Optimal-$k$ difference sequence in nonparametric regression

Wenlin Dai
Institute of Statistics and Big Data, Renmin University of China,
Beijing 100872, China
and
Xingwei Tong
School of Statistics, Beijing Normal University, Beijing 100875, China
and
Tiejun Tong
Department of Mathematics, Hong Kong Baptist University, Hong Kong
November 29, 2022

Abstract

Difference-based methods have been attracting increasing attention in nonparametric regression, in particular for estimating the residual variance. To implement the estimation, one needs to choose an appropriate difference sequence, mainly between the optimal difference sequence and the ordinary difference sequence. The difference sequence selection is a fundamental problem in nonparametric regression, and it remains a controversial issue for over three decades. In this paper, we propose to tackle this challenging issue from a very unique perspective, namely by introducing a new difference sequence called the optimal-$k$ difference sequence. The new difference sequence not only provides a better balance between the bias-variance trade-off, but also dramatically enlarges the existing family of difference sequences that includes the optimal and ordinary difference sequences as two important special cases. We further demonstrate, by both theoretical and numerical studies, that the optimal-$k$ difference sequence has been pushing the boundaries of our knowledge in difference-based methods in nonparametric regression, and it always performs the best in practical situations.
Keywords: Difference-based method; Nonparametric regression; Optimal difference sequence; Optimal-$k$ difference sequence; Ordinary difference sequence; Residual variance.
1 Introduction

Nonparametric regression models have been widely used in statistics and economics in the past several decades, mainly because of their flexibility in capturing the relationship between the dependent and independent variables. In this paper, we consider the nonparametric regression model

\[ Y_i = g(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \]  

(1)

where \( \{Y_i\} \) are the observations, \( g \) is a mean function, \( \{x_i\} \) are the design points, and \( \{\varepsilon_i\} \) are independent and identically distributed random errors with zero mean and residual variance \( \sigma^2 > 0 \).

In nonparametric regression, the estimation of the mean function \( g \) is a fundamental and important problem. An accurate estimate of the mean function is required for various purposes including, but not limited to, describing the relationship between responses and covariates, predicting observations at new experimental points, and imputing missing data values. In the existing literature, a large amount of effort has been made to obtain a reasonable estimate of \( g \) by, e.g., the kernel method (Härdle, 1990; Wand and Jones, 1995), the local linear method (Fan and Gijbels, 1996), and the smoothing spline method (Wang, 2011). Apart from the mean function, it is known that the residual variance also plays an important role and needs to be accurately estimated as well (Dette et al., 1998). For illustration, an estimate of the residual variance is needed for the purpose of constructing the confidence band for the mean function (Rice, 1984), testing the goodness of fit of the mean function (Carroll and Ruppert, 1988; Gasser et al., 1991), or assessing the discontinuities of the mean function (Müller and Stadtmüller, 1999).

To estimate the residual variance, there are two main classes of methods in the literature: residual-based methods and difference-based methods. Residual-based methods are the classical approaches for estimating the residual variance, which often make use of the summation of squared residuals as follows:

\[ \hat{\sigma}^2 = \frac{1}{n - \nu} \sum_{i=1}^{n} \{Y_i - \tilde{g}(x_i)\}^2, \]  

(2)

where \( \tilde{g} \) is a nonparametric fit of the mean function and \( \nu \) is the degrees of freedom associated with the fitted \( \tilde{g} \). By Hall and Marron (1990), residual-based estimators are capable to achieve the asymptotically optimal rate for the mean squared error (MSE) as \( \text{MSE}(\hat{\sigma}^2) = n^{-1} \text{var}(\varepsilon^2) + \)
Despite the good theoretical properties, it is noteworthy that the practical performance of residual-based estimators may not necessarily be acceptable due to a couple of reasons. First, the performance of $\hat{\sigma}^2$ in (2) is very sensitive to a delicate choice of the smoothing parameter. Second, the residual-based estimator $\hat{\sigma}^2$ is a second-step estimator that directly follows from the mean function estimator $\hat{g}$. In nonparametric regression, these two estimators are less likely to be independent of each other; and consequently, they may fail to provide a reliable confidence band for the mean function.

In view of the above limitations, difference-based methods have emerged that provided popular alternatives for estimating the residual variance. They are often constructed as a linear combination of squared differences from the neighbouring observations, do not require a nonparametric fit of the mean function, and hence are easy to implement. For simplicity, we assume that the design points are equally spaced on $[0, 1]$ with $x_i = i/n$ for $i = 1, \ldots, n$. For model (1), Rice (1984) proposed the first-order difference-based estimator as

$$\hat{\sigma}_{R}^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 / \left\{ 2(n-1) \right\}. \quad (3)$$

where the difference sequence, $(d_0, \ldots, d_r)$, is a sequence of real numbers satisfying

$$\sum_{j=0}^{r} d_j = 0 \quad \text{and} \quad \sum_{j=0}^{r} d_j^2 = 1 \quad (4)$$

with $d_0d_r \neq 0$ and $d_0 > 0$, and $r$ is the order of the difference sequence. When $r = 1$, the unique solution of the difference sequence under constraint (4) is $(d_0, d_1) = (2^{1/2}, -2^{-1/2})$, which results in the Rice estimator.

When $r \geq 2$, however, there are infinitely many solutions for the difference sequence under constraint (4). A natural question is then: which difference sequence is the best for estimator (3) to estimate the residual variance? To answer this question, there are two popular difference sequences available in the literature: the optimal difference sequence and the ordinary difference sequence. The optimal difference sequence is to minimize the asymptotic variance of estimator (3), whereas the ordinary difference sequence is to eliminate the estimation bias up to order $r - 1$. In other words, none of the two difference sequences took into account the bias-variance trade-off for the variance estimation. For the special case of $r = 2$, Hall et al. (1990) derived
the optimal difference sequence as \((d_0, d_1, d_2) = (0.809, -0.5, -0.309)\) so that the resulting estimator is 
\[
\hat{\sigma}^2_{HKT} = \frac{\sum_{i=1}^{n-2} (0.809Y_i - 0.5Y_{i+1} - 0.309Y_{i+2})^2}{(n-2)},
\]
and Gasser et al. (1986) derived the ordinary difference sequence as 
\((d_0, d_1, d_2) = (6^{-1/2}, -(2/3)^{1/2}, 6^{-1/2})\) so that the resulting estimator is
\[
\hat{\sigma}^2_{GSJ} = \frac{\sum_{i=1}^{n-2} (Y_i - 2Y_{i+1} + Y_{i+2})^2}{6(n-2)}.
\]

For further discussion on the two difference sequences, Dette et al. (1998) conducted a comparative study and concluded that the ordinary difference sequence should be recommended when the sample size is small and the signal-to-noise ratio is large; while for other situations, the optimal difference sequence can be the default choice for practical use. Although very simple to implement, the application of their rule is somewhat limited since the signal-to-noise ratio is rarely known in practice. As a consequence, the choice of the difference sequence remains, in fact, rather arbitrary in the subsequent literature. Inspired by this, Dai et al. (2017) proposed a unified framework for the variance estimation that combines the linear regression method with the higher-order difference-based estimators systematically. They further showed that, under the unified framework, the ordinary difference sequence can be consistently applied between the two difference sequences.

In this paper, we propose to further advance the difference sequence selection from another unique perspective. To achieve this, we first reformulate the existing difference sequences as solutions to an optimization problem that minimizes the variance of the estimator under certain constraints. Then under the optimization framework, we propose a new family of difference sequence, called the \textit{optimal-}k \textit{difference sequence}, by providing more flexible constraints on the estimation bias and variance. Moreover, we show that our newly proposed sequence is capable to achieve a better bias-variance trade-off in estimating the residual variance, and it also includes the optimal and ordinary difference sequences as two important special cases. Through theoretical and numerical studies, we demonstrate that the optimal- \textit{k} difference sequence has been pushing the boundaries of our knowledge in difference-based methods in nonparametric regression, and more importantly, it always performs the best in practical situations.

The rest of this paper is organized as follows. In Section 2, we review the two existing difference sequences under the optimization framework for estimating the residual variance. In Section 3, we define the optimal- \textit{k} difference sequence, investigate its properties, and provide a procedure for generating the new sequence. In Section 4, we apply the optimal- \textit{k} difference sequence and introduce a new difference-based estimator called the optimal- \textit{k} estimator. We further study its asymptotic properties and draw connections with other difference-based estimators. In Section
5, we conduct simulation studies to evaluate the finite sample performance of the new estimator and compare it with the existing methods. Finally, we conclude the paper with some discussion and future work in Section 6, and provide the technical details in the Appendix.

2 Optimal and ordinary difference sequences

To take a further look at the existing difference-based estimators, we first represent estimator (3) as a quadratic form of the observations as follows:
\[
\hat{\sigma}^2 = \frac{1}{n - r} Y^T D Y,
\]

where \( Y = (Y_1, \ldots, Y_n)^T \) with \( T \) denoting the transpose of a vector or a matrix, and \( D = \tilde{D}^T \tilde{D} \) with \( \tilde{D} \) being an \((n - r) \times n\) matrix of the structure
\[
\tilde{D} = \begin{pmatrix} d_0 & \ldots & d_r & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & \ddots & 0 & d_0 & \ldots & d_r \end{pmatrix}.
\]

According to Dette et al. (1998), the MSE of \( \hat{\sigma}^2 \) can be expressed as
\[
\text{MSE}(\hat{\sigma}^2) = \frac{1}{(n - r)^2} \left[ (g^T D g)^2 + 4\sigma^2 g^T D^2 g + 4g^T \{D \text{diag}(D) u\} \sigma^3 \gamma_3 + \sigma^4 \text{tr} \{\text{diag}(D)^2\} (\gamma_4 - 3) + 2\sigma^4 \text{tr} (D^2) \right],
\]

where \( g = (g(x_1), \ldots, g(x_n))^T \), \( \text{diag}(D) \) denotes the diagonal matrix of the matrix \( D \), \( u = (1, \ldots, 1)^T \), \( \gamma_i = E \{ (\varepsilon/\sigma)^i \} \) for \( i = 3 \) and 4, and \( \text{tr}(D) \) denotes the trace of the matrix \( D \). The first term in the right side of (6) is the squared bias, and the remaining four terms make up the variance of the estimator. When the random errors follow a normal distribution, the third and fourth terms in the right side of (6) will be zero so that the variance of the estimator can be further simplified as
\[
\text{var}(\hat{\sigma}^2) = \frac{1}{(n - r)^2} \left\{ 4\sigma^2 g^T D^2 g + 2\sigma^4 \text{tr}(D^2) \right\}.
\]
In what follows, we review in detail the optimal and ordinary difference sequences. Also from a unique perspective, we reformulate them as solutions to an optimization problem that minimizes the MSE of the estimator under different constraints.

### 2.1 Optimal difference sequence

Under some mild conditions, Hall et al. (1990) showed that the estimation bias of estimator (5) is asymptotically negligible compared to the estimation variance. They further derived the asymptotic MSE, or equivalently the asymptotic variance, of estimator (5) as

\[
\text{MSE}(\hat{\sigma}^2) = \frac{1}{n} \{ \text{var}(\varepsilon^2) + 4\sigma^4\delta(r) \} + o\left(\frac{1}{n}\right),
\]

where

\[
\delta(r) = \sum_{c=1}^{r} (\sum_{j=0}^{r-c} d_j d_{j+c})^2 \quad \text{for} \quad r \geq 1.
\]

It is clear that, besides the error moments and the sample size, the asymptotic MSE also depends on the choice of difference sequence \(d(r) = (d_0, \ldots, d_r)\) through \(\delta(r)\).

The optimal difference sequence, denoted by \(d_{\text{opt}}(r)\), was defined as the minimizer of the asymptotic MSE. This is equivalent to minimizing the quantity \(\delta(r)\) under the following optimization problem:

\[
\text{arg min}_{d(r) \in \mathbb{R}^{r+1}} \delta(r) \quad \text{subject to} \quad \sum_{j=0}^{r} d_j = 0 \quad \text{and} \quad \sum_{j=0}^{r} d_j^2 = 1. \tag{7}
\]

To solve (7), the Lagrange multiplier can be readily applied so that the optimal difference sequence satisfies

\[
\sum_{j=0}^{r-c} d_{\text{opt},j} d_{\text{opt},j+c} = -\frac{1}{2r}, \quad 1 \leq c \leq r.
\]

They further lead to \(\delta_{\text{opt}}(r) = \sum_{c=1}^{r} \{ -1/(2r) \}^2 = 1/(4r)\), which is the minimum value of \(\delta(r)\) associated with any difference sequence \(d(r)\).

We refer to estimator (5) with the optimal difference sequence as the optimal estimator, denoted by \(\hat{\sigma}^2_{\text{opt}}(r)\). When the design points are equally spaced and the mean function has a bounded first derivative, it can be further shown that

\[
\text{var}\{\hat{\sigma}^2_{\text{opt}}(r)\} = \frac{1}{n} \{ \text{var}(\varepsilon^2) + 4\sigma^4\delta_{\text{opt}}(r) \} + o\left(\frac{1}{n}\right) \quad \text{and} \quad \text{bias}\{\hat{\sigma}^2_{\text{opt}}(r)\} = O\left(\frac{1}{n^2}\right).
\]
This coincides with the result in [Hall et al. (1990)] that the estimation bias is asymptotically negligible compared to the estimation variance.

### 2.2 Ordinary difference sequence

When the sample size is small and the mean function is rough, it is known that the estimation bias of $\hat{\sigma}^2_{opt}$ is no longer negligible, or more seriously, it may even dominate the MSE. For the variance estimation in such scenarios, the ordinary difference sequence was then introduced that aims to eliminate the estimation bias as much as possible [Gasser et al., 1986; Buckley et al., 1988; Seifert et al., 1993].

By (6), the bias term is given as

$$\text{bias}(\hat{\sigma}^2) = \frac{g^T D g}{n - r} = \frac{1}{n - r} \sum_{i=1}^{n-r} \left( \sum_{j=0}^{r} d_j g(x_{i+j}) \right)^2.$$  

We further assume that the mean function has a bounded $r$th derivative. Then under the equidistant design, by the Taylor expansion it follows that

$$\sum_{j=0}^{r} d_j g(x_{i+j}) = \sum_{j=0}^{r} d_j \left( \sum_{p=0}^{r} \frac{j^p}{p! n^p} g^{(p)}(x_i) + o\left(\frac{1}{n^r}\right) \right) = \sum_{p=0}^{r} \frac{C_p}{n^p} g^{(p)}(x_i) + o\left(\frac{1}{n^r}\right),$$  

where $g^{(p)}$ denotes the $p$th derivative and $C_p = \sum_{j=0}^{r} j^p d_j / p!$ for $p = 0, 1, \ldots, r$. Finally, by plugging the approximate terms in (8) back to the bias formula, we have

$$\text{bias}(\hat{\sigma}^2) = \frac{1}{n - r} \sum_{i=1}^{n-r} \left( \sum_{p=0}^{r} \frac{C_p}{n^p} g^{(p)}(x_i) + o\left(\frac{1}{n^r}\right) \right)^2.$$  

Noting that the difference sequence $d(r)$ consists of $r + 1$ unknown quantities, one can (and only can) impose a maximum of $r + 1$ constraints on the difference sequence for the purpose of eliminating the estimation bias. To be more specific, apart from the two minimum requirements $C_0 = \sum_{j=0}^{r} d_j = 0$ and $\sum_{j=0}^{r} d_j^2 = 1$ in (4) for model identifiability, we can impose a maximum of $r - 1$ additional constraints $C_1 = \cdots = C_{r-1} = 0$ to further eliminate the bias term in (9). To conclude, the ordinary difference sequence can be redefined as the minimizer of the following
optimization problem:

$$\arg \min_{d(r) \in \mathbb{R}^{r+1}} \delta(r) \text{ subject to } C_0 = C_1 = \cdots = C_{r-1} = 0 \text{ and } \sum_{j=0}^{r} d_j^2 = 1. \quad (10)$$

It is also interesting to point out that, by solving the $r + 1$ constraints, it yields a unique solution of the ordinary difference sequence as $d_{ord}(r) = (d_{ord,0}, \ldots, d_{ord,r})$ where

$$d_{ord,j} = (-1)^j {r \choose j} \left( \frac{2r}{r} \right)^{-1/2}, \quad j = 0, \ldots, r. \quad (11)$$

In other words, the optimization problem in (10) is, in fact, a degenerate optimization problem and the minimum value of $\delta(r)$ is fixed as $\delta(d_{ord}(r))$.

Moreover, with the ordinary difference sequence in (11), we refer to estimator (5) as the ordinary estimator, denoted by $\hat{\sigma}^2_{ord}(r)$. Dette et al. (1998) also derived that

$$\text{var}\{\hat{\sigma}^2_{ord}(r)\} = \frac{1}{n} \left\{ \text{var}(\varepsilon^2) + 4\sigma^4 \delta_{ord}(r) \right\} + o\left(\frac{1}{n}\right) \quad \text{and} \quad \text{bias}\{\hat{\sigma}^2_{ord}(r)\} = O\left(\frac{1}{n^{2r}}\right),$$

where $\delta_{ord}(r) = [(2r)^{-2} (\frac{4r}{2r}) - 1] / 2$. As expected, the ordinary estimator can control the estimation bias up to order $O(n^{-2r})$, compared to the minimum control of $O(n^{-2})$ for the optimal estimator. As a trade-off, however, the ordinary estimator has a larger asymptotic variance than the optimal estimator, especially when the order $r$ is high.

3 Optimal-$k$ difference sequence

3.1 Definition

As shown in Section 2, the optimal and ordinary difference sequences can be derived as the unique solutions under the optimization problems (7) and (10), respectively. Specifically, the first one provides a minimum control on the estimation bias, and the second one may over-control the estimation bias so that, as a price to pay, the estimation variance is dramatically enlarged. To overcome the limitations on the two extreme cases, we propose a compromise solution that
defines the difference sequence as the solution(s) to the following unified optimization problem:

$$\arg \min_{d(r) \in \mathbb{R}^{r+1}} \delta(r) = \sum_{c=1}^{r} \left( \sum_{j=0}^{r-c} d_j d_{j+c} \right)^2$$

subject to \( C_0 = C_1 = \cdots = C_k = 0 \) and \( \sum_{j=0}^{r} d_j^2 = 1 \),

where \( 0 \leq k \leq r - 1 \) is an integer number. In the special case when \( k = 0 \), our unified optimization problem reduces to (7); and in the special case when \( k = r - 1 \), it reduces to (10).

When \( k \) increases from 0 to \( r - 1 \), with the increased number of constraints one can control the estimation bias at a higher order; yet on the other side, with a smaller space for \( d(r) \), the minimum possible variance of estimator (5) will be enlarged as a trade-off.

To further explore the effect of \( k \) on the optimization estimation, we also present the difference sequences in Figure 1 for various \( r \) values. When \( r = 1 \), there is only one unique sequence, represented by the black point in the corner, known also as the Rice difference sequence. When \( r = 2 \), there are two options of difference sequences, one is optimal and the other is ordinary. When \( r \geq 3 \), the optimization problem will produce a new difference sequence for

![Figure 1: The unified framework of optimal-\( k \) difference sequences, where the existing difference sequences (DS) are all located on the boundary of the triangle region as important special cases.](image-url)
each $1 \leq k \leq r - 2$. For ease of presentation, we refer to it as the optimal-$k$ difference sequence and denote it by $d_k(r)$. To summarize, the orange points with $k = 0$ represent the optimal difference sequences, the red points on the diagonal line $k = r - 1$ represent the ordinary difference sequences, and the blue points in the middle part represent the newly introduced optimal-$k$ difference sequences.

### 3.2 Properties and generating procedure

In this section, we propose to minimize the objective function, $\delta(r)$, in (12) and provide an algorithm for generating $d_k(r)$. To start with, we let

$$I_k = \sum_{c=1}^{r} c^k \text{ for } k = 0, 1, \ldots,$$

and

$$V_k = \begin{pmatrix}
I_0 & I_2 & \cdots & I_{2k} \\
I_2 & I_4 & \cdots & I_{2k+2} \\
\vdots & \vdots & \ddots & \vdots \\
I_{2k} & I_{2k+2} & \cdots & I_{4k}
\end{pmatrix}. \tag{13}$$

In the following theorem, we first show that $V_k$ is nonsingular and then derive the minimum value of $\delta(r)$ in closed form for arbitrary optimal-$k$ difference sequence.

**Theorem 1** (a) For any $0 \leq k \leq r - 1$, the constant matrix $V_k$ is an invertible matrix. (b) Let $V_k^{-1}$ be the inverse matrix of $V_k$ and $V_k^{-1}(i,j)$ be the element of $V_k^{-1}$ on the $i$th row and $j$th column. Then for any given $r \geq 1$, the minimum value of $\delta(r)$ under the optimization problem (12) with fixed $k$ is

$$\delta_k(r) = \frac{1}{4} V_k^{-1}(1, 1).$$

The proof of Theorem 1 is given in the Appendix A. By Theorem 1 and the fact that $V_k$ is a constant matrix for any given pair $(r, k)$, we can derive $\delta_k(r)$ for any optimal-$k$ difference sequence. When $k = 0$, we have $V_0 = I_0 = r$ and so $\delta_0(r) = 1/(4r)$, which is the same as $\delta_{\text{opt}}(r)$ in Hall et al. (1990); see also Section 2.2 for more details. When $k = r - 1 \geq 1$, it results in the ordinary difference sequence

$$\delta_{r-1}(r) = \frac{1}{4} V_{r-1}^{-1}(1, 1) = \frac{1}{2} \left\{ \frac{(2r)^2}{r} - \frac{(4r)^2}{2r} - 1 \right\},$$

11
which also coincides with the results in Dette et al. (1998). While for the middle $k$ values such that $1 \leq k \leq r - 2$ with $r \geq 3$, we let $S_{r,k}$ denote the feasible region of (12), i.e., the set of difference sequences satisfying the respective constraints. Then by noting that the constraints in (7), (10) and (12) are nested, we have $S_{r,r-1} \subset S_{r,k} \subset S_{r,0}$, and consequently, $\delta_0(r) \leq \delta_k(r) \leq \delta_{r-1}(r)$. According to Theorem 1, we calculate $\delta_k(r)$ for a range of difference sequences and find that $\delta_k(r)$ increases with $k$ for a fixed $r$, i.e., $\delta_k(r) < \delta_{k+1}(r)$ for $1 \leq k \leq r - 2$, which coincides with the previous analytical results. For a fixed $k$, $\delta_k(r)$ monotonically decreases to zero as $r$ is getting larger. Exceptionally, $\delta_{r-1}(r)$ monotonically increases with $r$ for the ordinary difference sequence.

Except for the ordinary difference sequence in (11), a closed-form solution may not exist for most difference sequences in the optimal-$k$ family. In the following theorem, we show that, for any given pair $(r, k)$, the optimal-$k$ difference sequence from the optimization problem (12) can be alternatively derived as the solution to a root-finding problem.

**Theorem 2** Let $R(t) = t^r \{D_c(t^r + t^{-r}) + \cdots + D_1(t + t^{-1}) + 1\}$ be a self-reciprocal polynomial, where $D_c = \sum_{j=0}^{r-c} d_j d_{j+c} = -\sum_{s=0}^{k} c^{2s} V_k^{-1}(s + 1, 1)/2$ for $c = 1, \ldots, r$. Let also $\{1, z_2, \ldots, z_r, 1, z_2^{-1}, \ldots, z_r^{-1}\}$ be the $2r$ roots from the equation $R(t) = 0$. We further choose one unit root and the $r - 1$ roots outside the unit circle, denoted by $\{t_1, \ldots, t_r\}$, and apply them to construct a new polynomial as

$$(x - t_1)(x - t_2) \cdots (x - t_r) = \sum_{k=0}^{r} a_k x^k z,$$

where $a_k$ are the coefficients. Then the optimal-$k$ difference sequence is given as

$$(d_0, \ldots, d_r) = (a_0, \ldots, a_r) / (\sum_{i=0}^{r} a_i^2)^{1/2}.$$
however, different options of $d_k(r)$ will lead to equivalent estimators and so they can be treated indifferently. We thus apply the algorithm in Theorem 2 to generate the optimal-$k$ difference sequence sequence for any part $(r, k)$ without loss of generality; while for easy reference, Table 1 also provides the numerical results of difference sequences for the order $r$ up to 5.

## 4 Optimal-$k$ estimator

In this section, we apply the optimal-$k$ difference sequence in Section 3 to estimate the residual variance in model (1). We also refer to the new estimator as the optimal-$k$ estimator, denoted by $\hat{\sigma}_k^2(r)$ for $0 \leq k \leq r - 1$. Then as two special cases, the optimal and ordinary estimators are given as $\hat{\sigma}_0^2(r)$ and $\hat{\sigma}_{r-1}^2(r)$, respectively. Now to study the effect of $(r, k)$ on the optimal estimation, we have the following theorem on the asymptotic bias and variance of the optimal-$k$ estimator, with the proof in the Appendix.

**Theorem 3** Assume that the mean function has a continuous $r$th derivative for any given $r \geq 1$

Table 1: Optimal-$k$ difference sequences $d_k(r) = (d_0, \ldots, d_r)$ for $r \leq 5$, where $d_0 > 0$ for identifiability and the entries are rounded to four decimal places.

| $(r, k)$ | $d_0$  | $d_1$  | $d_2$  | $d_3$  | $d_4$  | $d_5$  |
|----------|--------|--------|--------|--------|--------|--------|
| (1, 0)   | 0.7071 | -0.7071|        |        |        |        |
| (2, 0)   | 0.8090 | -0.5000| -0.3090|        |        |        |
| (2, 1)   | 0.4082 | -0.8165| 0.4082 |        |        |        |
| (3, 0)   | 0.8582 | -0.3832| -0.2809| -0.1942|        |        |
| (3, 1)   | 0.2673 | 0.0000 | -0.8018| 0.5345 |        |        |
| (3, 2)   | 0.2236 | -0.6708| 0.6708 | -0.2236|        |        |
| (4, 0)   | 0.8873 | -0.3099| -0.2464| -0.1901| -0.1409|        |
| (4, 1)   | 0.1982 | 0.1034 | -0.1855| -0.7322| 0.6160 |        |
| (4, 2)   | 0.1842 | -0.2271| -0.4242| 0.7928 | -0.3257|        |
| (4, 3)   | 0.1195 | -0.4781| 0.7171 | -0.4781| 0.1195 |        |
| (5, 0)   | 0.9064 | -0.2600| -0.2167| -0.1774| -0.1420| -0.1103|
| (5, 1)   | 0.1573 | 0.1151 | -0.0166| -0.2681| -0.6609| 0.6732 |
| (5, 2)   | 0.1553 | -0.0734| -0.3080| -0.1904| 0.8217 | -0.4053|
| (5, 3)   | 0.1143 | -0.2726| -0.0603| 0.6733 | -0.6470| 0.1923 |
| (5, 4)   | 0.0630 | -0.3150| 0.6299 | -0.6299| 0.3150 | -0.0630|
and $E(\varepsilon^4) < \infty$. Then under the equidistant design, we have
\[
\begin{align*}
\text{bias}\{\hat{\sigma}^2_k(r)\} &= \frac{C^2_{k+1}}{n^{2(k+1)}} J_{k+1} + o\left(\frac{1}{n^{2(k+1)}}\right), \\
\text{var}\{\hat{\sigma}^2_k(r)\} &= \frac{1}{n}\left\{\text{var}(\varepsilon^2) + \sigma^4 V^{-1}_k(1, 1)\right\} + o\left(\frac{1}{n}\right),
\end{align*}
\]
where $C_k = \sum_{j=0}^{r} j^k d_j/k!$ and $J_k = \int_0^1 \{g^{(k)}(x)\}^2 dx$, $k = 0, \ldots, r - 1$.

Note that Theorem 3 also includes the asymptotic results for the existing difference-based estimators. In the special case when $k = 0$, we have the optimal estimator in [Hall et al. (1990)] with the same asymptotic results as
\[
\begin{align*}
\text{bias}\{\hat{\sigma}^2_0(r)\} &= \frac{C^2_1}{n^2} J_1 + o\left(\frac{1}{n^2}\right), \\
\text{var}\{\hat{\sigma}^2_0(r)\} &= \frac{1}{n}\left\{\text{var}(\varepsilon^2) + \frac{1}{r}\sigma^4\right\} + o\left(\frac{1}{n}\right).
\end{align*}
\]
On the other side, when $k = r - 1$ with $r \geq 2$, we have the ordinary estimator in [Dette et al. (1998)] with the same asymptotic results as
\[
\begin{align*}
\text{bias}\{\hat{\sigma}^2_{r-1}(r)\} &= \frac{C^2_r}{n^{2r}} J_r + o\left(\frac{1}{n^{2r}}\right), \\
\text{var}\{\hat{\sigma}^2_{r-1}(r)\} &= \frac{1}{n}\left\{\text{var}(\varepsilon^2) + \left(\frac{2r}{r}\right)^2 \left(\frac{4r}{2r}\right)^{-2} - 2\right\} \sigma^4 \right\} + o\left(\frac{1}{n}\right).
\end{align*}
\]
Now for the new estimator, if we consider $k = 1$ with $r \geq 3$, we have $V^{-1}_1(1, 1) = 4I_4/(I_0 I_4 - I_2^2) \to 9/(4r)$ as $r \to \infty$ and thus
\[
\begin{align*}
\text{bias}\{\hat{\sigma}^2_1(r)\} &= \frac{C^2_2}{n^4} J_2 + o\left(\frac{1}{n^4}\right), \\
\text{var}\{\hat{\sigma}^2_1(r)\} &= \frac{1}{n}\left\{\text{var}(\varepsilon^2) + \frac{9}{4r}\sigma^4\right\} + o\left(\frac{1}{n}\right).
\end{align*}
\]
In what follows, we show that the optimal-1 estimator may provide a reasonable compromise between the optimal and ordinary estimators. For ease of explanation, we assume that $\{\varepsilon_i\}$ are normal errors so that $\text{var}(\varepsilon^2) = 2\sigma^4$. Then to compare the optimal-1 estimator and the optimal estimator in [Hall et al. (1990)], we note that $\text{var}\{\hat{\sigma}^2_1(r)\}/\text{var}\{\hat{\sigma}^2_0(r)\} = 1 + 5/(8r + 4) + o(1)$ and $\text{bias}\{\hat{\sigma}^2_1(r)\}/\text{bias}\{\hat{\sigma}^2_0(r)\} = (C^2_2 J_2)/(C^2_1 J_1) n^{-2} = O(n^{-2})$. To conclude, the optimal-1 estimator significantly reduces the asymptotic bias compared to the optimal estimator.
yet as a price to pay, there is a little sacrifice in the asymptotic variance, which however can be small when \( r \) is three or more. Next, for a comparison between the optimal-1 estimator and the ordinary estimator in Gasser et al. (1986), we note that

\[
\frac{\text{var}\{\hat{\sigma}_1^2(r)\}}{\text{var}\{\hat{\sigma}_{r-1}^2(r)\}} = \frac{(\pi r/2)^{-1/2} + o(r^{-1/2})}{\text{bias}\{\hat{\sigma}_1^2(r)\} / \text{bias}\{\hat{\sigma}_{r-1}^2(r)\}} = O(n^{2(r-2)}).
\]

To conclude, the optimal-1 estimator significantly reduces the asymptotic variance compared to the ordinary estimator; while for the asymptotic bias, given that the optimal-1 estimator has already controlled the order at \( n^{-4} \), a further reduction may not bring in significant improvement in practice.

For further comparison, we also compute the numerical values of the asymptotic variance and bias according to Theorem 3 for \( \{(r, k) : 1 \leq r \leq 10, \ 0 \leq k \leq \min(3, r - 1)\} \), and then transform them into the relative variance (RVAR), squared bias (RSB), and the relative mean squared error (RMSE) by dividing the MSE by the scaling factor \( \text{var}(\varepsilon^2)/n \). As shown in Figure 2, RVAR decreases to one for \( \hat{\sigma}_k^2(r) \) with \( 0 \leq k \leq 3 \) but increases for \( \hat{\sigma}_{r-1}^2(r) \) as \( r \) becomes larger, which coincides with the functional pattern of \( \delta_k(r) \). Due to the poorly-controlled estimation bias, \( \hat{\sigma}_0^2(r) \) breaks down when \( r \) is large. In contrast, thanks to a better balance on the bias-variance trade-off, the minimum RMSE for \( \hat{\sigma}_k^2(r) \) with \( 1 \leq k \leq 3 \) is always smaller than that for the optimal and ordinary estimators. To conclude, when estimating the residual variance, it is desirable to search for the whole optimal-\( k \) family rather than restricting the attention only to the
two extreme cases.

5 Simulation studies

This section compares the finite sample performance of the new and existing estimators within the optimal-\(k\) family. Noting that the optimal and ordinary estimators also belong to the optimal-\(k\) family, to avoid confusion we refer to the optimal-\(k\) estimators specifically as the optimal-\(k\) estimators with \(1 \leq k \leq r - 2\), i.e. the difference-based estimators associated with the blue points in Figure 1. By [Dette et al. (1998)](https://link.to/Dette1998), an order of \(r > 4\) is rarely used in practice due to its complexity and also unstable performance. This leads us to consider all the combinations of \((r,k)\) with \(r \leq 4\), which yields 10 difference-based estimators in total, including the Rice estimator, three optimal estimators, three ordinary estimators, and three optimal-\(k\) estimators.

To evaluate the performance of the difference-based estimators, we also follow the same mean function that is commonly used in the literature ([Hall et al., 1990](https://link.to/Hall1990), [Seifert et al., 1993](https://link.to/Seifert1993), [Dette et al., 1998](https://link.to/Dette1998)):

\[
g(x) = 5\sin(w\pi x),
\]

where \(w > 0\) controls the oscillation level of the mean function. Moreover, we let \(x_i = i/n\) be equally spaced design points, \(\{\varepsilon_i\}\) be a random sample of size \(n\) from \(N(0, \sigma^2)\), and \(n = 25, 100, 500\) represent three different sample sizes. We further take \(\{(\sigma, \omega) : \sigma = 0.2, 0.4, \ldots, 2 \text{ and } \omega = 0.5, 1, \ldots, 5\}\), which forms a total of 100 combinations for the signal-to-noise ratios for the simulated data. Finally, with 10000 simulations for each setting, we compute the RMSE for each estimator, defined as \((n/2\sigma^4)\text{MSE}\), and report the estimator with the smallest RMSE in Table 2.

From Table 2, it is evident that the optimal-\(k\) estimators provide the best performance in many settings, especially when the sample size is moderate to small. This coincides with the theoretical results, as well as the motivation, that the optimal-\(k\) difference sequence offers a good balance between the estimation bias and the estimation variance. Moreover, the comparison results between the Rice, optimal and ordinary estimators remain the same as observed in [Dette et al. (1998)](https://link.to/Dette1998). Specifically, when the sample size is large, the ordinary estimators are often suboptimal due to the large estimation variance. In contrast, when the sample size is small, the optimal estimators
Table 2: Optimal combination of $(r, k)$ under each of the 100 settings with the sample size $n = 500, 100$ and $25$. White cells: existing estimators; grey cells: new estimators. The $(r, k)$ value of the best option is provided in each cell.

| $n=500$ | $2-0.2$ | $2-0.4$ | $2-0.6$ | $2-0.8$ | $2-1-0.2$ | $2-1-0.4$ | $2-1-0.6$ | $2-1-0.8$ | $2-1-1$ | $2-1-1.2$ | $2-1-1.4$ | $2-1-1.6$ | $2-1-1.8$ | $2-1-2$ | $2-1-2.5$ | $2-1-3$ | $2-1-3.5$ | $2-1-4$ | $2-1-4.5$ | $2-1-5$ |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| $\sigma$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(4,0)$ |
| $\omega$ | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |

| $n=100$ | $2-0.2$ | $2-0.4$ | $2-0.6$ | $2-0.8$ | $2-1-0.2$ | $2-1-0.4$ | $2-1-0.6$ | $2-1-0.8$ | $2-1-1$ | $2-1-1.2$ | $2-1-1.4$ | $2-1-1.6$ | $2-1-1.8$ | $2-1-2$ | $2-1-2.5$ | $2-1-3$ | $2-1-3.5$ | $2-1-4$ | $2-1-4.5$ | $2-1-5$ |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| $\sigma$ | $(4,0)$ | $(4,0)$ | $(4,0)$ | $(3,0)$ | $(3,0)$ | $(3,0)$ | $(3,0)$ | $(3,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ | $(2,0)$ |
| $\omega$ | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |

| $n=25$ | $2-0.2$ | $2-0.4$ | $2-0.6$ | $2-0.8$ | $2-1-0.2$ | $2-1-0.4$ | $2-1-0.6$ | $2-1-0.8$ | $2-1-1$ | $2-1-1.2$ | $2-1-1.4$ | $2-1-1.6$ | $2-1-1.8$ | $2-1-2$ | $2-1-2.5$ | $2-1-3$ | $2-1-3.5$ | $2-1-4$ | $2-1-4.5$ | $2-1-5$ |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| $\sigma$ | $(3,0)$ | $(3,0)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ | $(4,1)$ |
| $\omega$ | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |

tend to be less satisfactory because of the uncontrolled estimation bias, especially when the mean function is also very rough. Lastly, the Rice estimator fails to provide the best performance in all of the settings. To sum up, with the newly introduced optimal-$k$ family, it has made possible for researchers to dramatically improve the existing difference-based estimation in nonparametric regression.

To further explore the optimal-$k$ estimators, our second simulation is to conduct a cross-sectional study for assessing the effect of $k$ on the variance estimation when the order $r$ is given.
We first consider $r = 3$, which contains three difference-based estimators, namely the optimal estimator $\hat{\sigma}_{opt}^2(3)$, the ordinary estimator $\hat{\sigma}_{ord}^2(3)$, and the optimal-$k$ estimator $\hat{\sigma}_1^2(3)$. We further take $(\sigma, \omega) = (0.5, 1), (1.5, 1), (0.5, 4) \text{ or } (1.5, 4)$, and $n = 25, 50, \ldots, 500$. All other settings remain the same as before. As shown in Figure 3 under the log scale, the optimal estimator is able to provide the smallest RMSE when the sample size is sufficiently large, but it tends to collapse easily when the sample size becomes smaller due to the poorly controlled estimation bias. In contrast, the RMSE of the ordinary estimator is very stable along with the sample size. As a consequence, the ordinary estimator may not be optimal in the asymptotic sense, but it can
provide a good estimate in the small sample size setting. Lastly, from Figure 3, it is also evident that the optimal-$k$ estimator performs consistently well in most settings and turns out to be the best choice among the three difference-based estimators with $r = 3$.

For additional information, we also present the cross-sectional study with $r = 4$ in Appendix B. From the results in Figure S1, it is evident again that both of the optimal-$k$ estimators outperform the two existing competitors. To further compare the two optimal-$k$ estimators, we note that $\hat{\sigma}_1^2(4)$ provides a smaller RMSE than $\hat{\sigma}_2^2(4)$ in most settings, showing that the bias correction with $k = 1$ is usually enough for the difference-based estimation. But if, in practice, small sample
sizes are a concern, then $\hat{\sigma}_2^2(4)$ can be preferred as well.

Finally, to identify the best difference-based estimator with $r \leq 4$, we present the five local winners, including $\hat{\sigma}_{opt}^2(2)$, $\hat{\sigma}_{ord}^2(2)$, $\hat{\sigma}_1^2(3)$, $\hat{\sigma}_1^2(4)$ and $\hat{\sigma}_2^2(4)$, in Figure 4 under the log scale of RMSE. As expected, the three optimal-$k$ estimators all perform very well and they are also significantly better than the two estimators with $r = 2$. Among the three new estimators, we also note that the RMSE of $\hat{\sigma}_1^2(3)$ is always in the middle of those for the two optimal-$k$ estimators with $r = 4$. In view of this, if we take into account both efficiency and robustness, $\hat{\sigma}_1^2(3)$ can be recommended as the final winner among all the candidate estimators. Another benefit for recommending $\hat{\sigma}_1^2(3)$ is that the second coefficient of $d_1(3) = (0.2673, 0, -0.8018, 0.5345)$ is zero, and subsequently this estimator can be implemented as simply as the difference-based estimators with $r = 2$. Taken together, we are now confident to recommend the optimal-$k$ difference sequence $d_1(3)$ for practical use, and for which it may also be considered as a rule of thumb.

6 Conclusion

Difference-based methods have been increasingly used in nonparametric regression, especially for estimating the residual variance. One fundamental problem, that is also of practical importance, is the difference sequence selection which has been studied for over three decades. The main contribution of this paper is to further advance the literature by introducing the optimal-$k$ difference sequence to better balance the bias-variance trade-off. More importantly, it also dramatically enlarges the existing family of difference sequences that includes the optimal and ordinary difference sequences as two important special cases. To highlight a few key steps, we first reformulate the existing difference sequences as solutions to an optimization problem that minimizes the variance of the estimator under certain constraints. We then derive the optimal-$k$ difference sequence under the proposed optimization framework by providing more flexible constraints on the estimation bias and variance. The asymptotic properties of the optimal-$k$ estimator are also established. Finally, we conduct extensive simulations to evaluate the difference-based estimators within the optimal-$k$ family, and more importantly identify the best difference sequence $d_1(3) = (0.2673, 0, -0.8018, 0.5345)$ for practical use.

Besides the new advances in finding the best difference sequence, another breakthrough in the difference-based methods is known to be the least squares estimator introduced by Tong and Wang (2005) and Tong et al. (2013). Specifically, they regressed the first order differences of
paired observations on the squared distances between the paired covariates via a simple linear regression model, and then applied in spirit the simulation-extrapolation (SIMEX) method to estimate the error variance as the intercept. Dai et al. (2017) extended their least squares estimator to form a unified framework by employing the higher-order differences among the observations as the regressors, and within the unified framework, they further concluded that the ordinary difference sequence can be consistently better than the optimal difference sequence. As an interesting future work, we will also incorporate the newly developed optimal-\(k\) difference sequence to the unified framework and see whether, or to what extent, the new difference sequence will alter the conclusion made in Dai et al. (2017).

Last but not least, to simplify the presentation of the main ideas, we have restricted our attention to the simple nonparametric regression model in (1) in the current paper. It is worth noting, however, that our proposed optimal-\(k\) difference sequence is general and can also be applied to the difference-based estimation in many other models. To name a few such scenarios, the difference-based methods have been extended to estimate the residual variance in multivariate nonparametric regression models (Munk et al., 2005), in partial linear models (Wang et al., 2011), and in nonparametric regression models with repeated measurements (Dai et al., 2015). Other than the residual variance, the difference-based methods have also been effectively applied to estimate other quantities including, for example, the variance function or variogram (Bliznyuk et al., 2012; Yang and Zhu, 2015), the autocovariance (Hall and Keilegom, 2003; Tecuapetla-Gómez and Munk, 2016; Cui et al., 2021), the error distribution Chang et al. (2018), the derivatives of the mean function (Dai et al., 2016; Liu and Brabanter, 2018; Wang et al., 2019), and the long-run variance in time series (Chan, 2022). Further research is needed to explore the potential applications of the optimal-\(k\) difference sequence under these more general settings.

**References**

Bliznyuk, N., Carroll, R. J., Genton, M. G., and Wang, Y. (2012). Variogram estimation in the presence of trend. *Statistics and Its Interface*, 5:159–168.

Buckley, M. J., Eagleson, G. K., and Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika*, 75:189–199.
Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.

Chan, K. W. (2022). Optimal difference-based variance estimators in time series: A general framework. *The Annals of Statistics*, 50:1376–1400.

Chang, J., Delaigle, A., Hall, P., and Tang, C. Y. (2018). A frequency domain analysis of the error distribution from noisy high-frequency data. *Biometrika*, 105:353–369.

Cui, Y., Levine, M., and Zhou, Z. (2021). Estimation and inference of time-varying autocovariance under complex trend: A difference-based approach. *Electronic Journal of Statistics*, 15:4264–4294.

Dai, W., Ma, Y., Tong, T., and Zhu, L. (2015). Difference-based variance estimation in nonparametric regression with repeated measurement data. *Journal of Statistical Planning and Inference*, 163:1–20.

Dai, W., Tong, T., and Genton, M. G. (2016). Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research*, 17(164):1–25.

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

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.

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

Gasser, T., Kneip, A., and Kohler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association*, 86:643–652.

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

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations, 3rd Ed*. Baltimore: Johns Hopkins University Press.
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 Keilegom, I. V. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *Journal of the Royal Statistical Society, Series B*, 65:443–456.

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

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

Jiu, L. and Li, Y. (2021). Hankel determinants of certain sequences of Bernoulli polynomials: A direct proof of an inverse matrix entry from statistics. *arXiv:2109.00772*

Liu, Y. and Brabanter, K. D. (2018). Derivative estimation in random design. *Advances in Neural Information Processing Systems*, 31:3445–3454.

Müller, H. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *The Annals of Statistics*, 27:299–337.

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

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

Tecuapetla-Gómez, I. and Munk, A. (2016). Autocovariance estimation in regression with a discontinuous signal and \(m\)-dependent errors: A difference-based approach. *Scandinavian Journal of Statistics*, 44:346–368.

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.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

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

Wang, W. W., Yu, P., Lin, L., and Tong, T. (2019). Robust estimation of derivatives using locally weighted least absolute deviation regression. *Journal of Machine Learning Research*, 20(60):1–49.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. New York: Chapman and Hall.

Yang, S. and Zhu, Z. (2015). Variance estimation and kriging prediction for a class of non-stationary spatial models. *Statistica Sinica*, 25:135–149.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. New York: Cambridge University Press.
Appendix A

To prove the main results, we first introduce a lemma on the equivalence of two groups of constraints, by which we can transform the nonlinear optimization problem (12) into a linear optimization problem so that a closed-form solution for minimizing $\delta(r)$ can be derived.

**Lemma 1** Assume that $C_0 = 0$. Then for any $1 \leq k \leq r - 1$, the constraints $C_1 = \cdots = C_k = 0$ are equivalent to

$$
\sum_{c=1}^{r} c^{2s} D_c = 0, \quad s = 1, \ldots, k,
$$

where $D_c = \sum_{j=0}^{r-c} d_j d_{j+c}$, for $c = 1, \ldots, r$, are the same as defined in Theorem 2.

Proof of Lemma 1. By the definition of $D_c$, for any integer $s \geq 1$, we have

$$
\sum_{c=1}^{r} c^{2s} D_c = \sum_{c=1}^{r} \sum_{j=0}^{r-l} c^{2s} d_j d_{j+c} = \sum_{k_1-k_2=1}^{r} (k_1 - k_2)^{2s} d_{k_1} d_{k_2}
$$

$$
= \sum_{k_1-k_2=1}^{r} \left[ \sum_{t=0}^{2s} \binom{2s}{t} k_1^t k_2^{2s-t} (-1)^{2s-t} \right] d_{k_1} d_{k_2}
$$

$$
= \sum_{t=0}^{2s} (-1)^{2s-t} \binom{2s}{t} \sum_{k_1-k_2=1}^{r} k_1^t k_2^{2s-t} d_{k_1} d_{k_2}
$$

$$
= \frac{1}{2} \sum_{t=0}^{2s} (-1)^{2s-t} \binom{2s}{t} \left( \sum_{k_1=0}^{r} k_1^t d_{k_1}^t \right) \left( \sum_{k_2=0}^{r} k_2^{2s-t} \right)
$$

$$
= \frac{1}{2} \sum_{t=0}^{2s} (-1)^{2s-t} \binom{2s}{t} C_t C_{2s-t}! (2s-t)!.
$$

To show the equivalence, we first assume that $C_1 = \cdots = C_k = 0$ holds. Then for any $s \leq k$, we have $\min(t, 2s - t) \leq s \leq k$ for any given $t$. Consequently,

$$
C_t = 0 \quad \text{or} \quad C_{2s-t} = 0, \quad t = 1, \ldots, 2s - 1.
$$

Moreover, by (14) and the assumption $C_0 = 0$, it yields that $\sum_{c=1}^{r} c^{2s} D_c = 0$ for $s = 1, \ldots, k$.

We now assume that $\sum_{c=1}^{r} c^{2s} D_c = 0$ holds for $s = 1, \ldots, k$. In what follows, we show that $C_1 = \cdots = C_k = 0$ by induction.
(a) When \( k = 1 \), we have \( \sum_{c=1}^{r} c^2 D_c = 0 \). Then by \((14)\), it yields that

\[
0 = \sum_{c=1}^{r} c^2 D_c = - \left( \sum_{k_1=0}^{r} k_1 d_{k_1} \right) \left( \sum_{k_2=0}^{r} k_2 d_{k_2} \right) = - \left( \sum_{k=0}^{r} k_1 d_{k_1} \right)^2,
\]

which leads to \( \sum_{k=0}^{r} k_1 d_{k_1} = 0 \), i.e. \( C_1 = 0 \).

(b) Assume that the conclusion holds for \( k = \tilde{k} \). That is, when \( \sum_{c=1}^{r} c^2 D_c = 0 \) for \( s = 1, \ldots, \tilde{k} \), it follows that \( C_1 = \cdots = C_{\tilde{k}} = 0 \). Then for \( k = \tilde{k} + 1 \), by \((14)\) and \( \sum_{c=1}^{r} c^2(\tilde{k}+1) D_c = 0 \), we have

\[
\frac{1}{2} \sum_{t=0}^{2(\tilde{k}+1)} (-1)^{2(\tilde{k}+1)-t} \binom{2(\tilde{k}+1)}{t} C_t C_{2(\tilde{k}+1)-t} t!(2(\tilde{k}+1) - t)! = 0. \quad (15)
\]

Moreover, by the assumptions that \( C_0 = C_1 = \cdots = C_{\tilde{k}} = 0 \), \((15)\) can be simplified as

\[
\sum_{c=1}^{r} c^2(\tilde{k}+1) D_c = \frac{1}{2} (-1)^{\tilde{k}+1} \binom{2(\tilde{k}+1)}{\tilde{k}+1} C_{\tilde{k}+1}^2 ((\tilde{k}+1)!)^2 = 0.
\]

This shows that \( C_{\tilde{k}+1} = 0 \) and thus the conclusion also holds for \( k = \tilde{k} + 1 \). This completes the proof of Lemma 1.

**Proof of Theorem 1**

(a) To prove the invertibility of \( V_k \), we first consider \( k = r - 1 \) and decompose the matrix as \( V_{r-1} = U_r^T U_r \), where \( U_r \) is the \( r \times r \) Vandermonde matrix with form

\[
U_r = \begin{pmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & 2^2 & 2^4 & \cdots & 2^{2(r-1)} \\
1 & 3^2 & 3^4 & \cdots & 3^{2(r-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & r^2 & r^4 & \cdots & r^{2(r-1)}
\end{pmatrix}.
\]

The decomposition holds by noting that

\[
V_{r-1}(i, j) = I_{2(i+j)-4} = \sum_{t=1}^{r} I_{2(i-1)t^2(j-1)} = U_r(\cdot, i)^T U_r(\cdot, j),
\]
where $U_r(\cdot, i)$ denotes the $i$th column of $U_r$. According to the properties of Vandermonde matrix (Golub and Van Loan, 1996; Jiu and Li, 2021), we further have

$$|U_r| = \prod_{1 \leq i < j \leq r} (j^2 - i^2) \neq 0,$$

so that $U_r$ is an invertible matrix. This shows that $V_{r-1}$ is a positive-definite matrix with all the principal submatrices invertible; that is $|V_k| \neq 0$ for any $0 \leq k \leq r - 1$.

(b) For the equidistant design, when $k \geq 1$ in (12), the optimization problem is given as

$$\arg \min_{d(r) \in \mathbb{R}^{r+1}} \delta(r) \quad \text{subject to} \quad C_0 = \cdots = C_k = 0 \quad \text{and} \quad \sum_{j=0}^{r} d_j^2 = 1. \quad (16)$$

By Lemma 1, the constraints in (16) can be equivalently expressed as

$$\sum_{c=1}^{r} c^{2s} D_c = 0, \quad s = 1, \ldots, k, \quad \text{and} \quad \sum_{c=1}^{r} D_c = -\frac{1}{2} \sum_{j=0}^{r} d_j^2 = -\frac{1}{2}. \quad (17)$$

Consequently, the optimization problem becomes

$$\arg \min_{d(r) \in \mathbb{R}^{r+1}} \sum_{c=1}^{r} D_c^2 \quad \text{subject to} \quad \sum_{c=1}^{r} c^{2s} D_c = 0, \quad s = 1, \ldots, k, \quad \text{and} \quad \sum_{c=1}^{r} D_c = -\frac{1}{2}. \quad (18)$$

To apply the method of Lagrange multipliers, we let

$$L(D_1, \ldots, D_r, \lambda_0, \ldots, \lambda_k) = \sum_{c=1}^{r} D_c^2 + \lambda_0 \left( \sum_{c=1}^{r} D_c + \frac{1}{2} \right) + \sum_{s=1}^{k} \left( \lambda_s \sum_{c=1}^{r} c^{2s} D_c \right).$$

Then by taking the partial derivatives of $L$ and setting them as zero, we have

$$\frac{\partial L}{\partial D_c} = 2D_c + \lambda_0 + \sum_{s=1}^{k} c^{2s} \lambda_s = 0, \quad c = 1, \ldots, r, \quad (17)$$

$$\frac{\partial L}{\partial \lambda_0} = \sum_{c=1}^{r} D_c + \frac{1}{2} = 0, \quad (18)$$

$$\frac{\partial L}{\partial \lambda_s} = \sum_{c=1}^{r} c^{2s} D_c = 0, \quad s = 1, \ldots, k. \quad (19)$$
Moreover, by taking the weighted sum of the \( r \) equations in (19), it yields that
\[
\sum_{c=1}^{r} c^{2t} \frac{\partial L}{\partial D_c} = 2 \sum_{c=1}^{r} c^{2t} D_c + \lambda_0 \sum_{c=1}^{r} c^{2t} + \sum_{s=1}^{k} \lambda_s \sum_{c=1}^{r} c^{2(s+t)} = 0, \quad t = 1, \ldots, k.
\]

By (18) and (19), the above equations can be expressed as
\[
V_k \Lambda = e_1,
\]
where \( \Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_k)^T \) and \( e_1 = (1, 0, \ldots, 0)^T \). By solving Equation (20), we get
\[
\Lambda = (V_k^{-1}(1, 1), \ldots, V_k^{-1}(k+1, 1))^T. \tag{21}
\]
Further by combining (17), (18) and (19), we get the minimum value of \( \delta(r) \) as
\[
\delta_k(r) = \sum_{c=1}^{r} D_c^2 = -\frac{1}{2} \sum_{c=1}^{r} D_c (\lambda_0 + \sum_{s=1}^{k} c^{2s} \lambda_s)
\]
\[
= -\frac{1}{2} \lambda_0 \sum_{c=1}^{r} D_c - \frac{1}{2} \sum_{s=1}^{k} \lambda_s \sum_{c=1}^{r} c^{2s} D_c
\]
\[
= \frac{\lambda_0}{4}.
\]
By (21), we have \( \lambda_0 = V_k^{-1}(1, 1) \) and hence \( \delta_k(r) = V_k^{-1}(1, 1)/4. \)

**Proof of Theorem 2**

To generate the optimal-\( k \) difference sequence, we first construct a polynomial \( R(t) \) in the following form
\[
R(t) = t^r \{ D_r(t^r + t^{-r}) + \cdots + D_1(t + t^{-1}) + 1 \}
\]
\[
= t^r(d_0 + d_1 t^1 + \cdots + d_r t^r)(d_0 + d_1 t^{-1} + \cdots + d_r t^{-r}).
\]

\( R(t) \) is a self-reciprocal polynomial, which means that if \( z \) is a root of \( R(t) \) then \( 1/z \) must also be a root of \( R(t) \). Also, if a complex number \( z \) is a root of \( R(t) \) then its conjugate \( \bar{z} \) is also a root of \( R(t) \) since \( D_c \)'s are real.

In \( R(t) \), the coefficients \( D_c \)'s can be calculated according to Theorem 1 and hence we can explicitly get all the \( 2r \) roots of \( R(t) \) via any root-finding algorithm. Note that \( \sum_{c=1}^{r} D_c = -1/2, \)
so we know that \( t = 1 \) is a double root of \( R(t) \) and the 2r root of \( R(t) \) can be expressed as

\[
RT = \{z_1, z_2, \ldots, z_r, z_1^{-1}, z_2^{-1}, \ldots, z_r^{-1}\} \text{ with } z_1 = z_1^{-1} = 1.
\]

To restore the polynomial \( d_0 + d_1 t + \cdots + d_r t^r \) and hence get the desired difference sequence as the normalized coefficients, we adopt the following two criteria to select \( r \) roots from \( RT \): (i) Choose any set of \( r \) roots with no repeat of the index. (ii) If a complex root \( z \in RT \) is selected, then its conjugate \( \bar{z} \in RT \) should also be selected. The first criterion is to ensure that the restored polynomial is proportional to the desired one, i.e.,

\[
\prod_{i=1}^{r} (t - z_i) = a_0 + a_1 t + \cdots + a_r t^r = (a_0^2 + \cdots + a_r^2)^{1/2} (d_0 + d_1 t + \cdots + d_r t^r).
\]

The second criterion is to ensure that the coefficients of the restored polynomial are real numbers.

Various valid sets of roots exist, among which we recommend simply separating the roots with the unit circle and choosing the roots on or outside the circle as described in Theorem 2, which satisfies the two criteria.

**Proof of Theorem**

The asymptotic bias of \( \hat{\sigma}^2_k(r) \) is an immediate result from equation (9). While for the asymptotic variance, according to (6), we have

\[
\begin{align*}
\var(\hat{\sigma}^2_k(r)) &= \frac{1}{(n-r)^2} \left[ 4\sigma^2 g^T D^2 g + 4g^T \{D \text{diag}(D)u\} \sigma^3 \gamma_3 \\
&+ \sigma^4 \text{tr}\{\text{diag}(D)^2\} (\gamma_4 - 3) + 2\sigma^4 \text{tr}(D^2) \right].
\end{align*}
\]

(22)

By Dette et al. (1998), together with the definition of \( D \) and the assumption that the mean function has a continuous \( r \)th derivative, we can derive that

\[
\begin{align*}
g^T D^2 g &= \|Dg\|_2^2 = O(1), \\
g^T \{D \text{diag}(D)u\} &= O(1), \\
\text{tr}\{\text{diag}(D)^2\} &= n + o(n), \\
\text{tr}(D^2) &= n\{1 + 2\delta(r) + o(1)\}.
\end{align*}
\]

29
Plugging them back to (22),

\[
\var(\hat{\sigma}^2_k(r)) = \frac{1}{n} \{\var(\varepsilon^2) + 4\sigma^4 \delta(r)\} + o\left(\frac{1}{n}\right).
\]

This proves the theorem by noting that \(\delta(r) = V_k^{-1}(1, 1)/4\) from Theorem 1.

**Appendix B**

This appendix presents the cross-sectional study with \(r = 4\), which includes a total of 4 estimators including the optimal estimator \(\hat{\sigma}^2_{\text{opt}}(4)\), the ordinary estimator \(\hat{\sigma}^2_{\text{ord}}(4)\), and two optimal-\(k\) estimators \(\hat{\sigma}^2_1(4)\) and \(\hat{\sigma}^2_2(4)\). For a fair comparison, we follow the same settings as in Figure 3 and report the logarithm of RMSE for each estimator in Figure S1.
Figure S1: Plots of log(RMSE) as functions of $n$ for the four estimators with $r = 4$: $\hat{\sigma}_{\text{opt}}^2(4)$ (dotted lines), $\hat{\sigma}_{\text{ord}}^2(4)$ (dashed lines), $\hat{\sigma}_1^2(4)$ (red solid lines) and $\hat{\sigma}_2^2(4)$ (red dash-dotted lines). The results are based on 10000 simulations.