Statistical Science
2017, Vol. 32, No. 3, 455–468
DOI: 10.1214/17-STS613

On the Choice of Difference Sequence in a Unified Framework for Variance Estimation in Nonparametric Regression

Wenlin Dai, Tiejun Tong and Lixing Zhu

Abstract. Difference-based methods do not require estimating the mean function in nonparametric regression and are therefore popular in practice. In this paper, we propose a unified framework for variance estimation that combines the linear regression method with the higher-order difference estimators systematically. The unified framework has greatly enriched the existing literature on variance estimation that includes most existing estimators as special cases. More importantly, the unified framework has also provided a smart way to solve the challenging difference sequence selection problem that remains a long-standing controversial issue in nonparametric regression for several decades. Using both theory and simulations, we recommend to use the ordinary difference sequence in the unified framework, no matter if the sample size is small or if the signal-to-noise ratio is large. Finally, to cater for the demands of the application, we have developed a unified R package, named VarED, that integrates the existing difference-based estimators and the unified estimators in nonparametric regression and have made it freely available in the R statistical program http://cran.r-project.org/web/packages/.

Key words and phrases: Difference-based estimator, nonparametric regression, optimal difference sequence, ordinary difference sequence, residual variance.

1. INTRODUCTION

We consider the nonparametric regression model:

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

where $\{Y_i\}$ are the observations, g is an unknown mean function, $\{x_i\}$ are the design points and $\{\varepsilon_i\}$ are the independent and identically distributed (i.i.d.) random errors with mean zero and variance σ^2. Needless to say, nonparametric regression models are very useful in statistics and have been extensively studied in the past several decades. There is a large body of literature on the estimation of the mean function g, for example, the kernel estimators, the local linear estimators and the smoothing spline estimators. Apart from the mean function, the variance estimation has also been recognized as an important problem in nonparametric regression. An accurate yet economic estimator of σ^2 is required in many aspects of nonparametric regression, for example, in the construction of confidence intervals, in testing the goodness of fit, and in choosing the amount of smoothing (Rice, 1984, Eubank and Spiegelman, 1990, Gasser, Kneip and Kohler, 1991, Härdle and Tsybakov, 1997).

To estimate the residual variance, researchers often apply the sum of squared residuals from a nonparametric fit, that is,

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

Wenlin Dai is Postdoctoral Fellow, CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia (e-mail: wenlin.dai@kaust.edu.sa). Tiejun Tong is Associate Professor and Corresponding Author, Department of Mathematics, Hong Kong Baptist University, Hong Kong (e-mail: tongt@hkbu.edu.hk). Lixing Zhu is Chair Professor, Department of Mathematics, Hong Kong Baptist University, Hong Kong (e-mail: lzhu@hkbu.edu.hk).

where $\hat{g}$ is the fitted mean function and ν is the number of degrees of freedom for the fitted model. Estimators with form (1) are referred to as residual-based estimators. With the optimal bandwidth, the minimum mean squared error (MSE) of $\hat{\sigma}^2$ is given as

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma}^2) &= \mathrm{var}(\hat{\sigma}^2) + \{E(\hat{\sigma}^2) - \sigma^2\}^2 \\ &= \frac{1}{n}\mathrm{var}(\varepsilon^2) + o\left(\frac{1}{n}\right).\end{aligned} \tag{2}$$

Hall and Marron (1990) showed that the MSE in (2) is asymptotically optimal in a minimax sense. Nevertheless, it is known that residual-based estimators have two major limitations. First, residual-based estimators depend heavily on the delicate choice of tuning parameters so that their practical applications are somewhat limited (Dette, Munk and Wagner, 1998). Second, residual-based estimators are completely determined by the fitted $\hat{g}$ from a nonparametric fit (Eubank and Spiegelman, 1990, Ye, 1998, Wang, 2011), and then consequently, the constructed confidence intervals and the goodness of fit test may not be reliable if such variance estimates are used. In a Bayesian framework, when the noninformative prior density $p(\sigma^2) \propto 1/\sigma^2$ is used, the variance estimation will also rely heavily on the estimated $\hat{g}$ (Berkey, 1982, Smith and Kohn, 1996).

For the demand of an estimate of σ^2 that is independent of the fitted mean function, researchers have proposed another class of estimators in the literature. These estimators use the differences between nearby observations to remove the trend in the mean function, and are the so-called difference-based estimators. For simplicity of notation, we assume an equally spaced design with $x_i = i/n$ for $i = 1, \ldots, n$. Let $r > 0$ be an integer number and $(d_0, \ldots, d_r)$ be a difference sequence with

$$\sum_{j=0}^{r} d_j = 0 \quad \text{and} \quad \sum_{j=0}^{r} d_j^2 = 1, \tag{3}$$

where $d_0 d_r \neq 0$, $d_0 > 0$, and $d_j = 0$ for $j < 0$ and $j > r$. Hall, Kay and Titterington (1990) proposed a general form of difference-based estimators:

$$\hat{\sigma}^2(r) = \frac{1}{n-r}\sum_{i=1}^{n-r}\left(\sum_{j=0}^{r} d_j Y_{j+i}\right)^2, \tag{4}$$

where r is the order of the variance estimator. Difference-based estimators do not require an estimate of the mean function and are attractive from a practical point of view. When $r = 1$, the unique solution of the difference sequence under the constraint (3) is $(d_0, d_1) = (2^{-1/2}, -2^{-1/2})$ and it yields the first-order difference-based estimator in Rice (1984),

$$\hat{\sigma}_{\mathrm{R}}^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}(Y_i - Y_{i+1})^2. \tag{5}$$

When $r \geq 2$, however, there are infinitely many solutions for $(d_0, \ldots, d_r)$ under the constraint (3). Among them, two commonly recommended difference sequences are: (i) *the optimal difference sequence* and (ii) *the ordinary difference sequence*.

The optimal difference sequence is obtained by minimizing the asymptotic MSE of the estimator with form (4). Under some smoothness conditions on g, Hall, Kay and Titterington (1990) showed that the effect of g on the estimation bias is asymptotically negligible. Then asymptotically, to minimize the MSE of the estimator is equivalent to minimizing the variance of the estimator. This leads to the optimal difference sequence satisfying $\sum_{j=0}^{r} d_j d_{j+i} = -1/2r$ for $1 \leq i \leq r$. For the special case of $r = 2$, the optimal difference sequence is $(d_0, d_1, d_2) = (0.809, -0.5, -0.309)$ and the resulting estimator is $\hat{\sigma}_{\mathrm{opt}}^2(2) = \sum_{i=1}^{n-2}(0.809Y_i - 0.5Y_{i+1} - 0.309Y_{i+2})^2/(n-2)$. We refer to the estimator (4) with the optimal difference sequence as $\hat{\sigma}_{\mathrm{opt}}^2(r)$.

When the sample size is small, however, the bias term of difference-based estimators may not be negligible, in particular when the mean function g is very rough. To reduce the bias, Dette, Munk and Wagner (1998) recommended to apply the following difference sequence:

$$d_j = (-1)^j \binom{2r}{r}^{-1/2} \binom{r}{j}, \quad j = 0, \ldots, r. \tag{6}$$

This difference sequence was commonly employed for numerical differentiation and was referred to as *the ordinary difference sequence* in Hall, Kay and Titterington (1990). With the difference sequence (6), the estimation bias of $\hat{\sigma}^2(r)$ vanishes for polynomials up to degree $r-1$. For this, we may also refer to it as *the debiased difference sequence* or *the polynomial weight sequence* (Dette, Munk and Wagner, 1998). For the special case of $r = 2$, the ordinary difference sequence is $(d_0, d_1, d_2) = (6^{-1/2}, -(2/3)^{1/2}, 6^{-1/2})$ and the resulting estimator is $\hat{\sigma}_{\mathrm{ord}}^2(2) = \sum_{i=1}^{n-2}(Y_i - 2Y_{i+1} + Y_{i+2})^2/[6(n-2)]$, which was first proposed in Gasser, Sroka and Jennen-Steinmetz (1986). We refer to the estimator (4) with the ordinary difference sequence as $\hat{\sigma}_{\mathrm{ord}}^2(r)$. When $r = 1$, both $\hat{\sigma}_{\mathrm{opt}}^2(1)$ and $\hat{\sigma}_{\mathrm{ord}}^2(1)$ reduce

to the Rice estimator $\hat{\sigma}_R^2$. Hence, without loss of generality, we assume $r \geq 2$ in $\hat{\sigma}_{\text{opt}}^2(r)$ and $\hat{\sigma}_{\text{ord}}^2(r)$ unless otherwise specified.

By Dette, Munk and Wagner (1998), the asymptotic variance of $\hat{\sigma}_{\text{ord}}^2(r)$ is always larger or much larger than that of $\hat{\sigma}_{\text{opt}}^2(r)$. Specifically, under some mild conditions, we have

$$\frac{\operatorname{var}(\hat{\sigma}_{\text{ord}}^2(r))}{\operatorname{var}(\hat{\sigma}_{\text{opt}}^2(r))} \to \frac{2r}{2r+1}\binom{4r}{2r}\binom{2r}{r}^{-2} \quad \text{as } n \to \infty. \tag{7}$$

The ratio in the right-hand side of (7) is 1.556 when $r = 2$, 1.980 when $r = 3$, 2.335 when $r = 4$, and approximately as large as $\sqrt{\pi r/2}$ when r is large. For the asymptotic bias, by noting that $E(\hat{\sigma}_{\text{ord}}^2(r)) = \sigma^2 + O(n^{-2r})$ and $E(\hat{\sigma}_{\text{opt}}^2(r)) = \sigma^2 + O(n^{-2})$, $\hat{\sigma}_{\text{ord}}^2(r)$ always provides a smaller asymptotic bias than $\hat{\sigma}_{\text{opt}}^2(r)$ for any $r \geq 2$. In view of the bias-variance tradeoff, Dette, Munk and Wagner (1998) suggested to use the ordinary difference sequence if the sample size is small and the signal-to-noise ratio is large; otherwise, the optimal difference sequence should be used. Although very easy to implement, their rule of thumb can be confusing in practice since the signal-to-noise ratio is rarely known. In addition, it is never known in practice when the sample size is large enough so that the bias term can be negligible. We note that the choice of the difference sequence is still rather arbitrary in the recent literature. For instance, Hall and Heckman (2000) and Shen and Brown (2006) used the Rice estimator; Munk et al. (2005), Einmahl and Van Keilegom (2008), and Dette and Hetzler (2009) used the ordinary estimators; Brown and Levine (2007), Benko, Härdle and Kneip (2009), and Pendakur and Sperlich (2010) used the optimal estimators.

To conclude, the difference sequence selection problem remains a controversial issue in nonparametric regression up to now. The main goal of the paper is to provide a smart solution for the very challenging difference sequence selection problem. To achieve this, we propose a unified framework for estimating σ^2 that combines the linear regression method in Tong and Wang (2005) with the higher-order difference estimators systematically. By this combination, the unified framework integrates the existing literature on the difference-based estimation and it has, but not limited to, the following major contributions: (i) the unified framework generates a very large family of estimators that includes most existing estimators as special cases; (ii) all existing difference-based estimators are shown to be suboptimal in the proposed family of estimators; and (iii) in the unified framework, the ordinary difference sequence can be widely used no matter if the sample size is small and/or the signal-to-noise ratio is large.

The rest of the paper is organized as follows. In Section 2, we propose a unified framework for estimating σ^2 by introducing the general methodology, drawing the connections between the unified estimators and the existing estimators. In Section 3, we tackle the challenging "*optimal or ordinary difference sequence*" problem in the unified framework and propose a smart way to solve it. In Section 4, we first provide two data-driven methods to choose the tuning parameters, and then conduct simulation studies to evaluate the finite sample performance of the unified estimators and compare them with existing methods. In Section 5, we conclude the paper with some discussions. Finally, an online supplement (Dai, Tong and Zhu, 2017) is also provided in which we have supplied the technical details of the unified framework, including the theoretical results of the unified estimators, their technical proofs, a numerical comparison on the bias terms and an alternative procedure for variance estimation under unequally spaced design.

2. A UNIFIED FRAMEWORK FOR VARIANCE ESTIMATION

2.1 Methodology

The aforementioned difference-based estimators, including the optimal estimators and ordinary estimators, are popular in practice owing to their independence of curve fitting and the ease of implementation. Noting, however, that

$$\begin{aligned}\operatorname{MSE}(\hat{\sigma}_{\text{opt}}^2(r)) &= \min_{d_0,\ldots,d_r} \operatorname{MSE}(\hat{\sigma}^2(r)) \\ &= n^{-1}(\operatorname{var}(\varepsilon^2) + r^{-1}\sigma^4) + o(n^{-1}),\end{aligned}$$

none of fixed-order difference-based estimators can attain the asymptotically optimal rate of MSE in (2), a property possessed usually by the residual-based estimators only. To improve the literature, Tong and Wang (2005) have proposed a new direction for estimating the residual variance, inspired by the fact that the Rice estimator is always positively biased. Their linear regression method not only eliminated the estimation bias, but also reduced the estimation variance dramatically and hence achieved the asymptotically optimal

rate of MSE for variance estimation. See also, for example, Park, Kim and Lee (2012), Tong, Ma and Wang (2013) and Dai et al. (2015).

To make the linear regression method a more effective tool and also to tackle the "*optimal or ordinary difference sequence*" problem, we propose a unified framework for estimating σ^2 that combines the linear regression method with the higher-order difference estimators systematically. Specifically, for any order-r difference sequence $d = (d_0, \ldots, d_r)$ satisfying (3), we define

$$s_k(r) = \frac{1}{n-rk} \sum_{i=1}^{n-rk} \left(\sum_{j=0}^{r} d_j Y_{i+jk} \right)^2. \tag{8}$$

When $k = 1$, $s_k(r)$ reduces to the difference-based estimator $\hat{\sigma}^2(r)$ in (4). For any fixed r and k, we have the expectation of $s_k(r)$ as

$$E\big[s_k(r)\big] = \sigma^2 + \frac{1}{n-rk} \sum_{i=1}^{n-rk} \left(\sum_{j=0}^{r} d_j g(x_{i+jk}) \right)^2.$$

This shows that $s_k(r)$ is always positively biased for estimating σ^2. To better quantify the size of the bias, we assume that g has a bounded first derivative and let $J(r) = (\sum_{j=0}^{r} j d_j)^2 \int_0^1 [g'(x)]^2 \, dx$. Then for any fixed r and $k = o(n)$, we have

$$E\big[s_k(r)\big] = \sigma^2 + \frac{k^2}{n^2} J(r) + o\left(\frac{k^2}{n^2}\right). \tag{9}$$

To eliminate the bias term in (9), we consider a linear regression model to a collection of $s_k(r)$ and then estimate σ^2 as the intercept. Specifically, by letting $\alpha = \sigma^2$, $\beta = J(r)$ and $h_k = k^2/n^2$, we have the approximately linear regression model $s_k(r) \approx \alpha + h_k \beta$. Then for the given $s_k(r)$, $k = 1, \ldots, m$ with $m = o(n)$, we fit the linear regression model by minimizing the following weighted sum of squares

$$\sum_{k=1}^{m} w_k \big(s_k(r) - \alpha - h_k \beta\big)^2, \quad \beta > 0,$$

where $w_k = (n - rk)/N$ are the corresponding weights with $N = \sum_{k=1}^{m} (n - rk) = nm - rm(m+1)/2$. Finally, we estimate σ^2 by the fitted intercept in the unified framework. This leads to the unified estimator as

$$\hat{\sigma}^2(r, m) = \hat{\alpha} = \sum_{k=1}^{m} b_k w_k s_k(r), \tag{10}$$

where $b_k = 1 - \bar{h}_w (h_k - \bar{h}_w) / (\sum_{k=1}^{m} w_k h_k^2 - \bar{h}_w^2)$ and $\bar{h}_w = \sum_{k=1}^{m} w_k h_k$.

Note that the weights $\{w_k\}$ in (10) are assigned in such a way that each $s_k(r)$ involves $(n - rk)$ pairs of observations and the regression weighs equally for each pair. By this, we have not only provided a simplified form for the final estimator, but also improves the finite-sample performance. Whereas for the asymptotic behavior, it can be readily shown that the estimator (10) is asymptotically equivalent to the estimator that minimizes the unweighted sum of squares $\sum_{k=1}^{m} (s_k(r) - \alpha - h_k \beta)^2$.

If the optimal difference sequence is used in (8), we refer to the unified estimator (10) as the unified optimal estimator, denoted by $\hat{\sigma}^2_{\text{opt}}(r, m)$. Otherwise, if the ordinary difference sequence is used, we refer to it as the unified ordinary estimator, denoted by $\hat{\sigma}^2_{\text{ord}}(r, m)$. By Theorems S2 and S3 in the online supplement, the unified estimator $\hat{\sigma}^2(r, m)$ is capable to control the bias to order $O(m^3/n^3)$ for the optimal difference sequence, and to order $O(m^{2r}/n^{2r})$ for the ordinary difference sequence. This demonstrates that the linear regression with the ordinary difference sequence provides a smaller asymptotic bias than the linear regression with the optimal difference sequence for any $r \geq 2$. In addition, by Theorem S4 in the online supplement, the unified estimator (10) can always achieve the asymptotically optimal rate of MSE, and hence is a consistent estimator of σ^2, no matter which difference sequence is used. From this point of view, the unified estimator has improved the classical difference-based estimators in Hall, Kay and Titterington (1990).

2.2 Unified Estimators

By a combination of the linear regression and the higher-order difference sequence, we have proposed a unified framework for variance estimation in nonparametric regression. In particular, with the tuning parameters r and m, we have generated a two-dimensional cone space, that is, $\mathcal{S} = \{(r, m) : r = 1, 2, \ldots; m = 1, 2, \ldots\}$, for locating the optimal variance estimator. The unified framework (see Figure 1) includes all existing difference-based estimators as special cases, which are all located in the edge of the two-dimensional cone space. When $m = 1$ and $r = 1$, the proposed estimator $\hat{\sigma}^2(r, m)$ results in the Rice estimator in (5), which is located on the corner of the cone space. If we fix $m = 1$ and allow $r \geq 2$, $\hat{\sigma}^2(r, m)$ results in the classical difference-based estimators including Gasser, Sroka and Jennen-Steinmetz (1986), Hall, Kay and Titterington (1990) and Dette, Munk and Wagner (1998). On the other side, if we fix $r = 1$ and allow $m \geq 2$,

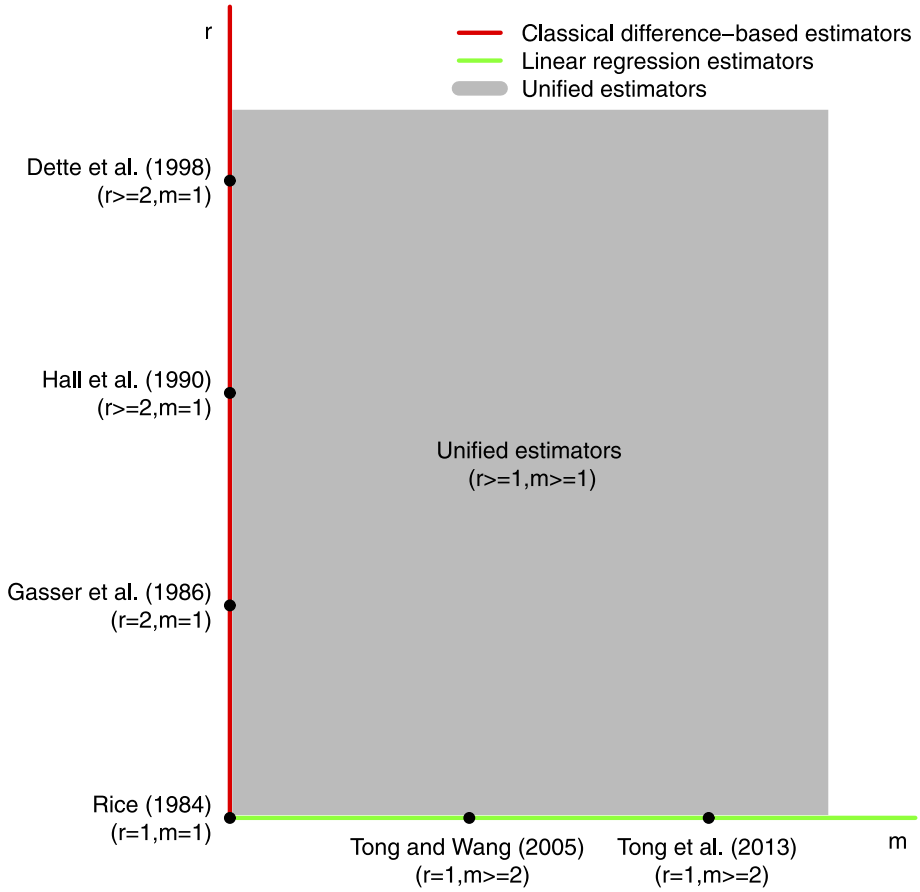

FIG. 1. *The unified framework for estimating σ^2 in nonparametric regression, where the existing difference-based estimators are all located in the edge of the two-dimensional cone space.*

then $\hat{\sigma}^2(r, m)$ results in the linear regression estimators in Tong and Wang (2005) and Tong, Ma and Wang (2013). From this point of view, the unified framework has greatly enriched the existing literature on the difference-based estimation in nonparametric regression.

Note that the difference-based estimator in (4) finds the optimal tuning parameters $(r_{\text{opt}}, m_{\text{opt}})$ only in the subspace $\mathcal{S}_1 = \{(r, 1) : r = 1, 2, \ldots\}$; whereas the linear regression estimator finds them only in the subspace $\mathcal{S}_2 = \{(1, m) : m = 1, 2, \ldots\}$. Neither of them may be globally optimal in the unified framework since $(r_{\text{opt}}, m_{\text{opt}})$ may also be located in the inner space $\mathcal{S} \setminus (\mathcal{S}_1 \cup \mathcal{S}_2)$. For the purpose of illustration, we present a numerical example to demonstrate our claim. Specifically, we consider $g(x) = 5\sin(2\pi x)$ with $n = 100$ and $\varepsilon \sim N(0, 4)$. With 1000 Monte Carlo simulations, we report in Figure 2 the simulated MSEs for the unified optimal estimator $\hat{\sigma}^2_{\text{opt}}(r, m)$. For the reported range with r from 1 to 5 and m from 1 to 10, the minimum MSE is 0.3615 which is located on $(r, m) = (2, 8)$. This numerically demonstrates that the optimal tuning parameters may not necessarily be in the edges of the two-dimensional cone space. Under the unified framework, we define the optimal variance estimator as

$$\tilde{\sigma}^2 = \hat{\sigma}^2(r_{\text{opt}}, m_{\text{opt}}),$$

where $(r_{\text{opt}}, m_{\text{opt}}) = \text{argmin}_{(r,m)\in\mathcal{S}} E(\hat{\sigma}^2(r, m) - \sigma^2)^2$ are the optimal parameters. Note that r_{opt} and m_{opt} are unknown in practice and need to be estimated. The methods for selecting the tuning parameters are given in Section 4.1.

Finally, we note that the unified framework can also be interpreted from the point of view of variance reduction. In estimating the mean function at a given point, Cheng, Peng and Wu (2007) and Paige, Sun and Wang (2009) formed a linear combination of the local linear estimators evaluated at several nearby points as the final estimate. The linear combination therein was constructed in such a way that maximizes the variance reduction while remaining the asymptotic bias unchanged. To our knowledge, there is little existing work

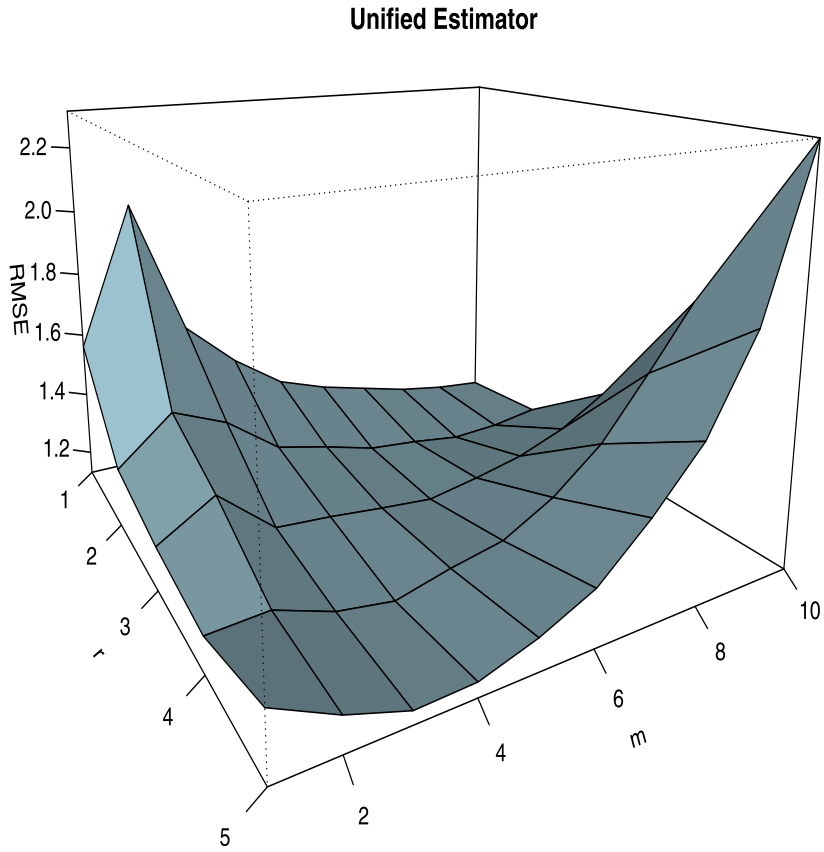

FIG. 2. *The perspective plot of the MSEs of the unified optimal estimator based on* 1000 *Monte Carlo simulations, where* $g(x) = 5\sin(2\pi x)$, $n = 100$, *and* $\varepsilon \sim N(0, 4)$. *The minimum MSE is* 0.3615 *which is located on* $(r, m) = (2, 8)$.

in the literature on variance reduction in nonparametric variance estimation. Note that $s_k(r)$ in (8) can be represented as a combination of several lag-k Rice estimators. Hence, in spirit to the simulation-extrapolation method of Cook and Stefanski (1994) and Stefanski and Cook (1995), and the empirical-bias bandwidth selection method of Ruppert (1997), the unified estimator $\hat{\sigma}^2(r, m)$ can be treated as a variance reduced estimator in comparison with the linear regression estimator in Tong and Wang (2005).

3. DIFFERENCE SEQUENCE SELECTION

Section 2 shows that the unified framework has generated a large family of new estimators and has greatly enriched the existing literature on variance estimation. In this section, we further show that the unified framework has provided a smart way to solve the challenging "*optimal or ordinary difference sequence*" problem.

For ease of exposition, we assume the random errors are normally distributed with mean zero and variance σ^2 in the remainder of the paper, then $\operatorname{var}(\varepsilon^2) = 2\sigma^4$. For the classical optimal estimator, we have

$$\operatorname{bias}\bigl(\hat{\sigma}^2_{\mathrm{opt}}(r, 1)\bigr) = O\bigl(n^{-2}\bigr) \quad \text{and}$$

$$\operatorname{var}\bigl(\hat{\sigma}^2_{\mathrm{opt}}(r, 1)\bigr) = \frac{V_1}{n}\operatorname{var}\bigl(\varepsilon^2\bigr) + o\bigl(n^{-1}\bigr),$$

where $V_1 = 1 + 1/2r$. For the classical ordinary estimator, we have

$$\operatorname{bias}\bigl(\hat{\sigma}^2_{\mathrm{ord}}(r, 1)\bigr) = O\bigl(n^{-2r}\bigr) \quad \text{and}$$

$$\operatorname{var}\bigl(\hat{\sigma}^2_{\mathrm{ord}}(r, 1)\bigr) = \frac{V_2}{n}\operatorname{var}\bigl(\varepsilon^2\bigr) + o\bigl(n^{-1}\bigr),$$

where $V_2 = \binom{4r}{2r}/\binom{2r}{r}^2$. Given that $V_1 > 1$ and $V_2 > 1$ for any $r \geq 1$, neither $\hat{\sigma}^2_{\mathrm{opt}}(r, 1)$ nor $\hat{\sigma}^2_{\mathrm{ord}}(r, 1)$ attains the asymptotically optimal rate of MSE in (2). Note also that $V_2 > V_1$ for any $r \geq 2$. Due to the bias-variance tradeoff, Dette, Munk and Wagner (1998) recommended to use the ordinary difference sequence when the sample size is small, and otherwise use the optimal difference sequence. In practice, however, it is rarely known whether the sample size is sufficiently small so that we can safely use the ordinary difference sequence or the other.

3.1 Bias-Variance Tradeoff

Let $m \to \infty$ and $m/n \to 0$. By Section S2 of the online supplement, the asymptotic bias and variance of the unified optimal estimator are

$$\operatorname{bias}\big(\hat{\sigma}^2_{\mathrm{opt}}(r,m)\big) = O\left(\frac{m^3}{n^3}\right) \quad \text{and}$$

$$\operatorname{var}\big(\hat{\sigma}^2_{\mathrm{opt}}(r,m)\big) = \frac{1}{n}\operatorname{var}\big(\varepsilon^2\big) + o\big(n^{-1}\big).$$

While for the unified ordinary estimator, we have

$$\operatorname{bias}\big(\hat{\sigma}^2_{\mathrm{ord}}(r,m)\big) = O\left(\frac{m^{2r}}{n^{2r}}\right) \quad \text{and}$$

$$\operatorname{var}\big(\hat{\sigma}^2_{\mathrm{ord}}(r,m)\big) = \frac{1}{n}\operatorname{var}\big(\varepsilon^2\big) + o\big(n^{-1}\big).$$

We note that, unlike those for the classical difference-based estimators, the asymptotic variances of $\hat{\sigma}^2_{\mathrm{opt}}(r,m)$ and $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ are the same and both attain the asymptotically optimal rate. Their difference appears only in the higher order terms, and hence is much alleviated compared to the difference $V_2 - V_1$ in the leading term for the classical difference-based estimators. Because of this, we have nearly gotten rid of the painful bias-variance tradeoff and can determine the appropriate difference sequence by the asymptotic bias of the estimators. As a conclusion, we recommend to use the ordinary difference sequence in the unified framework for any $r \geq 2$, no matter if the sample size is small or not.

To explore the advantage of the unified ordinary estimator over the existing estimators, we summarize in Table 1 the asymptotic biases and variances for the two unified estimators and the existing difference-based estimators. First, we conclude that $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ is better than $\hat{\sigma}^2(1,m)$ and $\hat{\sigma}^2_{\mathrm{opt}}(r,m)$, given that their asymptotic variances all attain the asymptotically optimal rate but $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ has the smallest asymptotic bias. Note also that a combination of optimal difference sequence and the linear regression method will not further reduce the asymptotic estimation bias. Second, if we choose $m = n^\tau$ with $0 < \tau < (2r-2)/2r$, then the asymptotic bias of $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ is of order $o(n^{-2})$, and hence it outperforms $\hat{\sigma}^2(1,1)$ and $\hat{\sigma}^2_{\mathrm{opt}}(r,1)$ in both mean and variance. Third, if we choose $m = n^\tau$ with $0 < \tau \to 0$, then the asymptotic bias of $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ converges to $O(n^{-2r})$ which is only beaten by the classical ordinary estimator $\hat{\sigma}^2_{\mathrm{ord}}(r,1)$. But on the other hand, the asymptotic variance of $\hat{\sigma}^2_{\mathrm{ord}}(r,m)$ is much smaller than that of $\hat{\sigma}^2_{\mathrm{ord}}(r,1)$, and hence the overall improvement is still quite significant. For more details, see the simulation results in Section 4.

3.2 Unified Ordinary Estimator

By Table 1, it is evident that a combination of the linear regression method and the ordinary difference sequence leads to a well behaved estimator for variance estimation in nonparametric regression, in which the linear regression method reduces the estimation variance and the ordinary difference sequence controls the estimation bias. In contrary, a combination of the linear regression method and the optimal difference sequence is less satisfactory, mainly because both techniques are to reduce the estimation variance so that the estimation bias may not be controlled sufficiently.

On the other hand, the TW estimator in Tong and Wang (2005) only uses the first-order estimators as regressors and, as a consequence, it is not possible to tackle the difference sequence selection problem. Note also that, without an effective mechanism in controlling the estimation bias, the TW estimator may also not provide a satisfactory performance when n is small and g is rough. For this, one may refer to the simulation results in Tong and Wang (2005) for small sample

TABLE 1
Asymptotic biases and variances for various estimators

		Asymptotic bias	Asymptotic variance
Rice (1984)	$\hat{\sigma}^2(1,1)$	$O(\frac{1}{n^2})$	$\frac{1.5}{n}\operatorname{var}(\varepsilon^2)$
Hall, Kay and Titterington (1990)	$\hat{\sigma}^2_{\mathrm{opt}}(r,1)$	$O(\frac{1}{n^2})$	$\frac{V_1}{n}\operatorname{var}(\varepsilon^2)$
Dette, Munk and Wagner (1998)	$\hat{\sigma}^2_{\mathrm{ord}}(r,1)$	$O(\frac{1}{n^{2r}})$	$\frac{V_2}{n}\operatorname{var}(\varepsilon^2)$
Tong and Wang (2005)	$\hat{\sigma}^2(1,m)$	$O(\frac{m^3}{n^3})$	$\frac{1}{n}\operatorname{var}(\varepsilon^2)$
Unified optimal estimator	$\hat{\sigma}^2_{\mathrm{opt}}(r,m)$	$O(\frac{m^3}{n^3})$	$\frac{1}{n}\operatorname{var}(\varepsilon^2)$
Unified ordinary estimator	$\hat{\sigma}^2_{\mathrm{ord}}(r,m)$	$O(\frac{m^{2r}}{n^{2r}})$	$\frac{1}{n}\operatorname{var}(\varepsilon^2)$

sizes. In addition, we have provided a numerical comparison study for the bias terms of the TW estimator and the proposed unified estimators in Section S3 of the online supplement. In summary, we recommend the unified ordinary estimator for practical use, no matter if the sample size is small or if the signal-to-noise ratio is large.

4. SIMULATION STUDIES

We first present two data-driven methods for selecting the tuning parameters r and m in the unified framework. We then assess the performance of the unified optimal and ordinary estimators, and make a recommendation between the two estimators for practical implementation. Finally, we compare the unified ordinary estimator with some existing competitors and demonstrate its superiority.

4.1 Choice of the Tuning Parameters

Apart from the difference sequence selection problem, the choices of the order r and the number of regression points m are also important in practice. For normally distributed errors, the optimal bandwidth is $m_{\text{opt}} = \sqrt{14}n^{1/2}$ when $r = 1$. And by Theorem S4 in the online supplement, the optimal bandwidth is $m_{\text{opt}} = \sqrt{A_1/(2A_2)}n^{1/2}$ when $r = 2$, where $A_1 = 9/4 + 9d_1^2(d_1^2 - 1/2)$ and $A_2 = 9/56 + 165d_1^2(1 - d_1^2)/448$. The optimal bandwidth is of order $O(n^{1/2})$ for any fixed value of $r \geq 1$. However, when n is small or moderate, as reported in Tong and Wang (2005), the theoretical bandwidth may be too large and is not applicable in practice.

For choosing the two tuning parameters in the unified estimation, we consider (i) the cross-validation (CV) method, and (ii) the plateau method. For the CV method, we first divide the whole data set into V disjoint sub-samples, $S_1, \ldots, S_V$. We then select the optimal pair of (r, m) that minimizes

$$\text{CV}(r, m) = \sum_{v=1}^{V} \big[\hat{\sigma}^2(r, m) - \hat{\sigma}_v^2(r, m)\big]^2,$$

where $\hat{\sigma}_v^2(r, m)$ denotes the unified estimate of σ^2 on the whole sample except for S_v with the tuning parameters r and m. This is also referred to as the V-fold CV method. We note that, however, the CV method is generally computationally expensive for choosing r and m simultaneously, especially when n and V are both large.

For large n, we propose another more effective method for choosing the tuning parameters. In essence, we follow the plateau method in Müller and Stadtmüller (1999) and propose the following criterion:

$$(\hat{r}, \hat{m}) = \arg\min_{r,m} \left\{ \frac{1}{2m_r + 1} \sum_{i=[m/r]-m_r}^{[m/r]+m_r} \big[\hat{\sigma}^2(r, i)\big]^2 - \left[\frac{1}{2m_r + 1} \sum_{i=[m/r]-m_r}^{[m/r]+m_r} \hat{\sigma}^2(r, i) \right]^2 \right\}, \tag{11}$$

where $m_0 = \max([n/50], 2)$ and $m_r = \max([m_0/r], 1)$. The expression in the curly brackets can be regarded as an approximation of the local variation of the estimator. For illustration, we present a numerical example to display the behavior of the unified ordinary estimator using the plateau method. Let $n = 500$, $\varepsilon \sim N(0, 0.25)$ and $g(x) = 5\sin(4\pi x)$. With 100 simulations, we report in Figure 3 the trend of the averaged $\hat{\sigma}^2_{\text{ord}}(r, m)$ along with the bandwidth m for $r = 1$, 2, 3 and 4, respectively. The true variance at $\sigma^2 = 0.25$ is also reported using the dashed lines for comparison. From Figure 3, we observe that the averaged $\hat{\sigma}^2_{\text{ord}}(r, m)$ stays around the true value of the residual variance within some range of bandwidth m, and then moves away monotonically.

4.2 Comparison Between $\hat{\sigma}^2_{\text{ord}}(r, m)$ and $\hat{\sigma}^2_{\text{opt}}(r, m)$

Our first simulation study is to conduct a comprehensive comparison for the finite sample performance of the unified ordinary and optimal estimators. We consider four mean functions with their shapes displayed in Figure 4:

$$g_1(x) = 5\sin(\pi x),$$
$$g_2(x) = 5\sin(4\pi x),$$
$$g_3(x) = 10\big[x + (2\pi)^{-1/2}\exp\{-100(x - 0.5)^2\}\big],$$
$$g_4(x) = 3\beta_{10,3}(x) + 2\beta_{3,11}(x),$$

where g_1 and g_2 are popularly used in the difference-based literature (Dette, Munk and Wagner, 1998, Tong and Wang, 2005), g_3 is a bell-shaped function used in Härdle (1990), and g_4 is a bimodal function used in Wahba (1983) with $\beta_{p,q}(x) = [\Gamma(p + q)/\Gamma(p)\Gamma(q)]x^{p-1}(1 - x)^{q-1}$. We consider the equidistant design with $x_i = i/n$, and generate ε_i independently from $N(0, \sigma^2)$. We further consider $n = 25$, 50 and 200, corresponding with small, moderate and large sample sizes, and $\sigma = 0.2$, 0.5 and 2, corresponding to low, moderate and high variances, respectively.

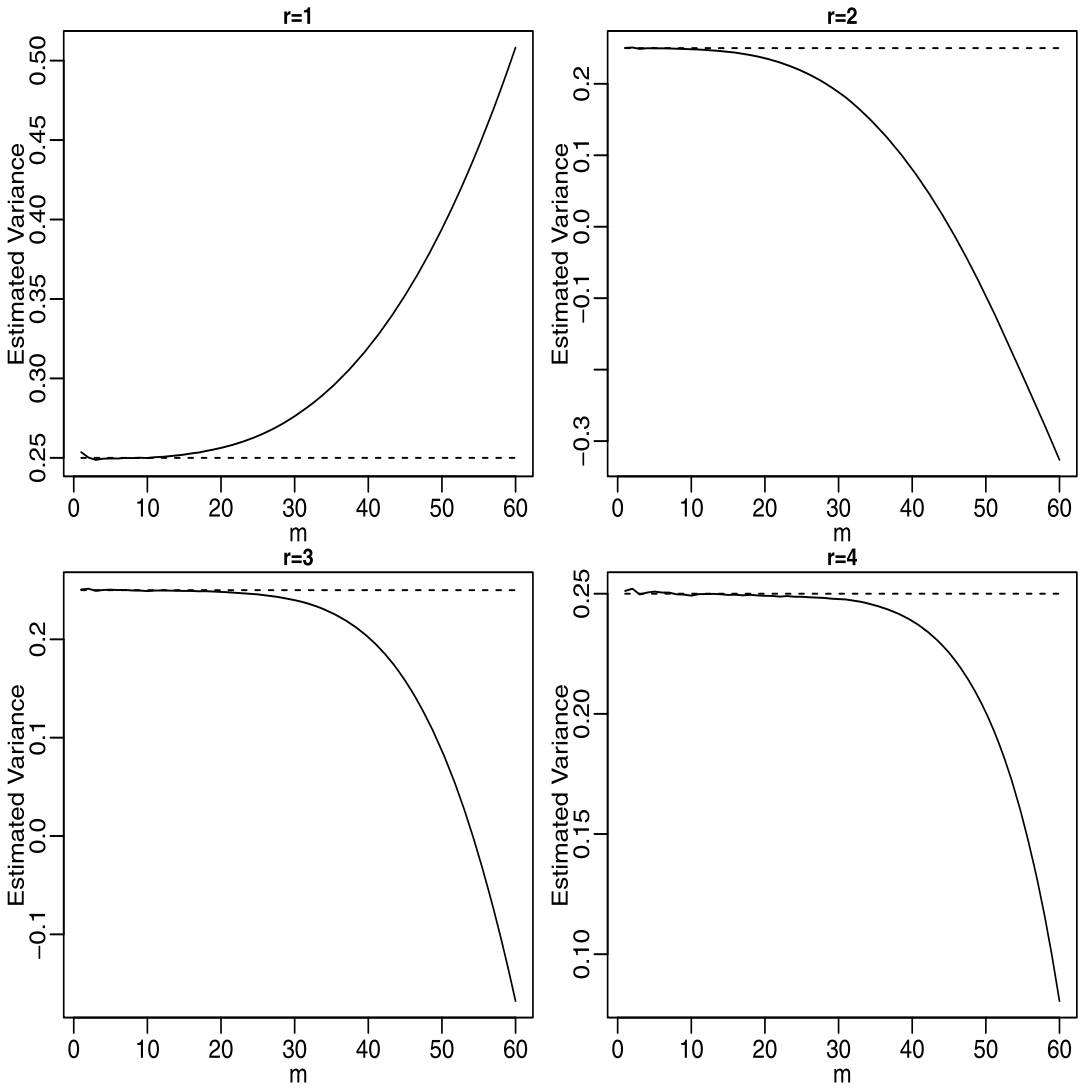

FIG. 3. *The trend of* $\hat{\sigma}^2_{\text{ord}}(r,m)$ *along with the bandwidth* m *for various* r *values, where* $n = 500$, $\varepsilon \sim N(0, 0.25)$ *and* $g(x) = 5\sin(4\pi x)$. *The solid lines represent the average values of* 100 *simulated* $\hat{\sigma}^2_{\text{ord}}(r,m)$, *and the dashed lines represent the true variance value.*

For the order of difference sequence, we focus on $r \leq 3$ for the unified estimators since an order of $r \geq 4$ is rarely recommended unless the mean function is enormously oscillating (Dette, Munk and Wagner, 1998). For the bandwidth m, we let $m'_s = \max([m_s/r], 1)$ for $\hat{\sigma}^2(r,m)$, where $m_s = n^{1/2}$ is the bandwidth suggested for $\hat{\sigma}^2(1,m)$ in Tong and Wang (2005). We refer to the resulting estimators with the fixed bandwidths as $\hat{\sigma}^2_{\text{ord}}(r,m'_s)$ and $\hat{\sigma}^2_{\text{opt}}(r,m'_s)$, respectively. Then with $r = 2$ and $r = 3$, we consider the following four estimators: $\hat{\sigma}^2_{\text{ord}}(2,m'_s)$, $\hat{\sigma}^2_{\text{opt}}(2,m'_s)$, $\hat{\sigma}^2_{\text{ord}}(3,m'_s)$, and $\hat{\sigma}^2_{\text{opt}}(3,m'_s)$. In addition, we also consider the unified estimators with the tuning parameters being selected by the data-driven methods. Specifically, we apply the leave-one-out CV method for $n = 25$ and 50, and the plateau method for $n = 200$. The tuning parameters are chosen from the space $\{(r,m) : rm \leq B \text{ with } r = 1,2,3\}$, where B is a pre-specified positive number. In our simulation, we set the value of B as follows: $B = 5$ for $n = 25$, $B = 8$ for $n = 50$, and $B = 15$ for $n = 200$. We refer to the unified estimators with the CV-based bandwidths as $\hat{\sigma}^2_{\text{ord}}(r_d,m_d)$ and $\hat{\sigma}^2_{\text{opt}}(r_d,m_d)$, where r_d and m_d are the selected tuning parameters.

In Table 2, we report the relative MSE, $(n/2\sigma^2)$ MSE, of the six estimators based on 1000 simulations for each setting. For the four estimators with fixed bandwidths, it is evident that the two unified ordinary estimators are more robust than the two uni-

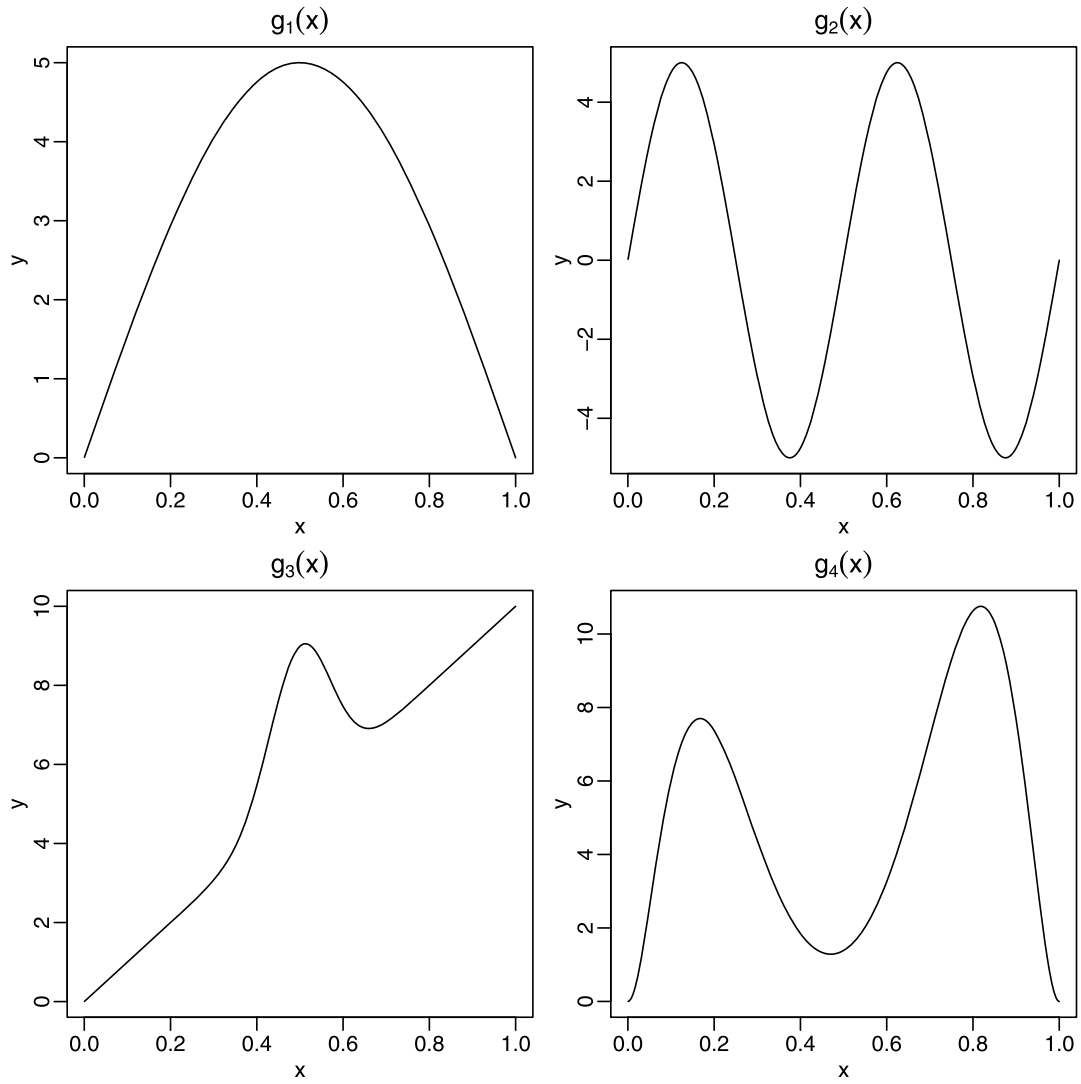

FIG. 4. *The four mean functions and their respective shapes.*

fied optimal estimators. Specifically, they provide a better control on the estimation bias, and consequently, RMSE$[\hat{\sigma}^2_{\text{ord}}(r, m'_s)]$ can be much smaller than RMSE$[\hat{\sigma}^2_{\text{opt}}(r, m'_s)]$ when the sample size is small or the signal-to-noise ratio is large. We note also that for the unified ordinary estimator, $\hat{\sigma}^2_{\text{ord}}(3, m'_s)$ is even better than $\hat{\sigma}^2_{\text{ord}}(2, m'_s)$ when the sample size is small. This coincides with the comparison results for the classical difference-based estimators. On the other side, the estimators with CV-based bandwidths (where r is not fixed at 2 or 3, and hence more flexible) always provide a comparable or even better performance than the estimators with fixed bandwidths. In particular, $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ outperforms all other five estimators including $\hat{\sigma}^2_{\text{opt}}(r_d, m_d)$ in most settings. We hence recommend $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ for practical use in the unified framework.

4.3 Comparison with Other Estimators

Our second simulation study is to compare the recommended unified ordinary estimator, $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$, with six existing competitors in the literature. Specifically, we consider $\hat{\sigma}^2_{\text{ord}}(2, 1)$ in Gasser, Sroka and Jennen-Steinmetz (1986), $\hat{\sigma}^2_{\text{ord}}(3, 1)$ in Dette, Munk and Wagner (1998), $\hat{\sigma}^2_{\text{opt}}(r, 1)$ in Hall, Kay and Titterington (1990) with $r = 2$ and 3, and $\hat{\sigma}^2_{\text{TW}}(m_{\text{CV}})$ in Tong and Wang (2005) with m_{CV} being selected by the CV method from the space $\{m : 1 \le m \le B\}$. For the sake of consistency, we remain the same simulation settings as those in Section 4.2. Note also that, to save space, we do not include the Rice estimator as it is always

TABLE 2
The relative mean squared errors (RMSEs) for three unified optimal estimators and three unified ordinary estimators, based on 1000 *simulations*

			Unified optimal estimators			Unified ordinary estimators		
n	σ	g	$(2, m_s')$	$(3, m_s')$	(r_d, m_d)	$(2, m_s')$	$(3, m_s')$	(r_d, m_d)
25	0.2	g_1	12.6	1239	12.0	3.01	2.35	2.11
		g_2	7144	281,974	9427	1693	2.89	2.89
		g_3	105	8260	150	38.6	2.54	2.54
		g_4	22,475	170,785	14,650	503	4.68	4.68
	0.5	g_1	1.90	33.6	1.86	2.94	2.35	2.03
		g_2	188	7236	245	47.2	2.37	2.75
		g_3	4.57	217	6.15	4.05	2.35	2.36
		g_4	582	4399	371	16.4	2.38	2.37
	2	g_1	1.44	1.45	1.35	2.93	2.35	1.54
		g_2	2.49	31.0	2.54	3.17	2.36	2.56
		g_3	1.49	2.34	1.37	2.96	2.35	1.56
		g_4	4.15	19.8	2.52	3.00	2.34	2.65
50	0.2	g_1	3.59	3.88	3.51	1.85	4.00	2.06
		g_2	1321	1175	2319	235	4.08	3.42
		g_3	15.0	14.8	58.7	7.58	4.02	2.83
		g_4	4328	3048	4974	76.9	4.28	3.21
	0.5	g_1	1.50	1.45	1.30	1.84	4.00	1.52
		g_2	36.2	32.3	62.0	7.72	4.00	4.00
		g_3	1.82	1.76	2.53	1.98	4.00	2.60
		g_4	115	81.3	123	3.73	4.03	2.55
	2	g_1	1.40	1.32	1.23	1.84	4.00	1.37
		g_2	1.61	1.52	1.53	1.87	4.00	1.72
		g_3	1.41	1.33	1.22	1.85	4.00	1.39
		g_4	1.98	1.74	1.72	1.85	4.01	1.97
200	0.2	g_1	1.40	1.36	1.45	1.39	1.91	1.35
		g_2	76.0	27.4	3.13	7.71	1.91	1.61
		g_3	1.48	1.34	1.35	1.59	1.91	1.51
		g_4	156	37.3	1.52	3.69	1.91	1.71
	0.5	g_1	1.26	1.30	1.33	1.39	1.91	1.34
		g_2	3.22	2.03	1.59	1.55	1.91	1.44
		g_3	1.25	1.30	1.32	1.39	1.91	1.42
		g_4	5.32	2.27	1.46	1.46	1.91	1.40
	2	g_1	1.26	1.30	1.29	1.39	1.91	1.33
		g_2	1.38	1.36	1.41	1.40	1.91	1.38
		g_3	1.25	1.30	1.31	1.39	1.91	1.34
		g_4	1.53	1.37	1.41	1.40	1.91	1.35

less satisfactory. In addition to the difference-based estimators, we also consider a residual-based estimator for comparison. Specifically, we fit the mean function using the cubic smoothing spline and compute the residual-based estimator, $\hat{\sigma}^2_{\text{ss}}$, as the average squared residuals. The tuning parameter is selected via the generalized cross validation. Finally, with 1000 simulations, we report the RMSEs of the seven estimators in Table 3 for each setting.

We first compare $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ with the five existing difference-base methods. When the sample size is 200, $\hat{\sigma}^2_{\text{opt}}(2, 1)$ and $\hat{\sigma}^2_{\text{opt}}(3, 1)$ are very sensitive to the different values of signal-to-noise ratio. For the remaining four estimators, we have $\text{MSE}(\hat{\sigma}^2_{\text{ord}}(r_d, m_d)) \simeq \text{MSE}(\hat{\sigma}^2_{\text{TW}}(m_{\text{CV}})) < \text{MSE}(\hat{\sigma}^2_{\text{ord}}(2, 1)) < \text{MSE}(\hat{\sigma}^2_{\text{ord}}(3, 1))$ for most cases. This coincides with the theoretical results that $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ and $\hat{\sigma}^2_{\text{TW}}(m_{\text{CV}})$ attain the asymptotically optimal rate of MSE, a property not possessed by the classical difference-based estimators. When the sample size is 50, the comparative results remain similar except that the two optimal estimators are getting even worse owing to their poor ability in con-

TABLE 3
The relative mean squared errors (RMSEs) for the unified ordinary estimator and six existing methods, based on 1000 *simulations*

n	σ	g	$\hat{\sigma}^2_{\text{ord}}(2,1)$	$\hat{\sigma}^2_{\text{opt}}(2,1)$	$\hat{\sigma}^2_{\text{ord}}(3,1)$	$\hat{\sigma}^2_{\text{opt}}(3,1)$	$\hat{\sigma}^2_{\text{TW}}(m_{\text{CV}})$	$\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$	$\hat{\sigma}^2_{\text{ss}}$
25	0.2	g_1	1.92	395	2.35	1239	34.2	2.11	2.92
		g_2	145	90,512	2.89	281,974	260	2.89	6.94
		g_3	6.47	2644	2.54	8260	3.64	2.54	5.75
		g_4	85.0	70,385	4.68	170,786	386	4.68	9.75
	0.5	g_1	1.93	11.6	2.35	33.6	2.15	2.03	2.89
		g_2	5.44	2323	2.37	7236	170	2.75	3.68
		g_3	5.03	70.4	2.35	217	9.08	2.36	3.47
		g_4	3.99	1810	2.38	4399	16.1	2.37	4.26
	2	g_1	1.93	1.37	2.35	1.45	1.25	1.54	2.83
		g_2	1.93	11.0	2.36	31.0	2.68	2.56	3.09
		g_3	1.92	1.68	2.35	2.34	1.30	1.56	3.14
		g_4	1.91	8.88	2.34	19.8	2.58	2.65	3.04
50	0.2	g_1	1.97	56.2	2.39	182	2.48	2.06	1.96
		g_2	3.18	13,539	2.40	44,210	47.6	3.42	2.38
		g_3	2.02	356	2.39	1208	1.91	2.83	2.39
		g_4	3.29	11,613	2.42	35,026	57.9	3.21	3.73
	0.5	g_1	1.97	2.74	2.39	5.97	1.57	1.52	1.91
		g_2	2.01	349	2.39	1135	3.46	4.00	2.21
		g_3	1.97	10.5	2.39	32.4	1.71	2.60	2.21
		g_4	2.03	300	2.40	901	3.16	2.55	2.44
	2	g_1	1.97	1.28	2.39	1.21	1.15	1.37	1.87
		g_2	1.97	2.71	2.39	5.79	1.70	1.72	2.06
		g_3	1.97	1.32	2.39	1.35	1.22	1.39	2.08
		g_4	1.97	2.52	2.39	4.90	2.08	1.97	2.06
200	0.2	g_1	2.03	2.31	2.37	4.47	1.37	1.35	1.19
		g_2	2.03	234	2.37	802	2.31	1.61	1.26
		g_3	2.03	6.98	2.37	20.5	1.33	1.51	1.28
		g_4	2.03	192	2.37	664	1.52	1.71	1.40
	0.5	g_1	2.03	1.41	2.37	1.40	1.27	1.34	1.18
		g_2	2.03	7.42	2.37	21.9	1.43	1.44	1.24
		g_3	2.03	1.53	2.37	1.81	1.28	1.42	1.25
		g_4	2.03	6.31	2.37	18.3	1.36	1.40	1.31
	2	g_1	2.03	1.38	2.37	1.32	1.24	1.33	1.18
		g_2	2.03	1.77	2.37	2.61	1.36	1.38	1.24
		g_3	2.03	1.39	2.37	1.34	1.24	1.34	1.23
		g_4	2.03	1.70	2.37	2.39	1.33	1.35	1.25

trolling the estimation bias. In addition, by comparing Tables 2 and 3, we note that the classical difference-based estimators are significantly improved by their respective unified estimators in most cases for moderate to large sample sizes. When the sample size is small at $n = 25$, however, all estimators are getting more sensitive to the change of the signal-to-noise ratio. In particular, the classical optimal estimators and the TW estimator fail to provide reasonable estimates, especially when the signal-to-noise ratio is large. Also as observed in Dette, Munk and Wagner (1998) and Tong and Wang (2005), the classical ordinary estimators, $\hat{\sigma}^2_{\text{ord}}(2, 1)$ and $\hat{\sigma}^2_{\text{ord}}(3, 1)$, still provide to be reliable estimates for σ^2. More interestingly, we note that the unified ordinary estimator $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ again provides to be the best estimator in most cases. Even in the most severe case with $\sigma = 0.2$ and $g = g_2$, $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ still performs as well as the $\hat{\sigma}^2_{\text{ord}}(3, 1)$. Finally, for the comparison between $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ and the residual-based estimator $\hat{\sigma}^2_{\text{ss}}$, we note that $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ is better when $n = 25$, $\hat{\sigma}^2_{\text{ss}}$ is better when $n = 200$, and they provide a comparable performance when $n = 50$.

It is noteworthy that the above findings coincide with the theoretical results and comparisons in Sections 2

and 3. In summary, we recommend to use the unified ordinary estimator $\hat{\sigma}^2_{\text{ord}}(r_d, m_d)$ in practice, no matter if the sample size is small or if the signal-to-noise ratio is large.

5. CONCLUSION

In this paper, we proposed a unified framework for variance estimation in nonparametric regression that combines the higher order difference sequence with the linear regression method. The unified framework has greatly enriched the existing literature on variance estimation with most existing estimators as special cases. In the unified framework, we derived the asymptotic results for the unified optimal and ordinary estimators, and also made a comprehensive comparison between the two estimators through both theoretical and numerical results. As a conclusion, we recommended to use the ordinary difference sequence in the unified framework for any difference order being at least 2, no matter if the sample size is small or if the signal-to-noise ratio is large. From this point of view, the unified framework has completely solved the challenging difference sequence selection problem that remains a long-standing controversial issue in nonparametric regression for several decades.

We note that the difference-based methods have been extended to estimate the residual variance in other regression models, including the nonparametric models with dependent errors (Hall and Keilegom, 2003, Bliznyuk et al., 2012), the heteroscedastic regression models (Brown and Levine, 2007, Zhou et al., 2015), the regression models with multivariate covariates (Munk et al., 2005), and the semiparametric regression models (Tabakan and Akdeniz, 2010, Wang, Brown and Cai, 2011). Further research is needed to address the difference sequence selection problem in such models. Recently, researchers have also applied the difference-based methods to estimate the derivatives of the mean function in nonparametric regression (Charnigo, Hall and Srinivasan, 2011, De Brabanter et al., 2013, Wang and Lin, 2015, Dai, Tong and Genton, 2016). To cater for the demands of the application, we have developed a unified R package that integrates the existing difference-based estimators and the unified estimators in nonparametric regression. And for the fast dissemination of research results, we have made the R package, named VarED, freely available in the R statistical program http://cran.r-project.org/web/packages/.

ACKNOWLEDGMENT

Tiejun Tong's research was supported by the National Natural Science Foundation of China grant (No. 11671338), and the Hong Kong Baptist University grants FRG1/14-15/044, FRG2/15-16/019 and FRG2/15-16/038. Lixing Zhu's research was supported by the Hong Kong Research Grants Council grant (No. HKBU202810). The authors thank the editor, the associate editor and two reviewers for their constructive comments that have led to a substantial improvement of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "On the Choice of Difference Sequence in a Unified Framework for Variance Estimation in Nonparametric Regression." (DOI: 10.1214/17-STS613SUPP; .pdf).

REFERENCES

BENKO, M., HÄRDLE, W. and KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. MR2488343

BERKEY, C. S. (1982). Bayesian approach for a nonlinear growth model. *Biometrics* **38** 953–961.

BLIZNYUK, N., CARROLL, R. J., GENTON, M. G. and WANG, Y. (2012). Variogram estimation in the presence of trend. *Stat. Interface* **5** 159–168. MR2928067

BROWN, L. D. and LEVINE, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Ann. Statist.* **35** 2219–2232. MR2363969

CHARNIGO, R., HALL, B. and SRINIVASAN, C. (2011). A generalized C_p criterion for derivative estimation. *Technometrics* **53** 238–253. MR2857702

CHENG, M.-Y., PENG, L. and WU, J.-S. (2007). Reducing variance in univariate smoothing. *Ann. Statist.* **35** 522–542. MR2336858

COOK, J. R. and STEFANSKI, L. A. (1995). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89** 1314–1328.

DAI, W., TONG, T. and GENTON, M. G. (2016). Optimal estimation of derivatives in nonparametric regression. *J. Mach. Learn. Res.* **17(164)** 1–25.

DAI, W., TONG, T. and ZHU, L. (2017). Supplement to "On the Choice of Difference Sequence in a Unified Framework for Variance Estimation in Nonparametric Regression." DOI:10.1214/17-STS613SUPP.

DAI, W., MA, Y., TONG, T. and ZHU, L. (2015). Difference-based variance estimation in nonparametric regression with repeated measurement data. *J. Statist. Plann. Inference* **163** 1–20.

DETTE, H. and HETZLER, B. (2009). A simple test for the parametric form of the variance function in nonparametric regression. *Ann. Inst. Statist. Math.* **61** 861–886. MR2556768

DETTE, H., MUNK, A. and WAGNER, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. Roy Statist. Soc. Ser. B.* **60** 751–764. MR1649480

DE BRABANTER, K., DE BRABANTER, J., DE MOOR, B. and GIJBELS, I. (2013). Derivative estimation with local polynomial fitting. *J. Mach. Learn. Res.* **14** 281–301.

EINMAHL, J. H. J. and VAN KEILEGOM, I. (2008). Tests for independence in nonparametric regression. *Statist. Sinica* **18** 601–615. MR2411617

EUBANK, R. L. and SPIEGELMAN, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Statist. Assoc.* **85** 387–392. MR1141739

GASSER, T., KNEIP, A. and KÖHLER, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86** 643–652. MR1147088

GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633. MR0897854

HALL, P. and HECKMAN, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.* **28** 20–39. MR1762902

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. MR1087842

HALL, P. and KEILEGOM, I. V. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *J. Roy. Statist. Soc. Ser. B* **65** 443–456. MR1983757

HALL, P. and MARRON, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77** 415–419. MR1064818

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press, Cambridge.

HÄRDLE, W. and TSYBAKOV, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics* **81** 223–242.

MÜLLER, H.-G. and STADTMÜLLER, U. (1999). Discontinuous versus smooth regression. *Ann. Statist.* **27** 299–337. MR1701113

MUNK, A., BISSANTZ, N., WAGNER, T. and FREITAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Statist. Soc. Ser. B* **67** 19–41. MR2136637

PAIGE, R. L., SUN, S. and WANG, K. (2009). Variance reduction in smoothing splines. *Scand. J. Stat.* **36** 112–126. MR2508334

PARK, C., KIM, I. and LEE, Y. (2012). Error variance estimation via least squares for small sample nonparametric regression. *J. Statist. Plann. Inference* **142** 2369–2385. MR2911851

PENDAKUR, K. and SPERLICH, S. (2010). Semiparametric estimation of consumer demand systems in real expenditure. *J. Appl. Econometrics* **25** 420–457. MR2752121

RICE, J. A. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230. MR0760684

RUPPERT, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **92** 1049–1062. MR1482136

SHEN, H. and BROWN, L. D. (2006). Non-parametric modelling of time-varying customer service times at a bank call centre. *Appl. Stoch. Models Bus. Ind.* **22** 297–311.

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

STEFANSKI, L. A. and COOK, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *J. Amer. Statist. Assoc.* **90** 1247–1256. MR1379467

TABAKAN, G. and AKDENIZ, F. (2010). Difference-based ridge estimator of parameters in partial linear model. *Statist. Papers* **51** 357–368. MR2665357

TONG, T., MA, Y. and WANG, Y. (2013). Optimal variance estimation without estimating the mean function. *Bernoulli* **19** 1839–1854. MR3129036

TONG, T. and WANG, Y. (2005). Estimating residual variance in nonparametric regression using least squares. *Biometrika* **92** 821–830. MR2234188

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150. MR0701084

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

WANG, L., BROWN, L. D. and CAI, T. (2011). A difference based approach to the semiparametric partial linear model. *Electron. J. Stat.* **5** 619–641. MR2813557

WANG, W. W. and LIN, L. (2015). Derivative estimation based on difference sequence via locally weighted least squares regression. *J. Mach. Learn. Res.* **16** 2617–2641. MR3450519

YE, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131. MR1614596

ZHOU, Y., CHENG, Y., WANG, L. and TONG, T. (2015). Optimal difference-based variance estimation in heteroscedastic nonparametric regression. *Statist. Sinica* **25** 1377–1397. MR3409072