
\end{aligned}$$

Consequently, we have $E(\bar{\sigma}_w^2) = \sum_{r=1}^{b} w_r E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] = \sigma^2 + J\bar{d}_w + O(b^3/n^3)$. In addition, it is easy to verify that

$$\bar{d}_w = \frac{1}{s_b}\sum_{r=1}^{b} c_r d_r = \frac{b^2}{3n^2} - \frac{b^3}{12n^3} - \frac{(m-1)b}{2mn^2} + o(\frac{b^3}{n^3}) + o(\frac{b}{n^2}), \tag{26}$$

and

$$\sum_{r=1}^{b} w_r(d_r - \bar{d}_w)^2 = \sum_{r=1}^{b} w_r d_r^2 - \bar{d}_w^2 = \frac{4b^4}{45n^4} + o(\frac{b^4}{n^4}). \tag{27}$$

This leads to

$$\begin{aligned}\sum_{r=1}^{b} w_r(d_r-\bar{d}_w)E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right] &= \sum_{r=1}^{b} w_r d_r\left[\sigma^2+Jd_r+O(\frac{r^3}{n^3})\right]-\bar{d}_w\left[\sigma^2+J\bar{d}_w+O(\frac{b^3}{n^3})\right]\\ &= J\left(\sum_{r=1}^{b} w_r d_r^2-\bar{d}_w^2\right)+O(\frac{b^5}{n^5}).\end{aligned}$$

Finally, we have

$$\begin{aligned}E(\hat{\sigma}^2_3) &= E(\bar{\sigma}^2_w)-\frac{\bar{d}_w}{\sum_{r=1}^{b} w_r(d_r-\bar{d}_w)^2}\sum_{r=1}^{b} w_r(d_r-\bar{d}_w)E\left[\hat{\sigma}^2_{\mathrm{Rt}}(r)\right]\\ &= \left[\sigma^2+J\bar{d}_w+O(\frac{b^3}{n^3})\right]-\left[J\bar{d}_w+O(\frac{b^3}{n^3})\right]\\ &= \sigma^2+O(\frac{b^3}{n^3}).\end{aligned}$$

To achieve the variance of $\hat{\sigma}^2_3$, we need Lemmas 2 and 3.

Lemma 2. *For the equally spaced design with $b\to\infty$ and $b/n\to 0$, we have*

(i) $\sum_{r=1}^{b}\tau_r = b-\frac{5b^2}{16n}+o(\frac{b^2}{n})+o(1).$

(ii) $\sum_{r=1}^{b}\tau_r^2 = \frac{9}{4}b+o(b).$

(iii) $\sum_{r=1}^{b} r\tau_r = \frac{3}{16}b^2+o(b^2).$

(iv) $\sum_{r=1}^{b} r^2\tau_r = o(b^3).$

(v) $\sum_{r=1}^{i}\tau_r = \frac{9}{4}i-\frac{5i^3}{4b^2}+o(i)+o(\frac{i^3}{b^2}),\ 1\le i\le b.$

(vi) $\sum_{r=1}^{i} r\tau_r = \frac{9}{8}i^2-\frac{15i^4}{16b^2}+o(i^2)+o(\frac{i^4}{b^2}),\ 1\le i\le b.$

Proof. (i) Let $\eta=\bar{d}_w/\sum_{r=1}^{b} w_r(d_r-\bar{d}_w)^2$. Then $\tau_r=1-\eta(d_r-\bar{d}_w)$. By (26) and (27), we have $\eta=15n^2/(4b^2)+o(n^2/b^2)$ and $\sum_{r=1}^{b}(d_r-\bar{d}_w)=\sum_{r=1}^{b} d_r-b\bar{d}_w=$

$b^4/(12n^3) + o(b^4/n^3) + o(b^2/n^2)$. This leads to

$$\sum_{r=1}^{b} \tau_r = b - \eta \sum_{r=1}^{b} (d_r - \bar{d}_w) = b - \frac{5b^2}{16n} + o(\frac{b^2}{n}) + o(1).$$

(ii) By $\sum_{r=1}^{b} d_r = b^3/(3n^2) + o(b^3/n^2)$ and $\sum_{r=1}^{b} d_r^2 = b^5/(5n^4) + o(b^5/n^4)$, we have $\sum_{r=1}^{b}(d_r - \bar{d}_w)^2 = \sum_{r=1}^{b} d_r^2 - 2\bar{d}_w \sum_{r=1}^{b} d_r + b\bar{d}_w^2 = 4b^5/(45n^4) + o(b^5/n^4)$. This leads to

$$\sum_{r=1}^{b} \tau_r^2 = b - 2\eta \sum_{r=1}^{b} (d_r - \bar{d}_w) + \eta^2 \sum_{r=1}^{b} (d_r - \bar{d}_w)^2 = \frac{9}{4} b + o(b).$$

(iii) Note that $\sum_{r=1}^{b} r d_r = b^4/(4n^2) + o(b^4/n^2)$. We have

$$\sum_{r=1}^{b} r\tau_r = (1 + \eta\bar{d}_w) \sum_{r=1}^{b} r - \eta \sum_{r=1}^{b} r d_r = \frac{3}{16} b^2 + o(b^2).$$

(iv) Note that $\sum_{r=1}^{b} r^2 d_r = b^5/(5n^2) + o(b^5/n^2)$. We have

$$\sum_{r=1}^{b} r^2\tau_r = (1 + \eta\bar{d}_w) \sum_{r=1}^{b} r^2 - \eta \sum_{r=1}^{b} r^2 d_r = o(b^3).$$

(v) Note that $\sum_{r=1}^{i} d_r = i^3/(3n^2) + o(i^3/n^2)$. For any $1 \le i \le b$, we have

$$\sum_{r=1}^{i} \tau_r = (1 + \eta\bar{d}_w) i - \eta \sum_{r=1}^{i} d_r = \frac{9}{4} i - \frac{5i^3}{4b^2} + o(i) + o(\frac{i^3}{b^2}).$$

(vi) Note that $\sum_{r=1}^{i} r d_r = i^4/(4n^2) + o(i^4/n^2)$. For any $1 \le i \le b$, we have

$$\sum_{r=1}^{i} r\tau_r = (1 + \eta\bar{d}_w) \sum_{r=1}^{i} r - \eta \sum_{r=1}^{i} r d_r = \frac{9}{8} i^2 - \frac{15i^4}{16b^2} + o(i^2) + o(\frac{i^4}{b^2}).$$

Lemma 3. *Under the same conditions as in Theorem 2, we have*

(i) $\boldsymbol{f}^T D^2 \boldsymbol{f} = O(\frac{b^5}{n^2}) + O(\frac{b^2}{n})$.

(ii) $\boldsymbol{f}^T [D \cdot \mathrm{diag}(D)\boldsymbol{u}] = O(\frac{b^4}{n}) + O(b^2)$.

(iii) $\mathrm{tr}[\mathrm{diag}(D)^2] = 4m^3 n b^2 - \frac{103m^3}{28} b^3 + o(b^3)$.

(iv) $\mathrm{tr}(D^2) = 4m^3 n b^2 - \frac{103m^3}{28} b^3 + \frac{9m^2}{2} nb + o(b^3) + o(nb)$.

Proof. (i) Noting that the matrix D is symmetric, we have

$$\boldsymbol{f}^T D^2 \boldsymbol{f} = \boldsymbol{f}^T D^T D \boldsymbol{f} = (D\boldsymbol{f})^T D\boldsymbol{f} = \boldsymbol{\xi}^T \boldsymbol{\xi},$$

where $\boldsymbol{\xi} = D\boldsymbol{f} = (\xi_1, \ldots, \xi_{nm})^T$. Let $l = (i-1)m + j$, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Note that f has a bounded second derivative. When $i \in [b+1, n-b]$, by Lemma 2 (i), (iii) and (iv), we have

$$\begin{aligned}
\xi_l &= (j-1)\sum_{r=1}^{b} \tau_r \left[\frac{r-1}{n} f_i' - \frac{(r-1)^2}{2n^2} f_i'' + o(\frac{r^2}{n^2})\right] \\
&\quad +(m-j+1)\sum_{r=1}^{b} \tau_r \left[\frac{r}{n} f_i' - \frac{r^2}{2n^2} f_i'' + o(\frac{r^2}{n^2})\right] \\
&\quad -(m-j)\sum_{r=1}^{b} \tau_r \left[\frac{r-1}{n} f_i' + \frac{(r-1)^2}{2n^2} f_i'' + o(\frac{r^2}{n^2})\right] \\
&\quad -j\sum_{r=1}^{b} \tau_r \left[\frac{r}{n} f_i' + \frac{r^2}{2n^2} f_i'' + o(\frac{r^2}{n^2})\right] \\
&= \frac{m-2j+1}{n} f_i' \sum_{r=1}^{b} \tau_r - \frac{1}{2n^2} f_i'' \sum_{r=1}^{b} \tau_r [2mr^2 - 2(m-1)r + (m-1)] + o(\frac{b^3}{n^2}) \\
&= O(\frac{b}{n}) + o(\frac{b^3}{n^2}).
\end{aligned}$$

When $i \in [1, b]$, by Lemma 2 (i), (iii), (v) and (vi), we have

$$\begin{aligned}
\xi_l &= (j-1)\sum_{r=1}^{i} \tau_r \left[\frac{r-1}{n} f_i' + o(\frac{r^2}{n^2})\right] + (m-j+1)\sum_{r=0}^{i-1} \tau_r \left[\frac{r}{n} f_i' + o(\frac{r^2}{n^2})\right] \\
&\quad -(m-j)\sum_{r=1}^{b} \tau_r \left[\frac{r-1}{n} f_i' + o(\frac{r^2}{n^2})\right] - j\sum_{r=1}^{b} \tau_r \left[\frac{r}{n} f_i' + o(\frac{r^2}{n^2})\right] \\
&= O(\frac{b^2}{n}).
\end{aligned}$$

When $i \in [n-b+1, n]$, similar argument leads to $\xi_l = O(b^2/n)$. Finally,

$$\boldsymbol{f}^T D^2 \boldsymbol{f} = \boldsymbol{\xi}^T \boldsymbol{\xi} = \sum_{l=1}^{mb} \xi_l^2 + \sum_{l=mb+1}^{m(n-b)} \xi_l^2 + \sum_{l=m(n-b)+1}^{nm} \xi_l^2 = O(\frac{b^5}{n^2}) + O(\frac{b^2}{n}).$$

(ii) Note that $\boldsymbol{f}^T[D\cdot\mathrm{diag}(D)\boldsymbol{u}]=\boldsymbol{\xi}^T\cdot\mathrm{diag}(D)\boldsymbol{u}$. By part (i) and Lemma 2 (i) and (v), we have

$$\begin{aligned}\boldsymbol{f}^T[D\cdot\mathrm{diag}(D)\boldsymbol{u}] &= \sum_{i=1}^{b}\sum_{j=1}^{m}\xi_{(i-1)m+j}\left[m\sum_{r=1}^{b}\tau_r+m\sum_{r=0}^{i-1}\tau_r+(j-1)\tau_i\right]\\ &\quad+\sum_{l=mb+1}^{m(n-b)}\xi_l\left(2m\sum_{r=1}^{b}\tau_r\right)\\ &\quad+\sum_{i=n-b+1}^{n}\sum_{j=1}^{m}\xi_{(i-1)m+j}\left[m\sum_{r=1}^{b}\tau_r+m\sum_{r=0}^{n-i}\tau_r+(m-j)\tau_i\right]\\ &= O(\frac{b^4}{n})+O(b^2).\end{aligned}$$

(iii) By Lemma 2 (i) and (v), we have

$$\begin{aligned}\mathrm{tr}[\mathrm{diag}(D)^2] &= \sum_{l=mb+1}^{m(n-b)}\left(2m\sum_{r=1}^{b}\tau_r\right)^2+2\sum_{i=1}^{b}\sum_{j=1}^{m}\left[m\sum_{r=1}^{b}\tau_r+m\sum_{r=0}^{i-1}\tau_r+(j-1)\tau_i\right]^2\\ &= 4m^3(n-2b)\left(b-\frac{5b^2}{16n}+o(\frac{b^2}{n})+o(1)\right)^2+2m^3\sum_{i=1}^{b}\left[b+\frac{9}{4}i-\frac{5i^3}{4b^2}+o(b)\right]^2\\ &= 4m^3nb^2-\frac{103m^3}{28}b^3+o(b^3).\end{aligned}$$

(iv) By part (iii) and Lemma 2 (ii), we have

$$\begin{aligned}\mathrm{tr}(D^2) &= \sum_{l=mb+1}^{m(n-b)}\left[\left(2m\sum_{r=1}^{b}\tau_r\right)^2+2m\sum_{r=1}^{b}\tau_r^2\right]\\ &\quad+2\sum_{i=1}^{b}\sum_{j=1}^{m}\left[\left(m\sum_{r=1}^{b}\tau_r+m\sum_{r=0}^{i-1}\tau_r+(j-1)\tau_i\right)^2+m\sum_{r=1}^{b}\tau_r^2+m\sum_{r=0}^{i-1}\tau_r^2+(j-1)\tau_i^2\right]\\ &= \mathrm{tr}[\mathrm{diag}(D)^2]+2m^2(n-2b)\sum_{r=1}^{b}\tau_r^2+2\sum_{i=1}^{b}\sum_{j=1}^{m}\left(m\sum_{r=1}^{b}\tau_r^2+m\sum_{r=0}^{i-1}\tau_r^2+(j-1)\tau_i^2\right)\\ &= \left[4m^3nb^2-\frac{103m^3}{28}b^3+o(b^3)\right]+2m^2(n-2b)\left[\frac{9}{4}b+o(b)\right]+O(b^2)\\ &= 4m^3nb^2-\frac{103m^3}{28}b^3+\frac{9m^2}{2}nb+o(b^3)+o(nb).\end{aligned}$$

Proof of (19): Note that the last four terms in (17) make up the variance of $\hat{\sigma}_3^2$. Note that $\sigma^4(\gamma_4-3)=\mathrm{Var}(\varepsilon^2)-2\sigma^4$, $\mathrm{tr}(D)=2s_b$, and $s_b=m^2nb-m^2b^2/2+o(b^2)$. By

Lemmas 2 and 3, we have

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}_3^2) &= \frac{1}{[\mathrm{tr}(D)]^2}\left\{4\sigma^2\boldsymbol{f}^T D^2\boldsymbol{f} + 4\boldsymbol{f}^T[D\cdot\mathrm{diag}(D)\boldsymbol{u}]\sigma^3\gamma_3\right. \\
&\quad \left.+\sigma^4(\gamma_4-3)\mathrm{tr}[\mathrm{diag}(D)^2] + 2\sigma^4\mathrm{tr}(D^2)\right\} \\
&= \frac{1}{4s_b^2}\left\{O(\frac{b^5}{n^2}) + O(\frac{b^2}{n}) + O(\frac{b^4}{n}) + O(b^2) + \sigma^4(\gamma_4-3)\left[4m^3nb^2 - \frac{103m^3}{28}b^3 + o(b^3)\right]\right. \\
&\quad \left.+2\sigma^4\left[4m^3nb^2 - \frac{103m^3}{28}b^3 + \frac{9m^2}{2}nb + o(b^3) + o(nb)\right]\right\} \\
&= \frac{1}{4s_b^2}\left[\left(4m^3nb^2 - \frac{103m^3}{28}b^3\right)\mathrm{Var}(\varepsilon^2) + 9m^2nb\sigma^4 + o(b^3) + o(nb)\right] \\
&= \frac{1}{mn}\mathrm{Var}(\varepsilon^2) + \frac{9}{4m^2nb}\sigma^4 + \frac{9b}{112mn^2}\mathrm{Var}(\varepsilon^2) + o(\frac{1}{nb}) + o(\frac{b}{n^2}).
\end{aligned}$$

Proof of (20): The MSE in (20) is an immediate result from (18) and (19).

Appendix 3: Proof of Theorem 3

Let $f_i = f(x_i)$, $i = 1, \ldots, n$. To prove the asymptotic normality for $\hat{\sigma}^2_{\mathrm{Rt}}(r)$, we first partition it into three parts, $\hat{\sigma}^2_{\mathrm{Rt}}(r) = L_1 + L_2 + L_3$, where

$$L_1 = \frac{1}{2c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})^2 + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})^2\right],$$

$$L_2 = \frac{1}{c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k}) + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})\right],$$

$$L_3 = \frac{1}{2c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k})^2 + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})^2\right].$$

(i) Note that $J = O(1)$ and $d_r = O(r^2/n^2)$. For L_1, by Taylor series we have $L_1 = Jd_r + o(r^2/n^2) = O(r^2/n^2)$. This shows that $L_1 = o(n^{-1/2})$ when $r = n^{\vartheta}$ with $0 \le \vartheta < 3/4$.

(ii) For L_2, by Cauchy-Schwarz inequality we have

$$\begin{aligned} L_2^2 &\leq \frac{2}{c_r^2}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k})\right]^2 \\ &\quad + \frac{2}{c_r^2}\left[\sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})\right]^2 \\ &\leq \frac{2}{c_r^2}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})^2\right]\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k})^2\right] \\ &\quad + \frac{2}{c_r^2}\left[\sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})^2\right]\left[\sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})^2\right]. \end{aligned}$$

This leads to

$$E(L_2^2) \leq \frac{4\sigma^2}{c_r}\left[\sum_{k=1}^{m-1}\sum_{i=r}^{n}\sum_{j=k+1}^{m}(f_i - f_{i-r+1})^2 + \sum_{k=1}^{m}\sum_{i=r+1}^{n}\sum_{j=1}^{k}(f_i - f_{i-r})^2\right] = O(\frac{r^2}{n^2}).$$

This shows that $L_2 = o_p(n^{-1/2})$ for any $r = n^{\vartheta}$ with $0 \leq \vartheta < 1/2$.

(iii) We represent the term L_3 as $L_3 = \sigma^2 + \sum_{i=r+1}^{n}\zeta_i(r)/(n-r) + O(1/n)$, where

$$\zeta_i(r) = \frac{1}{2m^2}\left[\sum_{k=1}^{m-1}\sum_{j=k+1}^{m}(\varepsilon_{ij} - \varepsilon_{i-r+1,j-k})^2 + \sum_{k=1}^{m}\sum_{j=1}^{k}(\varepsilon_{ij} - \varepsilon_{i-r,m-k+j})^2\right] - \sigma^2. \quad (28)$$

We have $E(\zeta_i(r)) = 0$. Treat $\{\zeta_i(r), i = r+1, \ldots, n\}$ as a stochastic process. With some straightforward algebra, we have (a) for $r = 1$,

$$\mathrm{Cov}(\zeta_i(r), \zeta_l(r)) = \begin{cases} [(8m^2 - 3m + 1)\gamma_4 - (8m^2 - 15m + 1)]\sigma^4/(12m^3), & l - i = 0, \\ (4m^2 + 3m - 1)(\gamma_4 - 1)\sigma^4/(24m^3), & l - i = 1, \\ 0, & l - i \geq 2; \end{cases}$$

(b) for $r = 2$,

$$\mathrm{Cov}(\zeta_i(r), \zeta_l(r)) = \begin{cases} [(5m^2 + 1)\gamma_4 - (5m^2 - 12m + 1)]\sigma^4/(12m^3), & l - i = 0, \\ (4m^2 - 3m - 1)(\gamma_4 - 1)\sigma^4/(24m^3), & l - i = 1, \\ (m + 1)(\gamma_4 - 1)\sigma^4/(8m^2), & l - i = 2, \\ 0, & l - i \geq 3; \end{cases}$$

and (c) for any $r \geq 3$,

$$\mathrm{Cov}(\zeta_i(r), \zeta_l(r)) = \begin{cases} [(5m^2 + 1)\gamma_4 - (5m^2 - 12m + 1)]\sigma^4/(12m^3), & l - i = 0, \\ (m^2 - 1)(\gamma_4 - 1)\sigma^4/(24m^3), & l - i = 1, \\ (m - 1)(\gamma_4 - 1)\sigma^4/(8m^2), & l - i = r - 1, \\ (m + 1)(\gamma_4 - 1)\sigma^4/(8m^2), & l - i = r, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the above covariances depend on i and l only through the difference $l-i$, regardless of the choice of r. This shows that for any given $r \geq 1$, $\{\zeta_i(r), i = r+1, \ldots, n\}$ is a strictly stationary sequence of random variables with mean zero and autocovariance function $C(\tau) = C(s, s+\tau) = \mathrm{Cov}(\zeta_s(r), \zeta_{s+\tau}(r))$. Also note that $\{\zeta_i(r), i = r+1, \ldots, n\}$ is an m-dependent sequence with $m = r$. Thus by Brockwell and Davis (1991), we have the following asymptotic normality for L_3,

$$\sqrt{n}(L_3 - \sigma^2) \xrightarrow{\mathcal{D}} N(0, \nu_r^2) \qquad \text{as } n \to \infty, \tag{29}$$

where $\nu_r^2 = C(0) + 2\sum_{\tau=1}^{r} C(\tau)$. For the covariance functions in (a)-(c), it is easy to verify that $\nu_1^2 = \nu_2^2 = \nu_r^2 = [\gamma_4/m - (m-1)/m^2]\sigma^4$ for any $r \geq 3$.

Finally, noting that $\hat{\sigma}^2_{\mathrm{Rt}}(r) = L_1 + L_2 + L_3 = L_3 + o_p(n^{-1/2})$, by (29) and Slutsky's theorem we have for any $r = n^{\vartheta}$ with $0 \leq \vartheta < 1/2$, $\sqrt{n}(\hat{\sigma}^2_{\mathrm{Rt}}(r) - \sigma^2) \xrightarrow{\mathcal{D}} N(0, [\gamma_4/m - (m-1)/m^2]\sigma^4)$ as $n \to \infty$.

Appendix 4: Proof of Theorem 4

To prove Theorem 4, we need the following lemma which was originated from Whittle (1964).

Lemma 4. *Assume that the matrix $A = (a_{ij})_{n\times n}$ satisfies $a_{ij} = a_{i-j}$ and $\sum_{-\infty}^{\infty} a_k^2 < \infty$. Also assume that $E(\varepsilon^2) = \sigma^2$ and $E(\varepsilon^{4+2\delta})$ is finite for some δ in $(0,1)$. Then*

$$\frac{1}{n}\boldsymbol{\varepsilon}^T A \boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i-j}\varepsilon_i\varepsilon_j \xrightarrow{\mathcal{D}} N(a_0\sigma^2, \sigma_A^2), \quad \text{as } n \to \infty,$$

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

Proof of Theorem 4: By $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$ and $\mathrm{tr}(D) = 2s_b$, we have

$$\hat{\sigma}_3^2 = \frac{1}{2s_b}\boldsymbol{f}^T D \boldsymbol{f} + \frac{1}{s_b}\boldsymbol{f}^T D\boldsymbol{\varepsilon} + \frac{1}{2s_b}\boldsymbol{\varepsilon}^T D\boldsymbol{\varepsilon}. \tag{30}$$

(i) For the first term in (30), noting that it corresponds to the bias term $E(\hat{\sigma}_3^2)$, By Theorem 2 we have $\boldsymbol{f}^T D\boldsymbol{f}/(2s_b) = O(b^3/n^3) = o(n^{-1/2})$ for any $b = n^{\vartheta}$ with $0 < \vartheta < 5/6$.

(ii) For the second term in (30), by Lemma 3 and the fact $s_b = O(nb)$ we have

$$E\left(\boldsymbol{f}^T D\boldsymbol{\varepsilon}/s_b\right)^2 = (\boldsymbol{f}^T D^2 \boldsymbol{f})\sigma^2/s_b^2 = O(b^3/n^4) + O(1/n^3).$$

This implies that $\boldsymbol{f}^T D\boldsymbol{\varepsilon}/s_b = o_p(n^{-1/2})$ for any $b = o(n)$.

(iii) Now we derive the asymptotic normality for the last term in (30). Let $(mn/2s_b)D = C - H$, where $C = (c_{ij})_{n\times n}$ is an $(mn) \times (mn)$ matrix with elements

$$c_{ij} = \begin{cases} m^2 n \sum_{r=1}^{b} \tau_r / s_b, & 1 \le i = j \le mn, \\ -mn\tau_a/(2s_b), & (a-1)m < |i-j| \le am \text{ with } a = 1, \ldots, b, \\ 0, & \text{otherwise,} \end{cases}$$

and $H = \text{diag}(h_1, h_2, \ldots, h_{mn})$ is an $(mn) \times (mn)$ diagonal matrix with elements $h_i = \{m^2 n \sum_{r=1}^{b} \tau_r - m^2 n \sum_{r=0}^{a-1} \tau_r - mn[i - 1 - (a-1)m]\tau_a\}/(2s_b)$ for $(a-1)m < i \le am$ with $a = 1, \ldots, b$; $h_i = 0$ for $(a-1)m < i \le am$ with $a = b+1, \ldots, n-b$; and $h_i = \{m^2 n \sum_{r=1}^{b} \tau_r - m^2 n \sum_{r=0}^{n-a} \tau_r - mn(am - i)\tau_{n+1-a}\}/(2s_b)$ for $(a-1)m < i \le am$ with $a = n - b + 1, \ldots, n$. Then

$$\frac{1}{2s_b}\boldsymbol{\varepsilon}^T D \boldsymbol{\varepsilon} = \frac{1}{mn}\boldsymbol{\varepsilon}^T C \boldsymbol{\varepsilon} - \frac{1}{mn}\boldsymbol{\varepsilon}^T H \boldsymbol{\varepsilon}. \tag{31}$$

For the symmetric matrix C, let $c_{ij} = c_{i-j}$ with $c_0 = m^2 n \sum_{r=1}^{b} \tau_r / s_b$; $c_{i-j} = c_{j-i} = -mn\tau_a/(2s_b)$ for $(a-1)m < |i-j| \le am$ with $a = 1, \ldots, b$; and $c_{i-j} = c_{j-i} = 0$ for $|i-j| > bm$. By Lemma 2, for any $b = n^\vartheta$ with $0 < \vartheta < 1$,

$$\sum_{-\infty}^{\infty} c_k^2 = c_0^2 + 2\sum_{k=1}^{bm} c_k^2 = \frac{m^4 n^2}{s_b^2}\left(\sum_{r=1}^{b} \tau_a\right)^2 + \frac{m^3 n^2}{2s_b^2}\sum_{a=1}^{b} \tau_a^2 = O(1) < \infty.$$

Now given that $\mathrm{E}(\varepsilon^{4+2\delta})$ is finite for some δ in $(0, 1)$, by Lemma 4 we have

$$\sqrt{mn}\left(\frac{1}{mn}\boldsymbol{\varepsilon}^T C \boldsymbol{\varepsilon} - c_0\sigma^2\right) \xrightarrow{\mathcal{D}} N(0, \sigma_c^2), \quad \text{as } n \to \infty,$$

where $\sigma_c^2 = (\gamma_4 - 3)\sigma^4 c_0^2 + 2\sigma^4 \sum_{i=1}^{mn} \sum_{j=1}^{mn} c_{i-j}^2/(mn)$. For the second term in (31), by Lemma 2 it is easy to verify that for any $b = n^\vartheta$ with $0 < \vartheta < 1/2$, $\mathrm{E}(\boldsymbol{\varepsilon}^T H \boldsymbol{\varepsilon}/mn)^2 = O(b^2/n^2)$ and further $\boldsymbol{\varepsilon}^T H \boldsymbol{\varepsilon}/(mn) = o_p(n^{-1/2})$.

By (i)-(iii) and Slutsky's theorem, we have for any $b = n^\vartheta$ with $0 < \vartheta < 1/2$,

$$\frac{\sqrt{mn}(\hat{\sigma}_3^2 - c_0\sigma^2)}{\sigma_c} \xrightarrow{\mathcal{D}} N(0, 1), \quad \text{as } n \to \infty. \tag{32}$$

Note that $c_0 = m^2 n \sum_{r=1}^{b} \tau_r / s_b = 1 + O(b/n)$. This leads to $\sqrt{mn}(c_0 - 1) = o(1)$ for any $b = n^\vartheta$ with $0 \le \vartheta < 1/2$. In addition, it is easy to verify that

$$\sigma_c^2 = \frac{m^4 n^2 (\gamma_4 - 1)\sigma^4}{s_b^2}\left(\sum_{r=1}^{b} \tau_r\right)^2 + \frac{m^3 n^2 \sigma^4}{s_b^2}\sum_{r=1}^{b} \tau_r^2 = (\gamma_4 - 1)\sigma^4 + o(1).$$

This leads to $(\gamma_4-1)\sigma^4/\sigma_c^2 \to 1$ as $n\to\infty$. Finally, by (32) and Slutsky's theorem,

$$\begin{aligned}\frac{\sqrt{mn}(\hat{\sigma}_3^2-\sigma^2)}{\sqrt{(\gamma_4-1)\sigma^4}} &= \frac{\sigma_c}{\sqrt{(\gamma_4-1)\sigma^4}}\left\{\frac{\sqrt{mn}(\hat{\sigma}_3^2-c_0\sigma^2)}{\sigma_c}+\frac{\sqrt{mn}(c_0-1)\sigma^2}{\sigma_c}\right\}\\ &\xrightarrow{\mathcal{D}} N(0,1), \quad \text{as } n\to\infty,\end{aligned}$$

for any $b=n^{\vartheta}$ with $0<\vartheta<1/2$.

Appendix 5: Proof of Lemma 1

For any $1\le r<p\le b$, by Appendix 3 it is easy to verify that

$$\mathrm{Cov}(\hat{\sigma}_{\mathrm{Rt}}^2(r),\hat{\sigma}_{\mathrm{Rt}}^2(p))=\frac{1}{(n-r)(n-p)}\sum_{i=r+1}^{n}\sum_{l=p+1}^{n}\mathrm{Cov}(\zeta_i(r),\zeta_l(p))+o(\frac{1}{n}), \tag{33}$$

where $\zeta_i(r)$ is defined in (28).

Let $Q=\sum_{i=r+1}^{n}\sum_{l=p+1}^{n}\mathrm{Cov}(\zeta_i(r),\zeta_l(p))$. When $r\ge 2$ and $p\ge r+2$, the term Q can be calculated as follows,

$$\begin{aligned}Q &= \sum_{i=r+p+1}^{n}\mathrm{Cov}(\zeta_i(r),\zeta_{i-r}(p))+\sum_{i=r+p}^{n}\mathrm{Cov}(\zeta_i(r),\zeta_{i-r+1}(p))+\sum_{i=p+1}^{n}\mathrm{Cov}(\zeta_i(r),\zeta_i(p))\\
&+\sum_{i=r+2}^{n-p+r+1}\mathrm{Cov}(\zeta_i(r),\zeta_{i+p-r-1}(p))+\sum_{i=r+1}^{n-p+r}\mathrm{Cov}(\zeta_i(r),\zeta_{i+p-r}(p))\\
&+\sum_{i=r+1}^{n-p+r-1}\mathrm{Cov}(\zeta_i(r),\zeta_{i+p-r+1}(p))+\sum_{i=r+1}^{n-p+1}\mathrm{Cov}(\zeta_i(r),\zeta_{i+p-1}(p))+\sum_{i=r+1}^{n-p}\mathrm{Cov}(\zeta_i(r),\zeta_{i+p}(p))\\
&=(\gamma_4-1)\sigma^4\left[\sum_{i=r+p+1}^{n}\frac{m+1}{8m^2}+\sum_{i=r+p}^{n}\frac{m-1}{8m^2}+\sum_{i=p+1}^{n}\frac{1}{4m}+\sum_{i=r+2}^{n-p+r+1}\frac{m^2-1}{24m^3}\right.\\
&\qquad\left.+\sum_{i=r+1}^{n-p+r}\frac{2m^2+1}{12m^3}+\sum_{i=r+1}^{n-p+r-1}\frac{m^2-1}{24m^3}+\sum_{i=r+1}^{n-p+1}\frac{m-1}{8m^2}+\sum_{i=r+1}^{n-p}\frac{m+1}{8m^2}\right]\\
&=\frac{1}{2m}(2n-2p-r)(\gamma_4-1)\sigma^4+O(1).\end{aligned} \tag{34}$$

Similarly, we can verify that $Q=(2n-2p-r)/(2m)+O(1)$ holds for $r\ge 2$ and/or $p=r+1$. We omit their derivations here for saving space. Plugging (34) into (33) leads to

$$\begin{aligned}\mathrm{Cov}(\hat{\sigma}_{\mathrm{Rt}}^2(r),\hat{\sigma}_{\mathrm{Rt}}^2(p)) &= \frac{2n-2p-r}{2m(n-r)(n-p)}(\gamma_4-1)\sigma^4+o(\frac{1}{n})\\ &= \frac{1}{mn}(\gamma_4-1)\sigma^4+o(\frac{1}{n}).\end{aligned}$$

Finally, we note that $\mathrm{Var}(\hat{\sigma}_{\mathrm{Rt}}^2(p))=[\gamma_4-1+1/m]\sigma^4/(mn)+o(1/n)$ for any $1\le p\le b$ is an immediate result from Theorem 3.