CrossMark

Optimal variance estimation based on lagged second-order difference in nonparametric regression

WenWu Wang[1] · **Lu Lin[1]** · **Li Yu[2]**

Abstract Differenced estimators of variance bypass the estimation of regression function and thus are simple to calculate. However, there exist two problems: most differenced estimators do not achieve the asymptotic optimal rate for the mean square error; for finite samples the estimation bias is also important and not further considered. In this paper, we estimate the variance as the intercept in a linear regression with the lagged Gasser-type variance estimator as dependent variable. For the equidistant design, our estimator is not only $n^{1/2}$-consistent and asymptotically normal, but also achieves the optimal bound in terms of estimation variance with less asymptotic bias. Simulation studies show that our estimator has less mean square error than some existing differenced estimators, especially in the cases of immense oscillation of regression function and small-sized sample.

1 Introduction

Consider the nonparametric regression model

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

✉ WenWu Wang
   wengewsh@sina.com

[1] Zhongtai Institute for Financial Studies and School of Mathematics, Shandong University, Jinan, China

[2] China Re Asset Management Company LTD, Beijing, China

where $0 \leq x_1 < \cdots < x_n \leq 1$ are design points, Y_i's are observed responses, $m(\cdot)$ is an unknown smooth mean function, and ϵ_i's are independent and identically distributed random errors with zero mean and variance σ^2.

For the estimation of variance σ^2, usually one fits the regression function m first by smoothing spline (Carter and Eagleson 1992; Carter et al. 1992) or kernel regression (Müller and Stadtmüller 1987; Hall and Carroll 1989; Neumann 1994), and then estimate variance σ^2 from residual sum of squares. However, the regression function estimation depends on the amount of smoothing (Dette et al. 1998), which requires knowledge about some unknown quantities such as $\int_0^1 \{m^{(1)}(x)\}^2 dx$ (Hall and Marron 1990), $\int_0^1 \{m^{(2)}(x)\}^2 dx$ (Buckley et al. 1988), and even $\int_0^1 \{m^{(l)}(x)\}^2 dx$ where l is a derivative order (Seifert et al. 1993; Eubank 1999; Wang 2011). Moreover, few methods are proposed to estimate the amount of smoothness $\int_0^1 \{m^{(l)}(x)\}^2 dx$. Thus in practical applications an estimator of σ^2 without estimating the regression function is preferable.

A class of estimators which bypass the estimation of regression function are the differenced estimators. Rice (1984) proposed the first-order differenced estimator

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

Later, Gasser et al. (1986) proposed the second-order differenced estimator

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

and Hall et al. (1990) generalized to the kth-order differenced estimator by minimizing the estimation variance. Based on former estimators, more estimators of σ^2 are proposed. With respect to the Rice-type estimator, Müller and Stadtmüller (1999) proposed the lagged Rice estimator; Müller et al. (2003) proposed a covariate-matched U-statistic; Tong and Wang (2005) further improved the lagged Rice estimator via estimating the variance as the intercept in a linear regression model, and the asymptotic optimal rate is discussed in Tong et al. (2013), Dai and Tong (2014) proposed a pairwise regression for models with jump discontinuities. As for the Gasser-type estimator, Seifert et al. (1993) generalized to a higher-order version, and Du and Schick (2009) proposed a covariate-matched U-statistic.

For the differenced estimators above, there exist the following problems. Most estimators keep the order k to be constant and don't achieve the asymptotic optimal rate for the mean square error (MSE), that is $\text{MSE}(\hat{\sigma}^2) = n^{-1}var(\varepsilon^2) + o(n^{-1})$, except the estimators in Müller et al. (2003), Tong and Wang (2005), and Du and Schick (2009). Zhou et al. (2015) indicated that Müller et al. (2003) and Du and Schick (2009) are not applicable in practice since they require an appropriate choice of bandwidth in advance. In Tong and Wang (2005), the bias caused by the regression function is non-ignorable for finite samples, especially in the cases of immense oscillation of regression function m and small sample size n. However, Seifert et al. (1993), and Du and Schick (2009) both showed their estimators based on the Gasser-

type estimator have better bias properties than the Rice-type and Hall-type estimators. Thus in practical applications, the method based on the Gasser-type estimator is preferred.

In this paper, we propose a new variance estimator based on the Gasser-type estimator. With equidistant designs, our estimator achieves the asymptotic optimal rate for the MSE. In finite samples, our estimator effectively reduces the estimation bias caused by the regression function, especially in the cases of immense oscillation of regression function and small-sized sample.

The remainder of this paper is organised as follows. In Sect. 2, we propose the lagged Gasser-type estimator, introduce the estimation methodology, and deduce the asymptotic results for our estimator. To assess the performance of our estimator, we conduct simulation studies and compare it with other differenced estimators in Sect. 3. All proofs of Theorems are given in "Appendices 1–4".

2 Main results

In this section, we define the lag-l Gasser-type estimator, and construct a new estimator based on these lagged estimators using least squares regression. It will be shown that the new estimator is more efficient than the Gasser et al. (1986)'s estimator, and reaches the optimal bound in terms of estimation variance. Meanwhile, the asymptotic normality is established.

2.1 Lagged Gasser-type estimator

Assume that x_i's are equidistantly designed, that is, $x_i = i/n$ for $1 \leq i \leq n$. For ease of notations, let $m_i = m(x_i)$, and $\gamma_4 = \mathrm{E}(\epsilon^4)/\sigma^4$. Define the lag-$l$ Gasser-type estimator as

$$s_l = \frac{2}{3(n-2l)} \sum_{i=l+1}^{n-l} \left(\frac{1}{2}Y_{i-l} - Y_i + \frac{1}{2}Y_{i+l} \right)^2 \quad (1 \leq l \leq k), \qquad (3)$$

which is similar to Müller and Stadtmüller (1999). The difference is that their methods focus on the Rice-type estimator while ours relies on the Gasser-type estimator to further reduce bias.

For the lag-l Gasser-type estimator s_l, we decompose it into three parts as

$$s_l = I_{1l} + I_{2l} + I_{3l},$$

where

$$I_{1l} = \frac{2}{3(n-2l)} \sum_{i=l+1}^{n-l} \left(\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} \right)^2,$$

$$I_{2l} = \frac{4}{3(n-2l)} \sum_{i=l+1}^{n-l} \left(\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} \right) \left(\frac{1}{2}\epsilon_{i-l} - \epsilon_i + \frac{1}{2}\epsilon_{i+l} \right),$$

$$I_{3l} = \frac{2}{3(n-2l)} \sum_{i=l+1}^{n-l} \left(\frac{1}{2}\epsilon_{i-l} - \epsilon_i + \frac{1}{2}\epsilon_{i+l} \right)^2.$$

Taking expectation of I_{2l} and I_{3l}, we have

$$\mathrm{E}(I_{2l}) = 0, \quad \mathrm{E}(I_{3l}) = \sigma^2.$$

Note that the lag-l Gasser-type estimator has a positive bias I_{1l}. Assume that $m(\cdot)$ has a bounded second-order derivative. According to the Taylor expansion of $m_{i\pm l}$ at x_i, we have

$$\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} = \frac{m_i^{(2)}}{2}\frac{l^2}{n^2} + o\left(\frac{l^2}{n^2}\right),$$

and

$$I_{1l} = \frac{2}{3(n-2l)} \sum_{i=1+l}^{n-l} \left(\frac{m_i^{(2)}}{2}\frac{l^2}{n^2} + o\left(\frac{l^2}{n^2}\right) \right)^2$$

$$= \frac{2}{3(n-2l)} \sum_{i=1+l}^{n-l} \left(\frac{m_i^{(2)}}{2} \right)^2 \frac{l^4}{n^4} + o\left(\frac{l^4}{n^4}\right)$$

$$\approx J\frac{l^4}{n^4},$$

where $J = (1/6)\int_0^1 \{m^{(2)}(x)\}^2 \mathrm{d}x$. For $l = o(n)$, the expectation of s_l is

$$\mathrm{E}(s_l) \approx \sigma^2 + J\frac{l^4}{n^4}.$$

2.2 Estimation methodology

For any fixed $k = o(n)$, we construct a linear regression model

$$s_l = \beta_0 + \beta_1 d_l + \delta_l \quad (l = 1, \ldots, k), \tag{4}$$

where $\beta = (\beta_0, \beta_1)^T = (\sigma^2, J)^T$, $d_l = l^4/n^4$, and $\delta_l = o(l^4/n^4) + I_{2l} + (I_{3l} - \sigma^2)$ with $\mathrm{E}(\delta_l) \approx 0$ and $var(\delta_l) \approx (\gamma_4 + 8/9)\sigma^4/n$ (see Theorem 2). The method is similar in spirit to the variance estimation (Müller and Stadtmüller 1999; Tong and Wang 2005) and the derivative estimation (Wang and Lin 2015).

Since the asymptotic variances of δ_l's are all the same, we use the least squares method to obtain the parameter estimation

$$\hat{\beta} = \arg\min_{\beta_0, \beta_1} \sum_{l=1}^{k} (s_l - \beta_0 - \beta_1 d_l)^2$$

$$= (X^T X)^{-1} X^T S,$$

where

$$X^T = \begin{pmatrix} 1 & \cdots & 1 \\ 1^4/n^4 & \cdots & k^4/n^4 \end{pmatrix}_{2 \times k}, \quad S^T = \left(s_1, \cdots, s_k \right)_{1 \times k}.$$

That is

$$\hat{\beta}_0 = \bar{s} - \hat{\beta}_1 \bar{d}, \quad \hat{\beta}_1 = \frac{\sum_{l=1}^{k} s_l (d_l - \bar{d})}{\sum_{l=1}^{k} (d_l - \bar{d})^2},$$

where $\bar{s} = (1/k) \sum_{l=1}^{k} s_l$ and $\bar{d} = (1/k) \sum_{l=1}^{k} d_l$. We define the proposed variance estimator as $\hat{\sigma}^2 = \hat{\beta}_0$. The following theorem gives some properties of the variance estimator $\hat{\sigma}^2$.

Theorem 1 *For equidistant designs, we have the following:*

1. $\hat{\sigma}^2$ *is unbiased when m is a linear or quadratic function regardless of the choice of k;*
2. $\hat{\sigma}^2$ *can be represented as a linear combination of $\{s_l\}_{l=1}^{k}$, i.e., $\hat{\sigma}^2 = (1/k) \sum_{l=1}^{k} b_l s_l$, where $b_l = 1 - \frac{k \bar{d} (d_l - \bar{d})}{\sum_{l=1}^{k} (d_l - \bar{d})^2} \approx \frac{25}{16} - \frac{45}{16} \frac{l^4}{k^4}$ for $l = 1, \dots, k$;*
3. $\hat{\sigma}^2$ *can be written as a quadratic form $\hat{\sigma}^2 = Y^T D Y$, where $D = (1/k) \sum_{l=1}^{k} b_l D_l$ is an $n \times n$ matrix with $tr(D) = 1$, and D_l is an $n \times n$ matrix corresponding to $s_l = Y^T D_l Y$ with $tr(D_l) = 1$.*

Remark 1 From the second point of Theorem 1, $\hat{\sigma}^2$ is a linear combination of $\{s_l\}_{l=1}^{k}$. The weighted average has two superiorities: it decreases asymptotic estimation variance from $(\gamma_4 + 8/9)\sigma^4$ to $(\gamma_4 - 1)\sigma^4$ (see Theorems 2 and 4); in addition, it eliminates the bias terms $\{Jl^4/n^4\}_{l=1}^{k}$ since $\sum_{l=1}^{k} b_l l^4/n^4 = 0$. Owing to $\sum_{l=1}^{k} b_l = k$ and $b_l < 0$ for $l > (5/9)^{1/4} k \approx 0.86k$, D is not guaranteed to be a positive definite matrix, and it is possible to take a negative value for $\hat{\sigma}^2$. However, our simulation indicates that this rarely happens. If negative estimates do happen in practice, then we recommend replacing them by zero as Tong and Wang (2005) suggested.

For finite samples, since s_l is the average of $(n - 2l)$ lag-l differences, it may be better to assign weight $w_l = (n - 2l)/N$ to s_l with $N = \sum_{l=1}^{k} (n - 2l)$. So we obtain the weighted least squares variance estimation

$$\hat{\beta}_0 = \bar{s}_w - \hat{\beta}_1 \bar{d}_w, \quad \hat{\beta}_1 = \frac{\sum_{l=1}^{k} w_l s_l (d_l - \bar{d}_w)}{\sum_{l=1}^{k} w_l (d_l - \bar{d}_w)^2},$$

where $\bar{s}_w = \sum_{l=1}^{k} w_l s_l$ and $\bar{d}_w = \sum_{l=1}^{k} w_l d_l$.

In the construction of the least squares estimator and its weighted version, we ignore the correlation between s_l's. It is well known that the generalized least squares estimator $\hat{\beta}_\Sigma = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} S$ is the best linear unbiased estimator (BLUE), where Σ is the asymptotic covariance matrix. "BLUE = OLSE" in linear regression if and only if the errors have equal variances and equal nonnegative correlation coefficients

by McElroy (1967). From the following Theorem 2 and Lemma 3, the generalized least squares estimator is identical to the ordinary least squares estimator. Thus the above three estimators are asymptotically equivalent.

2.3 Asymptotic results

Next, we establish the asymptotic normality for the lag-l Gasser-type estimator and our estimator.

Theorem 2 *Assume that m has a bounded second-order derivative. If $l = o(n^{7/8})$ as $n \to \infty$, then the lag-l Gasser estimator satisfies*

$$\sqrt{n}\left(s_l - \sigma^2\right) \xrightarrow{d} N\left(0, \left(\gamma_4 + \frac{8}{9}\right)\sigma^4\right).$$

With normal errors, the asymptotic variances for the lagged Rice-type and Gasser-type estimators are $3\sigma^4/n$ (Tong et al. 2013) and $(3 + 8/9)\sigma^4/n$ respectively; thus the Rice-type estimator is more efficient than the Gasser-type estimator theoretically. However, for finite samples the Gasser-type estimator would be better if the bias caused by regression function is controlled in the case of immense oscillation of regression function.

Lemma 3 *Assume that m has a bounded second-order derivative. If $v = o(n^{3/4})$ as $n \to \infty$ and $v > u$, then the covariance between lag-l Gasser-type estimators is*

$$\text{Cov}(s_u, s_v) = \begin{cases} (\gamma_4 - 1)\sigma^4/n + o(1/n), & v \neq 2u, \\ (\gamma_4 - \frac{13}{9})\sigma^4/n + o(1/n), & v = 2u. \end{cases}$$

Theorem 4 *Assume that m has a bounded second-order derivative. For any $k = O(n^r)$ with $0 < r < 3/4$ as $n \to \infty$, we have*

$$\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) \xrightarrow{d} N\left(0, (\gamma_4 - 1)\sigma^4\right).$$

Remark 2 There are two kinds of differenced methods: the direct-difference method and integral-based method. Most direct-difference estimators do not achieve the asymptotic optimal variance, i.e., Gasser et al. (1986)'s estimator has the asymptotic variance $(\gamma_4 + 8/9)\sigma^4$ from Theorem 2. As for the integral-based estimator, our estimator achieves the same optimal variance $(\gamma_4 - 1)\sigma^4$ from Theorem 4 as Tong and Wang (2005) did, but ours has less bias ($o(k^4/n^4)$) than Tong and Wang (2005)'s ($o(k^2/n^2)$) under the same conditions of equidistant design and bounded second-order derivative.

3 Simulations

3.1 Finite-sample choice of the bandwidth

In our comparisons, we evaluate the performance of the estimators of Rice (1984), Gasser et al. (1986), Hall et al. (1990), Tong and Wang (2005), and ours, which are notated as $\hat{\sigma}_R^2, \hat{\sigma}_{GSJ}^2, \hat{\sigma}_{HKT}^2, \hat{\sigma}_{TW}^2$, and $\hat{\sigma}^2$, respectively. For Hall et al. (1990)'s estimator, we set $k = 2$ in our simulations, that is

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

As for Tong and Wang (2005)'s estimator and ours, the bandwidth is chosen as $k = n^{1/3}, n^{1/2}$ for small and large n respectively, as Tong and Wang (2005) did. The further explanation is given recently by Wang and Lin (2015), Wang and Yu (2016).

3.2 Simulation results

The oscillation of periodic function depends on frequency and amplitude. We choose the regression function in our simulations to be

$$m(x) = A \sin(2\pi f x),$$

with design points $x_i = i/n$, and the errors are independent and identically distributed normal random variables with zero mean and variance σ^2. The regression function is similar to those in Seifert et al. (1993), Dette et al. (1998), and Tong and Wang (2005). We consider two sample sizes $n = 50, 500$, corresponding to small, large sample sizes, three standard deviations $\sigma = 0.1, 0.5, 2$, two frequencies $f = 0.5, 2$, and two amplitudes $A = 5, 50$.

For each simulation setting, we generate observations and compute the estimators $\hat{\sigma}_R^2, \hat{\sigma}_{GSJ}^2, \hat{\sigma}_{HKT}^2, \hat{\sigma}_{TW}^2, \hat{\sigma}^2$. We repeat this process 1000 times and compute relative MSEs for these estimators.

Table 1 lists relative MSEs of all estimators, $nMSE/(2\sigma^4)$, since the asymptotically optimal variance is $2\sigma^4/n$ with normal errors. In general, $\hat{\sigma}^2$ has smaller relative MSE in most settings. With normal errors, the relative MSEs are 1.25, 1.5, and 1.94 corresponding to $\hat{\sigma}_{HKT}^2, \hat{\sigma}_R^2$, and $\hat{\sigma}_{GSJ}^2$. However, $\hat{\sigma}_{GSJ}^2$ is robust in most settings because it reduces more bias than $\hat{\sigma}_R^2$ and $\hat{\sigma}_{HKT}^2$. $\hat{\sigma}_{TW}^2$ and $\hat{\sigma}^2$ both achieve the asymptotic optimal rate, but $\hat{\sigma}^2$ has less relative MSE than $\hat{\sigma}_{TW}^2$ in all settings except that they have the similar performance in the case within large sample and variance, low frequency, and small amplitude occur simultaneously.

Next we explain the reason why there are some quite large values in Table 1. The asymptotic MSE is composed of bias and variance, that is $MSE = Bias^2 + Var$. By Theorems 2 and 4, we known that the variance can be controlled as the sample size n becomes large. However, the bias is fixed which is determined by the used method and the oscillation of regression function. For example, $AMSE(\hat{\sigma}_R^2) = 1.47E2$ is large

Table 1 Relative MSEs of various estimators. Note that $1.47\text{E}2 = 1.47 * 10^2$

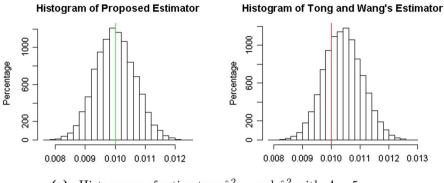| n	A	f	σ	$\hat{\sigma}_R^2$	$\hat{\sigma}_{GSJ}^2$	$\hat{\sigma}_{HKT}^2$	$\hat{\sigma}_{TW}^2$	$\hat{\sigma}^2$
50	5	0.5	0.1	1.47E2	1.88	8.70E2	2.61	1.46
			0.5	1.72	1.98	2.67	1.51	1.50
			2	1.45	1.96	1.19	1.50	1.45
		2	0.1	3.70E4	2.01E1	2.16E5	7.10E2	3.14
			0.5	6.11E1	2.10	3.49E2	2.95	1.70
			2	1.64	1.91	2.52	1.47	1.44
	50	0.5	0.1	1.46E6	4.64	8.72E6	9.40E3	1.94
			0.5	2.34E3	1.95	1.39E4	1.81E1	1.56
			2	1.07E1	2.00	5.62E1	1.70	1.55
		2	0.1	3.70E8	1.83E5	2.17E9	7.08E6	1.68E4
			0.5	5.93E5	2.94E2	3.46E6	1.13E4	2.87E1
			2	2.31E3	2.96	1.35E4	4.61E1	1.52
500	5	0.5	0.1	1.75	2.00	2.29	1.48	1.06
			0.5	1.51	1.95	1.27	1.06	1.07
			2	1.49	1.93	1.24	1.04	1.06
		2	0.1	4.05E1	1.88	2.44E2	2.62E2	1.07
			0.5	1.65	2.06	1.71	1.51	1.10
			2	1.50	1.95	1.26	1.06	1.07
	50	0.5	0.1	1.52E3	1.93	9.44E3	4.08E3	1.10
			0.5	3.85	2.03	1.61E1	7.73	1.13
			2	1.53	1.92	1.37	1.10	1.09
		2	0.1	3.88E5	1.88E2	2.41E6	2.59E6	2.13E2
			0.5	6.21E2	2.07	3.86E3	4.14E3	1.45
			2	3.94	1.91	1.65E1	1.74E1	1.05

enough for $n = 50, A = 5, f = 0.5, \sigma = 0.1$. We compute the expectation of $\hat{\sigma}_R^2$ such that

$$
\begin{aligned}
\text{E}(\hat{\sigma}_R^2) &= \text{E}\left(\frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_i - Y_{i-1})^2 \right) \\
&= \text{E}\left(\frac{1}{2(n-1)} \sum_{i=2}^{n} (m_i - m_{i-1})^2 + (\epsilon_i - \epsilon_{i-1})^2 + 2(m_i - m_{i-1})(\epsilon_i - \epsilon_{i-1}) \right) \\
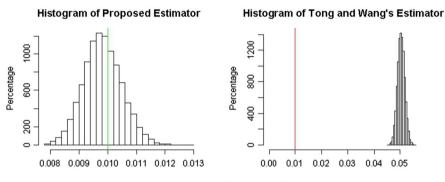&= \sigma^2 + \frac{1}{2(n-1)} \sum_{i=2}^{n} (m_i - m_{i-1})^2.
\end{aligned}
$$

Thus the bias is

$$
\text{Bias}(\hat{\sigma}_R^2) = \frac{1}{2(n-1)} \sum_{i=2}^{n} (m_i - m_{i-1})^2 \approx 0.025,
$$

(a) Histograms of estimators $\hat{\sigma}_{TW}^2$ and $\hat{\sigma}^2$ with $A = 5$.

(b) Histograms of estimators $\hat{\sigma}_{TW}^2$ and $\hat{\sigma}^2$ with $A = 50$.

Fig. 1 Histograms of the variance estimators $\hat{\sigma}_{TW}^2$ and $\hat{\sigma}^2$ for the cases $(n, A, f, \sigma^2) = (500, 5, 0.5, 0.01)$ and $(500, 50, 0.5, 0.01)$ in 10,000 simulations. **a** Histograms of estimators $\hat{\sigma}_{TW}^2$ and $\hat{\sigma}^2$ with $A = 5$. **b** Histograms of estimators $\hat{\sigma}_{TW}^2$ and $\hat{\sigma}^2$ with $A = 50$

and now

$$nMSE/(2\sigma^4) \approx 50 * Bias^2(\hat{\sigma}_R^2)/(2\sigma^4) \approx 156,$$

which is close to 147.

To further show the importance of bias-correction, we consider the two cases $(n, A, f, \sigma^2) = (500, 5, 0.5, 0.01)$ and $(500, 50, 0.5, 0.01)$. Figure 1 indicates that $\hat{\sigma}_{TW}^2$ and $\hat{\sigma}^2$ both have asymptotic normality with variance $\sigma^2 = 0.01$. However, $\hat{\sigma}_{TW}^2$ has small bias (0.0004) while $\hat{\sigma}^2$ almost has no bias with $A = 5$; as amplitude A varies from 5 to 50, $\hat{\sigma}_{TW}^2$ has larger bias (0.04) while $\hat{\sigma}^2$ has little bias (0.00018). So the new estimator based on the Gasser-type estimator controls bias much better than Tong and Wang (2005)'s estimator based on the Rice-type estimator.

4 Discussion

In this paper, we propose a new variance estimator in nonparametric regression model. The new estimator achieves the asymptotic optimal rate for the MSE, meanwhile it is less-biased than most differenced estimators.

This work concentrates on the equidistant design. We can generalize the idea to non-equidistant design. Following one reviewer's comment, we assume that $x_i = g(i/n)$ for $i = 1, \ldots, n$, where the function $g(\cdot)$ has a positive derivative function and $c \leq g(x) \leq 1$ for some constant $0 < c < 1$. We expand $x_{i+l} - x_i$, $x_{i+l} - x_{i-1}$, and $x_i - x_{i-1}$ on the same design point i/n such that

$$x_{i+l} - x_i \approx g^{(1)}(i/n)l/n,$$
$$x_{i+l} - x_{i-l} \approx 2g^{(1)}(i/n)l/n,$$
$$x_i - x_{i-l} \approx g^{(1)}(i/n)l/n.$$

Thus in the sense of linear approximation, we have the same result as the equidistant design such that

$$\tilde{s}_l = \frac{1}{n-2l} \sum_{i=l+1}^{n-l} \frac{\{(x_{i+l} - x_i)Y_{i-l} - (x_{i+l} - x_{i-l})Y_i + (x_i - x_{i-l})Y_{i+l}\}^2}{(x_{i+l} - x_i)^2 + (x_{i+l} - x_{i-l})^2 + (x_i - x_{i-l})^2}$$

$$\approx \frac{2}{3(n-2l)} \sum_{i=l+1}^{n-l} \left(\frac{1}{2}Y_{i-l} - Y_i + \frac{1}{2}Y_{i+l}\right)^2.$$

Further, assume that the function $g(\cdot)$ is two times continuous differentiable. We expand $x_{i+l} - x_i$, $x_{i+l} - x_{i-1}$, and $x_i - x_{i-1}$ on the point i/n such that

$$x_{i+l} - x_i \approx g^{(1)}(i/n)l/n + \frac{g^{(2)}(i/n)}{2}(l/n)^2,$$
$$x_{i+l} - x_{i-l} \approx g^{(1)}(i/n)2l/n,$$
$$x_i - x_{i-l} \approx g^{(1)}(i/n)l/n - \frac{g^{(2)}(i/n)}{2}(l/n)^2.$$

Thus we have

$$\tilde{s}_l = \frac{1}{n-2l} \sum_{i=l+1}^{n-l} \frac{\{(x_{i+l} - x_i)Y_{i-l} - (x_{i+l} - x_{i-l})Y_i + (x_i - x_{i-l})Y_{i+l}\}^2}{(x_{i+l} - x_i)^2 + (x_{i+l} - x_{i-l})^2 + (x_i - x_{i-l})^2}$$

$$\approx \frac{1}{n-2l} \sum_{i=l+1}^{n-l} \frac{\{g^{(1)}(i/n)(Y_{i-l} - 2Y_i + Y_{i+l})(l/n) + g^{(2)}(i/n)(Y_{i-l} - Y_{i+l})(l/n)^2/2\}^2}{6[g^{(1)}(i/n)]^2(l/n)^2 + [g^{(2)}(i/n)]^2(l/n)^4/2}$$

$$\approx \frac{2}{3(n-2l)} \sum_{i=l+1}^{n-l} \left(\frac{1}{2}Y_{i-l} - Y_i + \frac{1}{2}Y_{i+l}\right)^2$$

$$+ \frac{1}{6(n-2l)} \sum_{i=l+1}^{n-l} \frac{g^{(2)}(i/n)}{g^{(1)}(i/n)}(Y_{i-l} - 2Y_i + Y_{i+l})(Y_{i-l} - Y_{i+l})(l/n).$$

To eliminate the last term of the above equation, we need the conditions that $g^{(1)}(i/n)$ is away from 0 and $g^{(2)}(i/n)$ is close to 0. Or else the last term is non-ignorable and our proposed method is not applied directly. In particular case $g(x) = x$ corresponding to the equidistant design, we have that $g^{(1)}(i/n) \equiv 1$ and $g^{(2)}(i/n) \equiv 0$, which satisfy the above conditions.

In addition, our proposed method concentrates on univariate x, while some differenced methods in the literature focused on multivariate x. Hall et al. (1991) generalized the idea in Hall et al. (1990) to the bivariate lattice design. Munk et al. (2005) further proposed a differenced estimator for multivariate regression (the dimension number $d \leq 4$). Our method can be generalized to the bivariate lattice design as Hall et al. (1991) did. The generalization may be more efficient in some particular cases. For example, when the surface has the similar variation trend in the horizontal direction, difference in the vertical direction can eliminate the trend to retain the error information.

Appendix 1: Proof of Theorem 1

Let $m(x) = a + bx + cx^2$. We have

$$E\left(\left(\frac{1}{2}Y_{i-l} - Y_i + \frac{1}{2}Y_{i+l}\right)^2\right) = \frac{3}{2}\sigma^2 + c^2\frac{l^4}{n^4},$$

$$E(s_l) = \sigma^2 + \frac{2c^2}{3}\frac{l^4}{n^4}.$$

Since it is an accurate linear regression model, the coefficient estimation is unbiased. Thus,

$$E(\hat{\beta}_1) = \beta_1, \quad E(\hat{\beta}_0) = \beta_0.$$

Variance estimator $\hat{\sigma}^2$ is a linear combination of $\{s_l\}_{l=1}^k$, that is

$$\hat{\sigma}^2 = \bar{s} - \hat{\beta}_1\bar{d}$$
$$= 1/k\left(\sum_{l=1}^k s_l - \frac{\sum_{l=1}^k ks_l(d_l - \bar{d})\bar{d}}{\sum_{l=1}^k (d_l - \bar{d})^2}\right)$$

$$= 1/k \sum_{l=1}^{k} b_l s_l,$$

where $b_l = 1 - \frac{k\bar{d}(d_l - \bar{d})}{\sum_{l=1}^{k}(d_l - \bar{d})^2} \approx \frac{25}{16} - \frac{45}{16} \frac{l^4}{k^4}$.

According to the proof of (2) and $\sum_{l=1}^{k} b_l = k$, (3) is straightforward.

Appendix 2: Proof of Theorem 2

For lag-l Gasser-type estimator,

$$s_l = I_{1l} + I_{2l} + I_{3l},$$

Applying Taylor expansion to I_{1l}, we can show that $I_{1l} = (l^4/n^4)J + o(l^4/n^4) = o_p(n^{-1/2})$ if $n^{-1}l^{8/7} \to 0$ as $n \to \infty$. For I_{2l}, we have

$$
\begin{aligned}
E(I_{2l}^2) &= \frac{16}{9(n-2l)^2} \sum_{i=l+1}^{n-l} \sum_{j=l+1}^{n-l} \left(\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} \right) \left(\frac{1}{2}m_{j-l} - m_j + \frac{1}{2}m_{j+l} \right) \\
&\quad \times E\left\{ \left(\frac{1}{2}\epsilon_{i-l} - \epsilon_i + \frac{1}{2}\epsilon_{i+l} \right) \left(\frac{1}{2}\epsilon_{j-l} - \epsilon_j + \frac{1}{2}\epsilon_{j+l} \right) \right\} \\
&= \frac{16}{9(n-2l)^2} \left\{ \sum_{i=j=l+1}^{n-l} \frac{3\sigma^2}{2} \left(\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} \right)^2 \right. \\
&\quad - \sum_{i=j+l=2l+1}^{n-l} \sigma^2 \left(\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} \right) \\
&\quad \times \left(\frac{1}{2}m_i - m_{i+l} + \frac{1}{2}m_{i+2l} \right) - \sum_{i=j-l=l+1}^{n-2l} \\
&\quad \times \sigma^2 \left(\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} \right) \left(\frac{1}{2}m_{i-2l} - m_{i-l} + \frac{1}{2}m_i \right) \\
&\quad + \sum_{i=j+2l=3l+1}^{n-l} \sigma^2/4 \left(\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} \right) \left(\frac{1}{2}m_{i-3l} - m_{i-2l} + \frac{1}{2}m_{i-l} \right) \\
&\quad \left. + \sum_{i=j-2l=l+1}^{n-3l} \sigma^2/4 \left(\frac{1}{2}m_{i-l} - m_i + \frac{1}{2}m_{i+l} \right) \left(\frac{1}{2}m_{i+3l} - m_{i+2l} + \frac{1}{2}m_{i+l} \right) \right\} \\
&= O(l^4/n^5)
\end{aligned}
$$

which implies that $I_{2l} = o_p(n^{-1/2})$ for any $l = o(n)$.

Rewrite I_{3l} as $I_{3l} = \sigma^2 + 1/(n-2l) \sum_{i=l+1}^{n-l} \eta_i(l)$, where $\eta_i(l) = 2/3(1/2\epsilon_{i-l} - \epsilon_i + 1/2\epsilon_{i+l})^2 - \sigma^2$. For any l, $\{\eta_i(l)\}_{i=l+1}^{n-l}$ is a strictly stationary sequence of random

variables with mean zero and autocovariance function

$$\gamma(t) = \gamma(u, u+t) = \begin{cases} [(\gamma_4 + 1)/2]\sigma^4 & t = 0, \\ [2(\gamma_4 + 1)/9]\sigma^4 & t = l, \\ [(\gamma_4 - 1)/36]\sigma^4 & t = 2l, \\ 0 & otherwise. \end{cases}$$

Note that the sequence $\{\eta_i(l)\}_{i=l+1}^{n-l}$ is $2l$-dependent. Thus by the central limit theorem in Whittle (1964), $\sqrt{n}(I_{3l} - \sigma^2) \xrightarrow{d} N(0, \zeta^2)$, where $\zeta^2 = \gamma(0) + 2\sum_{t=1}^{2l}\gamma(t) = (\gamma_4 + 8/9)\sigma^4$. Using $I_{1l} = o_p(n^{-1/2})$, $I_{2l} = o_p(n^{-1/2})$ and $s_l = I_{1l} + I_{2l} + I_{3l}$, we have

$$\sqrt{n}(s_l - \sigma^2) \xrightarrow{d} N\left(0, (\gamma_4 + \frac{8}{9})\sigma^4\right).$$

Appendix 3: Proof of Lemma 3

For $1 \le u < v = o(n)$, we have

$$E(s_u s_v) = \frac{4}{9(n-2u)(n-2v)} E[L_1(u,v) + L_2(u,v) + L_3(u,v) \\ + L_4(u,v) + L_5(u,v)],$$

where

$$L_1(u,v) = \sum_{i=u+1}^{n-u}\left(\frac{1}{2}\epsilon_{i-u} - \epsilon_i + \frac{1}{2}\epsilon_{i+u}\right)^2 \sum_{j=v+1}^{n-v}\left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v}\right)^2,$$

$$L_2(u,v) = \sum_{i=u+1}^{n-u}\left(\frac{1}{2}m_{i-u} - m_i + \frac{1}{2}m_{i+u}\right)^2 \sum_{j=v+1}^{n-v}\left(\frac{1}{2}m_{j-v} - m_j + \frac{1}{2}m_{j+v}\right)^2,$$

$$L_3(u,v) = \sum_{i=u+1}^{n-u}\left(\frac{1}{2}\epsilon_{i-u} - \epsilon_i + \frac{1}{2}\epsilon_{i+u}\right)^2 \sum_{j=v+1}^{n-v}\left(\frac{1}{2}m_{j-v} - m_j + \frac{1}{2}m_{j+v}\right)^2,$$

$$L_4(u,v) = \sum_{i=u+1}^{n-u}\left(\frac{1}{2}m_{i-u} - m_i + \frac{1}{2}m_{i+u}\right)^2 \sum_{j=v+1}^{n-v}\left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v}\right)^2.$$

$$L_5(u,v) = 2\sum_{i=u+1}^{n-u}\left(\frac{1}{2}m_{i-u} - m_i + \frac{1}{2}m_{i+u}\right)\left(\frac{1}{2}\epsilon_{i-u} - \epsilon_i + \frac{1}{2}\epsilon_{i+u}\right)$$

$$\times \sum_{j=v+1}^{n-v}\left[\left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v}\right)^2\right.$$

$$\left. + \left(\frac{1}{2}m_{j-v} - m_j + \frac{1}{2}m_{j+v}\right)^2 + 2\left(\frac{1}{2}m_{j-v} - m_j + \frac{1}{2}m_{j+v}\right)\right.$$

$$\times \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right) \Bigg]$$

$$+ 2 \sum_{j=v+1}^{n-v} \left(\frac{1}{2}m_{j-v} - m_j + \frac{1}{2}m_{j+v} \right) \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)$$

$$\times \sum_{i=u+1}^{n-u} \left[\left(\frac{1}{2}\epsilon_{i-u} - \epsilon_i + \frac{1}{2}\epsilon_{i+u} \right)^2 \right.$$

$$\left. + \left(\frac{1}{2}m_{i-u} - m_i + \frac{1}{2}m_{i+u} \right)^2 \right].$$

Applying the second-order Taylor expansion, we have

$$E[L_2(u, v)] = \frac{9J^2 u^4 v^4}{4n^6} + o\left(\frac{u^4 v^4}{n^6} \right),$$

$$E[L_3(u, v)] = \frac{9J\sigma^2}{4} \frac{v^4}{n^2} + o\left(\frac{v^4}{n^2} \right),$$

$$E[L_4(u, v)] = \frac{9J\sigma^2}{4} \frac{u^4}{n^2} + o\left(\frac{u^4}{n^2} \right),$$

$$E[L_5(u, v)] = O\left(\frac{u^2 v^2}{n^2} \right).$$

Next, we decompose $L_1(u, v)$ into ten parts such as

$$L_1(u, v) = L_{11} + L_{12} + L_{13} + L_{14} + L_{15} + L_{16} + L_{17} + L_{18} + L_{19} + L_{10},$$

where

$$L_{11} = \sum_{j=v+1}^{n-v} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_{j-u} - \epsilon_j + \frac{1}{2}\epsilon_{j+u} \right)^2,$$

$$L_{12} = \sum_{j=v+1}^{n-v} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_{j+v-2u} - \epsilon_{j+v-u} + \frac{1}{2}\epsilon_{j+v} \right)^2,$$

$$L_{13} = \sum_{j=v+1}^{n-v} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_{j-v+u} + \frac{1}{2}\epsilon_{j-v+2u} \right)^2,$$

$$L_{14} = \sum_{j=v+1}^{n-2u-v} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_{j+v} - \epsilon_{j+u+v} + \frac{1}{2}\epsilon_{j+2u+v} \right)^2,$$

$$L_{15} = \sum_{j=2u+v+1}^{n-v} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_{j-u-2v} - \epsilon_{j-u-v} + \frac{1}{2}\epsilon_{j-v} \right)^2,$$

$$L_{16} = \sum_{j=u+v+1}^{n-v} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_{j-u-v} - \epsilon_{j-v} + \frac{1}{2}\epsilon_{j+u-v} \right)^2,$$

$$L_{17} = \sum_{j=v+1}^{n-u-v} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_{j-u+v} - \epsilon_{j+v} + \frac{1}{2}\epsilon_{j+u+v} \right)^2,$$

$$L_{18} = \sum_{j=v+1}^{n-\max\{v,2u\}} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_j - \epsilon_{j+u} + \frac{1}{2}\epsilon_{j+2u} \right)^2,$$

$$L_{19} = \sum_{j=\max\{v,2u\}+1}^{n-v} \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2 \left(\frac{1}{2}\epsilon_{j-2u} - \epsilon_{j-u} + \frac{1}{2}\epsilon_j \right)^2,$$

$$L_{10} = \sum_{(i,j)\in\Gamma} \left(\frac{1}{2}\epsilon_{i-u} - \epsilon_i + \frac{1}{2}\epsilon_{i+u} \right)^2 \left(\frac{1}{2}\epsilon_{j-v} - \epsilon_j + \frac{1}{2}\epsilon_{j+v} \right)^2,$$

and $\Gamma = \{(i,j) | u+1 \leq i \leq n-u; v+1 \leq j \leq n-v; \{i-u, i, i+u\} \cap \{j-v, j, j+v\} = \emptyset\}$. When $v \neq 2u$, it is easy to verify the following

$$\mathrm{E}(L_{11}) = \frac{(16\gamma_4 + 20)\sigma^4 n}{16} + o(n), \quad \mathrm{E}(L_{10}) = \frac{9\sigma^4[(n-2u)(n-2v) - 9n]}{4} + o(n),$$

$$\mathrm{E}(L_{12}) = \mathrm{E}(L_{13}) = \mathrm{E}(L_{14}) = \mathrm{E}(L_{15}) = \frac{(\gamma_4 + 35)\sigma^4 n}{16} + o(n),$$

$$\mathrm{E}(L_{16}) = \mathrm{E}(L_{17}) = \mathrm{E}(L_{18}) = \mathrm{E}(L_{19}) = \frac{(4\gamma_4 + 32)\sigma^4 n}{16} + o(n).$$

Therefore,

$$\mathrm{E}(s_u s_v) = \sigma^4 + \frac{(\gamma_4 - 1)\sigma^4}{n} + O\left(\frac{v^4}{n^4}\right) + o\left(\frac{1}{n}\right).$$

When $v = 2u$, note that $L_{12} = L_{18}$, $L_{13} = L_{19}$ and $\mathrm{E}(L_{12}) = (5\gamma_4 + 23)\sigma^4 n/16 + o(n)$. Thus,

$$\mathrm{E}(s_u s_v) = \sigma^4 + \frac{(\gamma_4 - \frac{13}{9})\sigma^4}{n} + O\left(\frac{v^4}{n^4}\right) + o\left(\frac{1}{n}\right).$$

Finally, if $v = o(n^{3/4})$ as $n \to \infty$, then

$$\mathrm{Cov}(s_u, s_v) = \begin{cases} (\gamma_4 - 1)\sigma^4/n + o(1/n), & v \neq 2u, \\ (\gamma_4 - \frac{13}{9})\sigma^4/n + o(1/n), & v = 2u. \end{cases}$$

Appendix 4: Proof of Theorem 4

When $k = O(n^r)$ with $0 < r < 3/4$ as $n \to \infty$, the estimation bias is negligible, and Theorems 1 and 3 are established simultaneously. Therefore $\hat{\sigma}^2$ is consistent and asymptotically normal since it is a least squares estimator. Next we derive the asymptotic variance

$$
\begin{aligned}
\text{Var}(\hat{\sigma}^2) &= \frac{1}{k^2} \sum_{u=1}^{k} \sum_{v=1}^{k} b_u b_v cov(s_u, s_v) \\
&= \frac{\gamma_4 - 1}{n} \sigma^4 + o\left(\frac{1}{n}\right).
\end{aligned}
$$

So we have

$$
\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, (\gamma_4 - 1)\sigma^4).
$$

References

Buckley MJ, Eagleson GK, Silverman BW (1988) The estimation of residual variance in nonparametric regression. Biometrika 75:189–199

Carter CK, Eagleson GK (1992) A comparison of variance estimators in nonparametric regression. J R Stat Soc B 54:773–780

Carter CK, Eagleson GK, Silverman BW (1992) A comparison of the reinsch and speckman splines. Biometrika 71:81–91

Dai W, Tong T (2014) Variance estimation in nonparametric regression with jump discontinuities. J Appl Stat 41:530–545

Dette H, Munk A, Wagner T (1998) Estimating teh variance in nonparametric regression-what is a reasonable choice? J Am Stat Assoc 60:751–764

Du J, Schick A (2009) A covariate-matched estimator of the error variance in nonparametric regression. J Nonparametr Stat 21(3):263–285

Eubank RL (1999) Nonparametric regression and spline smoothing. Marcel Dekker Inc, New York

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

Hall P, Carroll RJ (1989) Variance function estimation in regression: the effect of estimating the mean. J R Stat Soc B 51:3–14

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

Hall P, Kay JW, Titterington DM (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika 77:521–528

Hall P, Kay JW, Titterington DM (1991) On estimation of noise variance in two-dimensional signal processing. Appl Probab Trust 23(3):476–495

McElroy F (1967) A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased. J Am Stat Assoc 62(320):1302–1304

Müller HG, Stadtmüller U (1987) Estimation of heteroscedasticity in regression analysis. Ann Stat 15:610–635

Müller HG, Stadtmüller U (1999) Discontinuous versus smooth regression. Ann Stat 27:1215–1230

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

Munk A, Bissantz N, Wagner T, Freitag G (2005) On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. Aust N Z J Stat 44:479–488

Neumann MH (1994) Fully data-driven nonparametric variance estimators. Statistics 25:189–212

Rice JA (1984) Bandwidth choice for nonparametric regression. Ann Stat 12:1215–1230

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

Tong T, Wang Y (2005) Estimating residual variance in nonparametric regression using least squares. Biometrika 92(4):821–830

Tong T, Ma Y, Wang Y (2013) Optimal variance estimation without estimating the mean function. Bernoulli 19(5A):1839–1854

Wang Y (2011) Smoothing splines: methods and applications. CRC Press, Boca Raton

Wang WW, Lin L (2015) Derivative estimation based on difference sequence via locally weighted least squares regression. J Mach Learn Res 16:2617–2641

Wang WW, Yu P (2016) Asymptotically optimal differenced estimators of error variance in nonparametric regression. Comput Stat Data Anal. Technical report, The University of Hong Kong. http://web.hku.hk/~pingyu/KernelD.pdf

Whittle P (1964) On the convergence to normality of quadratic forms in independent variables. Theory Probab Appl 9:103–108

Zhou Y, Cheng Y, Wang L, Tong T (2015) Optimal difference-based variance estimation in heteroscedastic nonparametric regression. Stat Sin 25:1377–1397