Estimating residual variance in nonparametric regression using least squares

By TIEJUN TONG and YUEDONG WANG

Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA

tong@pstat.ucsb.edu yuedong@pstat.ucsb.edu

Abstract

We propose a new estimator for the error variance in a nonparametric regression model. We estimate the error variance as the intercept in a simple linear regression model with squared differences of paired observations as the dependent variable and squared distances between the paired covariates as the regressor. Our method can be applied to nonparametric regression models with multivariate functions defined on arbitrary subsets of normed spaces, possibly observed on unequally spaced or clustered designed points. No ordering is required for our method. We develop methods for selecting the bandwidth. For the special case of one dimensional domain with equally spaced design points, we show that our method reaches an asymptotic optimal rate which is not achieved by some existing methods. We conduct extensive simulations to evaluate finite sample performance of our method and compare it with existing methods. We illustrate our method using a real data set.

Some key words: Bandwidth; Difference-based estimator; Least square; Nonparametric regression; Quadratic forms; Residual variance.

1. Introduction

Consider a nonparametric regression model

$$y_i = g(x_i) + \epsilon_i, \quad 1 \leq i \leq n, \tag{1}$$

where y_i's are observations, g is an unknown mean function, and ϵ_i's are independent and identically distributed random errors with zero mean and variance σ^2.

Usually one fits the mean function g first and then estimates the variance σ^2 from residual sum of squares (Wahba 1990, Müller and Stadtmüller 1987, Hall and Carroll 1989, Carter and Eagleson 1992, Neumann 1994). However, it is often desirable to have an accurate estimate of σ^2, independent of that obtained by curve fitting, for the purpose of testing the goodness of fit or choosing the amount of smoothing (Eubank and Spiegelman 1990, Rice 1984, Gasser, Kneip and Kohler 1991, Kulasekera and Gallagher 2002). An accurate estimate of σ^2 can also be used to estimate the detection limits of immunoassay (Carroll 1987, Carroll and Ruppert 1988).

Most estimators of σ^2 proposed in the literature are quadratic forms of the response vector $\mathbf{y} = (y_1, \cdots, y_n)^T$,

$$\hat{\sigma}_D^2 = \mathbf{y}^T \mathbf{D} \mathbf{y} / tr(\mathbf{D}). \tag{2}$$

These estimators usually fall into two classes. The first class of estimators are based on the residual sum of squares from some nonparametric fits to g. Specifically, one first estimates g by a nonparametric method such as kernel smoothing or spline smoothing (Wahba 1990, Hastie and Tibshirani 1990). For linear smoothers the fitted values $\hat{\mathbf{y}} = \mathbf{A}\mathbf{y}$, where $\mathbf{A}$ is a smoother matrix. Then an estimator of variance has the form (1.2) with $\mathbf{D} = (\mathbf{I} - \mathbf{A})^T(\mathbf{I} - \mathbf{A})$ (Hastie and Tibshirani 1990). We call estimators in the first class as residual-based estimators. Residual-based estimators depend critically on the amount of smoothing (Dette, Munk and Wagner 1998). Some methods require knowledge about some unknown quantities such as $\int_0^1 g'(t)^2 dt$ (Hall and Marron 1990) or $\int_0^1 g''(t)^2 dt$ (Buckley, Eagleson and Silverman 1988).

The second class of estimators use differences to remove trend in the mean function, an idea originated from time series analysis. This kind of method does not require an estimate of the mean function and are often called the difference-based estimators. Almost all difference-based methods in the literature were developed for univariate x only with the exception of Kulasekera and Gallagher (2002) who extended the differenced-based method to multivariate $x \in [0, 1]^d$. However, Kulasekera and Gallagher's method requires an artificial ordering of the design points in $x \in [0, 1]^d$. In this paper we will propose a new method which can be applied to any domain $x \in \mathcal{T}$, where $\mathcal{T}$ is an arbitrary subset of a normed space. Interesting examples of $\mathcal{T}$ are the Euclidean d-space $\mathcal{R}^d$, unit circle and unit sphere.

Before introducing our new estimator, we review some popular difference-based estimators. Assume that x is univariate and $0 \leq x_1 \leq \cdots \leq x_n \leq 1$. Rice (1984) proposed the first order difference-based estimator

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2. \tag{3}$$

Gasser, Sroka and Jennen-Steinmetz (1986) proposed the following second order difference-based estimator (referred to as the GSJ estimator in this article)

$$\hat{\sigma}_{GSJ}^2 = \frac{1}{(n-2)} \sum_{i=2}^{n-1} c_i^2 \hat{\epsilon}_i^2, \tag{4}$$

where $\hat{\epsilon}_i$ is the difference between y_i and the value at x_i of the line joining $(x_{i-1},\ y_{i-1})$ and $(x_{i+1},\ y_{i+1})$. The coefficients c_i are chosen such that $E c_i^2 \hat{\epsilon}_i^2 = \sigma^2$ for all i when g is linear.

For equidistant design points, $\hat{\sigma}^2_{GSJ}$ reduces to

$$\hat{\sigma}^2_{GSJ} = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} \left(\frac{1}{2} y_{i-1} - y_i + \frac{1}{2} y_{i+1} \right)^2 . \tag{5}$$

Hall, Kay and Titterington (1990) introduced another difference-based estimator (referred to as the HKT estimator)

$$\hat{\sigma}^2_{HKT}(m) = \frac{1}{n-m} \sum_{i=m_1+1}^{n-m_2} \left(\sum_{k=-m_1}^{m_2} d_k y_{k+i} \right)^2 , \tag{6}$$

where m_1 and m_2 are non-negative integers, $m = m_1 + m_2$ is called the order, and the difference sequence $\{d_i\}_{i=-m_1,\cdots,m_2}$ satisfies $\sum_{j=-m_1}^{m_2} d_j = 0$, $\sum_{j=-m_1}^{m_2} d_j^2 = 1$ and $d_{-m_1} d_{m_2} \neq 0$.

None of the difference-based estimators achieves the asymptotic optimal rate for the mean squared error (MSE) (Dette et al. 1998)

$$MSE(\hat{\sigma}^2) \triangleq E(\hat{\sigma}^2 - \sigma^2)^2 = n^{-1} var(\epsilon^2) + o(n^{-1}). \tag{7}$$

In practice, the choice of the order m and an appropriate difference sequence which minimizes the finite sample MSE is rather complicated. Dette et al. (1998) showed that for a finite sample size, a proper choice of the order m depends sensitively on the oscillation of the mean function g and the sample size n. That is, the order m acts as a tuning parameter. In this paper we propose a new estimator which is the estimated intercept of a linear model. When design points are equally spaced in $[0, 1]$, using the optimal bandwidth, we can reduce the asymptotic rate of MSE to

$$MSE(\hat{\sigma}^2) = n^{-1} var(\epsilon^2) + O(n^{-3/2}).$$

And more importantly, our method extends naturally to functions defined on a general domain.

In Sections 2 and 3 we consider equally spaced designs in $[0, 1]$. We present our estimator, asymptotic results and the choice of the optimal bandwidth in Section 2. We compare the performance of our estimator with several popular difference-based estimators in Section 3. We extend the proposed method to general domain $\mathcal{T}$ in Section 4. We apply our method to a real data set in Section 5. We conclude the paper with a brief discussion in Section 6.

2. Main results

In this and the next sections we assume that $x_i = i/n$ for $1 \leq i \leq n$. In Section 2.1 we provide the motivation to our method. In Sections 2.2 and 2.3 we present the methodology and some asymptotic results. Then we discuss how to choose the bandwidth in Section 2.4.

2·1. *Motivation*

Taking expectation to the Rice estimator,

$$E(\hat{\sigma}_R^2) = \frac{1}{2(n-1)} \sum_{i=2}^{n} E(y_i - y_{i-1})^2 = \sigma^2 + \frac{1}{2(n-1)} \sum_{i=2}^{n} (g(x_i) - g(x_{i-1}))^2. \tag{8}$$

This means that Rice's estimator is always positively biased. Suppose that g has a bounded first derivative. Then from (8), we have

$$\begin{aligned} E(\hat{\sigma}_R^2) &= \sigma^2 + \frac{1}{2(n-1)} \sum_{i=2}^{n} \left(\frac{1}{n} g'(x_i) + o(\frac{1}{n}) \right)^2 \\ &= \sigma^2 + \frac{1}{2n^2} \frac{1}{(n-1)} \sum_{i=2}^{n} g'(x_i)^2 + o(\frac{1}{n^2}) \\ &= \sigma^2 + \frac{1}{n^2} J + o(\frac{1}{n^2}), \end{aligned} \tag{9}$$

where $J = \frac{1}{2} \int_0^1 g'(x)^2 dx$. Rice's estimator uses differences of all consequent observations. We define a lag-k Rice estimator $\hat{\sigma}_R^2(k)$ as

$$\hat{\sigma}_R^2(k) = \frac{1}{2(n-k)} \sum_{i=k+1}^{n} (y_i - y_{i-k})^2, \qquad k = 1, 2, \cdots, n-1.$$

Similar calculations as in (9) give

$$E(\hat{\sigma}_R^2(k)) = \sigma^2 + \frac{k^2}{n^2} J + O(\frac{k^3}{n^2(n-k)}) + o(\frac{1}{n^2}).$$

When $m = o(n)$, we have

$$E(\hat{\sigma}_R^2(k)) \approx \sigma^2 + J d_k, \qquad 1 \le k \le m, \tag{10}$$

where $d_k = k^2/n^2$. Treating (10) as a simple linear regression model with d_k as the independent variable, we can estimate σ^2 as the intercept. Throughout this paper, we take the integer part of m whenever necessary.

2·2. *Methodology*

Let $d_k = k^2/n^2$ and $s_k = \sum_{i=k+1}^{n} (y_i - y_{i-k})^2 / 2(n-k)$, where $1 \le k \le m$. As discussed in Section 2.1, we regress s_k on d_k to estimate σ^2 as the intercept. We will discuss the choice of m in Section 2.4. Since s_k is the average of $(n-k)$ lag-k differences, we assign weight $w_k = (n-k)/N$ to the observation s_k, where $N = (n-1) + (n-2) + \cdots + (n-m) = nm - m(m+1)/2$. Specifically, we fit the following linear model

$$s_k = \alpha + \beta d_k + e_k, \qquad k = 1, 2, \cdots, m, \tag{11}$$

using the weighted least square $\sum_{k=1}^m w_k(s_k - \alpha - \beta d_k)^2$.

Let $\bar{s}_w = \sum_{k=1}^m w_k s_k$ and $\bar{d}_w = \sum_{k=1}^m w_k d_k$. Then

$$\hat{\sigma}^2 = \hat{\alpha} = \bar{s}_w - \hat{\beta}\bar{d}_w, \tag{12}$$

where

$$\hat{\beta} = \frac{\sum_{k=1}^m w_k s_k (d_k - \bar{d}_w)}{\sum_{k=1}^m w_k (d_k - \bar{d}_w)^2}$$

is the estimate of the intercept β. When necessary, the dependence of $\hat{\sigma}^2$ on m, $\hat{\sigma}^2(m)$, will be expressed explicitly.

The following theorem shows that, as the GSJ estimator, $\hat{\sigma}^2$ is unbiased when g is linear. Though we derived our estimator by the least squares method, the following theorem also shows that $\hat{\sigma}^2$ can be represented as the quadratic form (2).

Theorem 1. *For the equally spaced design, we have*

(a) $\hat{\sigma}^2$ *is unbiased when* g *is a linear function regardless of the choice of* m.

(b) $\hat{\sigma}^2$ *can be written in a quadratic form* $\hat{\sigma}^2 = \mathbf{y}^T \mathbf{D} \mathbf{y} / tr(\mathbf{D})$, *where*

$$\mathbf{D} = \begin{pmatrix}
\sum_{k=1}^m b_k & -b_1 & \cdots & -b_m & & & & 0 & \\
-b_1 & \sum_{k=1}^m b_k + b_1 & -b_1 & \cdots & -b_m & & & & \\
\vdots & \ddots & \ddots & \ddots & & \ddots & & & \\
-b_m & \cdots & -b_1 & 2\sum_{k=1}^m b_k & -b_1 & \cdots & -b_m & & \\
 & \ddots & & \ddots & \ddots & \ddots & & \ddots & \\
 & & -b_m & \cdots & -b_1 & 2\sum_{k=1}^m b_k & -b_1 & \cdots & -b_m \\
 & & & \ddots & & \ddots & \ddots & \ddots & \vdots \\
 & & & & -b_m & \cdots & -b_1 & \sum_{k=1}^m b_k + b_1 & -b_1 \\
 & 0 & & & & -b_m & \cdots & -b_1 & \sum_{k=1}^m b_k
\end{pmatrix}$$

and

$$b_k = 1 - \frac{\bar{d}_w (d_k - \bar{d}_w)}{\sum_{k=1}^m w_k (d_k - \bar{d}_w)^2}.$$

Notice that $\mathbf{D}$ is a symmetric matrix with both row and column sums equal to zero. Our estimator is different from existing residual-based and difference-based estimators. Most existing difference-based estimators require the design points to be ordered with some conditions such as $max|x_i - x_{i-1}| = O(n^{-1+\delta})$, where $0 < \delta < 1/2$ for the HKT estimator and $\delta = 0$ for other estimators. It is thus difficult to extend these methods to high dimensional domains or general domains since there is no clear ordering in these situations. Furthermore, unequally spaced designs may have clusters or even tied design points, and/or large gaps between some neighboring design points. In other word, the assumption

$max|x_i - x_{i-1}| = O(n^{-1+\delta})$ may not hold. Our method can be extended naturally to general domains with unequally spaced design points (Section 4).

2·3. *Asymptotic results*

Using the fact that $\hat{\sigma}^2$ has a quadratic form, we have the following formula for the MSE (Dette et al. 1998),

$$\begin{aligned} MSE(\hat{\sigma}^2) \;\; = \;\; & \left\{(\mathbf{g}^T\mathbf{D}\mathbf{g})^2 + 4\sigma^2\mathbf{g}^T\mathbf{D}^2\mathbf{g} + 4\mathbf{g}^T(\mathbf{D}diag(\mathbf{D})\mathbf{1})\sigma^3\gamma_3 \right. \\ & \left. + \sigma^4 tr\{diag(\mathbf{D})^2\}(\gamma_4 - 3) + 2\sigma^4 tr(\mathbf{D}^2)\right\}/tr(\mathbf{D})^2, \end{aligned} \tag{13}$$

where $\mathbf{g} = (g(x_1), \cdots, g(x_n))^T$, $diag(\mathbf{D})$ denotes the diagonal matrix of the diagonal elements of $\mathbf{D}$, $\mathbf{1} = (1, \cdots, 1)^T$ and $\gamma_i = E[(\epsilon/\sigma)^i]$, $i = 3, 4$. The first term in (13) is the squared bias and the last four terms make up the variance. When the random errors are normally distributed, the second and the third terms are both equal to zero. In Appendixes B, we will show

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

$$Bias(\hat{\sigma}^2) \triangleq E(\hat{\sigma}^2 - \sigma^2) = O(\frac{m^3}{n^3}), \tag{14}$$

$$var(\hat{\sigma}^2) = \frac{1}{n}var(\epsilon^2) + \frac{9}{4nm}\sigma^4 + \frac{9m}{112n^2}var(\epsilon^2) + o(\frac{1}{nm}) + o(\frac{m}{n^2}), \tag{15}$$

$$MSE(\hat{\sigma}^2) = \frac{1}{n}var(\epsilon^2) + \frac{9}{4nm}\sigma^4 + \frac{9m}{112n^2}var(\epsilon^2) + o(\frac{1}{nm}) + o(\frac{m}{n^2}) + O(\frac{m^6}{n^6}). \tag{16}$$

Theorem 2 indicates that $\hat{\sigma}^2$ is a consistent estimator of σ^2. The asymptotical optimal bandwidth is $m_{opt} = (28n\sigma^4/var(\epsilon^2))^{1/2}$. Substituting this optimal bandwidth into (16) leads to

$$MSE(\hat{\sigma}^2(m_{opt})) = \frac{1}{n}var(\epsilon^2) + \frac{9}{28}\left(7\sigma^4 var(\epsilon^2)\right)^{1/2} n^{-3/2} + o(n^{-3/2}), \tag{17}$$

which satisfies (7).

2·4. *The choice of the bandwidth in practice*

For simplicity of notation, we assume that random errors are normally distributed with mean zero and variance σ^2. Then $var(\epsilon^2) = 2\sigma^4$ and $m_{opt} = (14n)^{1/2}$. This optimal bandwidth is obtained under the conditions that g has a bounded second derivative, $m \to \infty$ and $m/n \to 0$. Note that m_{opt} does not depend on g. However, some slightly higher order terms ignored in the MSE (16) do depend on the smoothness of the function. Therefore, the asymptotic optimal bandwidth applies for very large n only. For small to median n, we find that $m_{opt} = (14n)^{1/2}$ is too large. We now discuss two strategies for selecting m in these situations.

Note that the dominant term in (16), $var(\epsilon^2)/n$, cannot be reduced. Let

$$h(m) = \frac{9}{4nm}\sigma^4 + \frac{9m}{56n^2}\sigma^4$$

be the two higher order terms. Our first strategy is to select the smallest $m = cn^{1/2}$ such that $h(m)/h(m_{opt}) \leq 1+\lambda$, where $100\lambda\%$ is the percentage of increase in the higher order terms. It is easy to check that $m = \left(1+\lambda-(\lambda^2+2\lambda)^{1/2}\right)(14n)^{1/2}$. Note that the convergence rate of MSE remains the same. Our simulations in Section 3 indicate that $m = n^{1/2}$ with $\lambda \approx 1$ works very well. Denote $m_s = n^{1/2}$. Note that the increases of MSE are in the higher order terms. Thus, the increase of the overall MSE is usually not large. For example, $MSE(\hat{\sigma}^2(m_s))/MSE(\hat{\sigma}^2(m_{opt}))$ equal $1.099,\ 1.079,\ 1.057$ and 1.026 for $n = 30$, $n = 50$, $n = 100$ and $n = 500$ respectively. Therefore, the increases of MSEs are between 10% to 3% for these sample sizes.

Simulations in Section 3 indicate that $m_s = n^{1/2}$ is still too large when n is small and g is rough. The poor performance in these situations is usually caused by large bias. Our second strategy for selecting m is to control bias such that $Bias(\hat{\sigma}^2) = O(n^{-2})$. Consider the power form $m = cn^{\tau}$. Then from (14), $Bias(\hat{\sigma}^2) = O(n^{-3+3\tau})$. It is easy to see that the largest τ to have $Bias(\hat{\sigma}^2) = O(n^{-2})$ is $\tau = 1/3$. Therefore, another choice of m is $m_t = n^{1/3}$. Simulations in the next section indicate that m_t performs well when n is small and g is rough. For $m_t = n^{1/3}$,

$$MSE(\hat{\sigma}^2(m_t)) = \frac{2}{n}\sigma^4 + \frac{9}{4}n^{-4/3}\sigma^4 + o(n^{-4/3}), \tag{18}$$

which still satisfies (7) and has a better convergence rate than the existing difference-based estimators.

3. Simulations and comparisons with other estimators

In this section we present some simulation results based on equally spaced designs. We use the same simulation setting as in Seifert, Gasser and Wolf (1993) and Dette et al. (1998): $g(x) = 5\sin(\omega\pi x)$, where ω is the frequency of the mean function, $x_i = i/n$ and $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$. We consider three different frequencies, $\omega = 1,\ 2,\ 4$, which corresponds to low, median and high oscillations respectively. We consider three standard deviations, $\sigma = 0.5,\ 1.5,\ 4$, for different signal-to-noise ratios, and three choices of n, $n = 50$, $n = 250$ and $n = 800$, for small, median and large sample sizes. Therefore, we have 27 combinations of simulation settings.

For each simulation setting, we generate observations and compute the Rice estimator $\hat{\sigma}^2_R$, the GJS estimator $\hat{\sigma}^2_{GSJ}$, the HKT estimator $\hat{\sigma}^2_{HKT}(m)$, and our estimator $\hat{\sigma}^2(m_s)$. We repeat this process 20000 times and compute MSEs for each method. The order m in $\hat{\sigma}^2_{HKT}(m)$ acts as a tuning parameter which depends on the unknown function g. We set

$m = 2$ in our simulations and consequently,

$$\hat{\sigma}^2_{HKT}(2) = \frac{1}{n-2}\sum_{i=1}^{n-2}(0.8090y_i - 0.5y_{i+1} - 0.3090y_{i+2})^2.$$

Table 1: MSEs of various estimators and percentages of reductions.

n	ω	σ	$MSE_{\hat{\sigma}^2_R}$	$MSE_{\hat{\sigma}^2_{GSJ}}$	$MSE_{\hat{\sigma}^2_{HKT}(2)}$	$MSE_{\hat{\sigma}^2(m_s)}$	$R(\hat{\sigma}^2_R)$	$R(\hat{\sigma}^2_{GSJ})$	$R(\hat{\sigma}^2_{HKT}(2))$
50	1	0.5	0.0045	0.0051	0.0069	0.0036	20.0%	29.4%	47.8%
		1.5	0.311	0.405	0.267	0.246	20.9%	39.3%	7.9%
		4	15.774	20.745	13.429	12.372	21.6%	40.4%	7.9%
	2	0.5	0.0134	0.0051	0.0593	0.0115	14.2%	-125%	80.6%
		1.5	0.319	0.408	0.324	0.254	20.4%	37.7%	21.6%
		4	15.42	20.22	13.32	12.26	20.5%	39.4%	8.0%
	4	0.5	0.153	0.0051	0.871	0.337	-120%	-6508%	61.3%
		1.5	0.465	0.411	1.157	0.601	-29.2%	-46.2%	48.1%
		4	15.65	20.55	14.27	12.70	18.8%	38.2%	11.0%
250	1	0.5	0.00076	0.00099	0.00065	0.00054	28.9%	45.5%	16.9%
		1.5	0.0613	0.0797	0.0517	0.0451	26.4%	43.4%	12.8%
		4	3.120	4.001	2.607	2.256	27.7%	43.6%	13.5%
	2	0.5	0.00076	0.00097	0.00072	0.00057	25.0%	41.2%	20.8%
		1.5	0.0610	0.0802	0.0515	0.0453	25.7%	43.5%	12.0%
		4	3.114	4.013	2.598	2.291	26.4%	42.9%	11.8%
	4	0.5	0.00101	0.00097	0.00216	0.00118	-16.8%	-21.6%	45.4%
		1.5	0.0610	0.0801	0.0528	0.0462	24.3%	42.3%	12.5%
		4	3.092	3.953	2.581	2.270	26.6%	42.6%	12.0%
800	1	0.5	0.00023	0.00031	0.00020	0.00017	26.1%	45.2%	15.0%
		1.5	0.0190	0.0248	0.0157	0.0133	30.0%	46.4%	15.3%
		4	0.969	1.267	0.806	0.673	30.5%	46.9%	16.5%
	2	0.5	0.00024	0.00030	0.00020	0.00016	33.3%	46.7%	20.0%
		1.5	0.0189	0.0248	0.0159	0.0134	29.1%	46.0%	15.7%
		4	0.967	1.257	0.802	0.675	30.2%	46.3%	15.8%
	4	0.5	0.00024	0.00030	0.00021	0.00018	25.0%	40.0%	14.3%
		1.5	0.0188	0.0249	0.0158	0.0133	29.3%	46.6%	15.8%
		4	0.964	1.255	0.797	0.671	30.4%	46.5%	15.8%

Table 1 lists MSEs for all methods under all simulation settings. We define $R(\hat{\sigma}^2_R) = 1 - MSE_{\hat{\sigma}^2(m_s)}/MSE_{\hat{\sigma}^2_R}$ to measure the reduction in MSE of our estimator over the Rice estimator. We define $R(\hat{\sigma}^2_{GSJ})$ and $R(\hat{\sigma}^2_{HKT}(2))$ similarly. A positive R represents a reduction in MSE while a negative R represents an increase. These reductions in percentages are also listed in Table 1. In general, $MSE_{\hat{\sigma}^2(m_s)} < MSE_{\hat{\sigma}^2_{HKT}(2)} < MSE_{\hat{\sigma}^2_R} < MSE_{\hat{\sigma}^2_{GSJ}}$ for most cases, especially when $n = 250$ or $n = 800$. To visualize the comparative results, we plot MSEs versus sample sizes in Figure 1. We conclude that $\hat{\sigma}^2(m_s)$ has smaller MSEs than $\hat{\sigma}^2_{HKT}(2)$ in all situations. The comparative performance of $\hat{\sigma}^2(m_s)$ depends on the smoothness of g, the sample size and the signal-to-noise ratio. $\hat{\sigma}^2(m_s)$ has smaller MSEs except for four cases, $(n, \omega, \sigma) = (50, 2, 0.5)$, $(50, 4, 0.5)$, $(50, 4, 1.5)$ or $(250, 4, 0.5)$, where g is rough, sample size is small and standard deviation is small. Same comparative results on three existing methods have been reached in Seifert et al. (1993) and Dette et al. (1998): the HKT estimator performs better when g is flat and/or n is large, while the GSJ estimator performs better when the opposite is true.

To take a closer look at why $\hat{\sigma}^2(m_s)$ fails when g is rough and n is small, we list squared biases of $\hat{\sigma}^2(m_s)$, $\hat{\sigma}^2_R$, $\hat{\sigma}^2_{GSJ}$ and $\hat{\sigma}^2_{HKT}(2)$ in Table 2 with $n = 50$. It is clear that MSEs of $\hat{\sigma}^2(m_s)$ and $\hat{\sigma}^2_{HKT}(2)$ are dominated by biases when g is rough and n is small. The GSJ estimator has much smaller biases, thus much smaller MSEs in these situations. As discussed in Section 2.4, the approximate optimal rate of m, $n^{1/2}$, requires a large n or smooth g. When g is rough and n is small, $m_s = n^{1/2}$ is too large which leads to large biases. One option is to control bias using $m_t = n^{1/3}$, as discussed in Section 2.4. Table 2 also lists squared biases, variances and MSEs of $\hat{\sigma}^2(m_t)$. As expected, $\hat{\sigma}^2(m_t)$ reduces the bias with small increase in the variance. Though the performance of $\hat{\sigma}^2(m_t)$ is a little worse than $\hat{\sigma}^2(m_s)$ for other cases, it performs well when $\hat{\sigma}^2(m_s)$ fails. $\hat{\sigma}^2(m_t)$ has smaller MSEs than $\hat{\sigma}^2_{GSJ}$ for all cases except one case when $\omega = 4$ and $\sigma = 0.5$. Therefore, we recommend $\hat{\sigma}^2(m_t)$ when sample size is small, and g is rough or little information about g is available.

For the equally spaced design, it is clear that $\hat{\sigma}^2(1) = \hat{\sigma}^2_R$. That is, the Rice estimator is a special case of our estimator with $m = 1$. One interesting observation from simulations is that $\hat{\sigma}^2(2) \approx \hat{\sigma}^2_{GSJ}$ when σ^2 is not very small. In theory it is easy to show that the dominant term of $MSE(\hat{\sigma}^2(2))$ is $35\sigma^4/9n$, which is exactly the same as that of $\hat{\sigma}^2_{GSJ}$. The simulated MSEs of $\hat{\sigma}^2(2)$ are list in the last column of Table 2 which is almost the same as those of $\hat{\sigma}^2_{GSJ}$. We have performed many more simulations with different mean functions, signal-to-noise ratios and sample sizes. Comparative results remain the same.

4. Extension to the general domain

In this section we extend our method to a general domain $\mathcal{T}$, where $\mathcal{T}$ is an arbitrary subset of a normed space. Let $d_{ij} = \|x_i - x_j\|^2$ and $s_{ij} = \frac{1}{2}(y_i - y_j)^2$ for all pairs i and j, where $1 \le i < j \le n$. We fit the following simple linear model

$$s_{ij} = \alpha + \beta d_{ij} + e_{ij}, \qquad d_{ij} \le M, \tag{19}$$

using the least squares where $M > 0$ is the bandwidth. The estimate of σ^2 is $\hat{\sigma}^2 = \hat{\alpha}$. For $\mathcal{T} = [0, 1]$, $x_i = i/n$ and $M = (m/n)^2$, $\hat{\sigma}^2$ reduces to the weighted least squares estimate proposed in Section 2.2.

We now discuss how to choose the bandwidth M. For unequally spaced designs on $\mathcal{T} = [0, 1]$, we may choose $M = (m/n)^2$ as in Section 2.4. This leads to two choices of M:

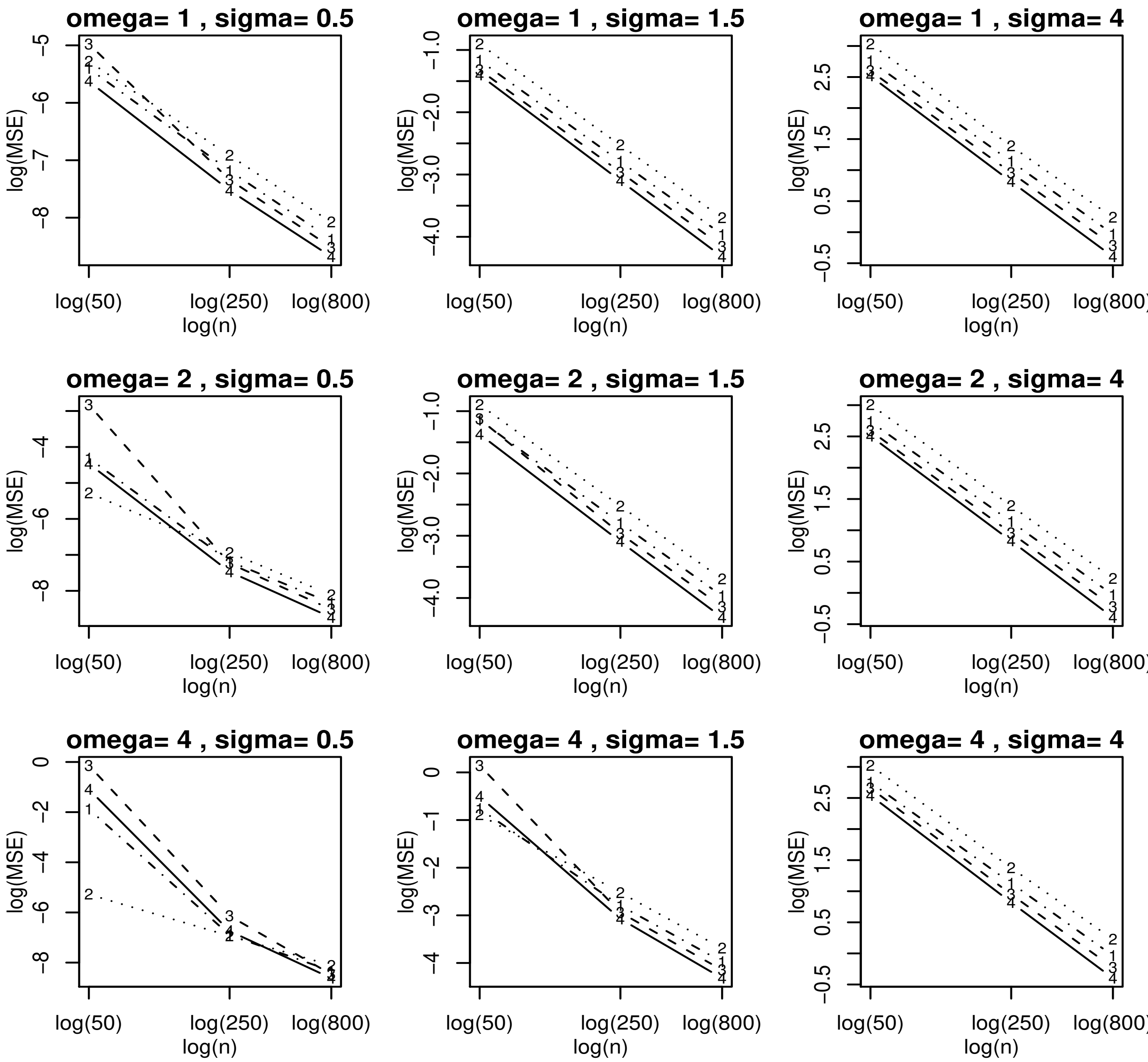

Figure 1: Plot of $\log(MSE)$ versus $\log(n)$. Simulation settings are marked above each plot. Four lines marked as "1", "2", "3" and "4" in each plot correspond to Rice, GSJ, HKT and our estimators respectively.

Table 2: MSEs, Squared Biases (BSQ) and Variances (VAR). $n = 50$.

ω	σ		$\hat{\sigma}^2_R$	$\hat{\sigma}^2_{GSJ}$	$\hat{\sigma}^2_{HKT}(2)$	$\hat{\sigma}^2(m_s)$	$\hat{\sigma}^2(m_t)$	$\hat{\sigma}^2(2)$
1	0.5	MSE	0.0045	0.0051	0.0069	0.0036	0.0038	0.0050
		BSQ	0.0006	0.0000	0.0035	0.0004	0.0000	0.0000
		VAR	0.0039	0.0051	0.0034	0.0032	0.0038	0.0050
	1.5	MSE	0.311	0.405	0.267	0.246	0.307	0.402
		BSQ	0.001	0.000	0.004	0.000	0.000	0.000
		VAR	0.311	0.405	0.264	0.246	0.307	0.402
	4	MSE	15.774	20.745	13.429	12.372	15.42	20.231
		BSQ	0.001	0.000	0.007	0.000	0.00	0.000
		VAR	15.773	20.745	13.422	12.372	15.42	20.231
2	0.5	MSE	0.0134	0.0051	0.0593	0.0115	0.0040	0.0050
		BSQ	0.0095	0.0000	0.0557	0.0078	0.0001	0.0000
		VAR	0.0039	0.0051	0.0036	0.0037	0.0039	0.0050
	1.5	MSE	0.319	0.408	0.324	0.254	0.310	0.401
		BSQ	0.010	0.000	0.058	0.008	0.000	0.000
		VAR	0.309	0.408	0.266	0.246	0.310	0.401
	4	MSE	15.42	20.22	13.32	12.26	15.72	20.24
		BSQ	0.01	0.00	0.00	0.01	0.00	0.00
		VAR	15.41	20.22	13.26	12.25	15.72	20.24
4	0.5	MSE	0.153	0.0051	0.871	0.337	0.0070	0.0055
		BSQ	0.149	0.0001	0.865	0.331	0.0027	0.0003
		VAR	0.004	0.0050	0.006	0.006	0.0043	0.0052
	1.5	MSE	0.465	0.411	1.157	0.601	0.316	0.405
		BSQ	0.150	0.000	0.869	0.332	0.003	0.000
		VAR	0.315	0.411	0.288	0.269	0.313	0.405
	4	MSE	15.65	20.55	14.27	12.70	15.62	20.40
		BSQ	0.15	0.00	0.89	0.34	0.00	0.00
		VAR	15.50	20.55	13.38	12.36	15.62	20.40

$M_s = (m_s/n)^2 = n^{-1}$ and $M_t = (m_t/n)^2 = n^{-4/3}$. Note that these two choices of M applies to $\mathcal{T} = [0, 1]$ only. For a general domain, similar to the idea of nearest neighbor estimators, we can select M such that the number of pairs involved in the linear regression (19) equals to N, where $N = nm - m(m+1)/2$ is the number of pairs involved in the regression (11). For $m = m_s$ and $m = m_t$, we denote the resulting M as M_S and M_T respectively.

We now conduct a small scale simulation to evaluate performance of our method and compare it with existing methods. We use the same settings as in Section 3 for the mean function g and error standard deviation σ. For $n = 50$, we generate design points from the density function $0.9f_1 + 0.1f_2$, where f_1 and f_2 are density functions of uniform random variables on $[0.2, 0.3]$ and $[0, 1]$ respectively. For each generated design points, we generate observations and compute $\hat{\sigma}^2_R$, $\hat{\sigma}^2_{GSJ}$, $\hat{\sigma}^2_{HKT}(2)$ and our estimators. We repeat this process 20000 times and compute MSEs for all methods. Table 3 shows that $\hat{\sigma}^2(M_s)$ and $\hat{\sigma}^2(M_S)$ have the similar performance when g is flat. $\hat{\sigma}^2(M_S)$ trends to performs better when g becomes rougher. $\hat{\sigma}^2(M_T)$ and $\hat{\sigma}^2(M_t)$ have similar performance for all cases. $\hat{\sigma}^2(M_S)$ and $\hat{\sigma}^2(M_t)$ always have smaller MSEs than the three existing methods, especially when ω is large and/or σ is small. One possible explanation is that the design points are clustered. Our method uses design points which are actually close to each other rather than consecutive design points.

Table 3: MSEs of various estimators.

ω	σ	$MSE_{\hat{\sigma}^2_R}$	$MSE_{\hat{\sigma}^2_{GSJ}}$	$MSE_{\hat{\sigma}^2_{HKT}(2)}$	$MSE_{\hat{\sigma}^2(M_S)}$	$MSE_{\hat{\sigma}^2(M_s)}$	$MSE_{\hat{\sigma}^2(M_T)}$	$MSE_{\hat{\sigma}^2(M_t)}$
1	0.5	0.0144	0.0055	0.0147	0.0032	0.0044	0.0036	0.0032
	1.5	0.329	0.413	0.285	0.254	0.265	0.289	0.253
	4	15.64	20.82	13.35	12.89	12.98	14.61	12.74
2	0.5	0.378	0.0165	0.724	0.0031	0.0082	0.0036	0.0031
	1.5	0.759	0.434	1.072	0.254	0.270	0.291	0.255
	4	16.35	20.57	14.47	12.42	12.48	14.12	12.26
4	0.5	0.720	0.201	0.737	0.036	0.647	0.0048	0.0056
	1.5	1.150	0.654	1.102	0.287	1.065	0.282	0.265
	4	16.95	21.08	14.53	12.51	14.47	14.13	12.43

5. An application

We now apply our method to the lake acidity data derived by Douglas and Delampady (1990) from the Eastern Lakes Survey of 1984. It contains measurements of 1789 lakes in three eastern US regions: northeast, upper Midwest and southeast. As in Gu and Wahba (1993) and Wang and Ke (2002), we use a subset of 112 lakes in the southern Blue Ridge mountain areas.

Of interest is the dependence of the water pH level (y) on the calcium concentration in $\log_{10}$ milligrams per liter (t_1) and the geographical location ($\mathbf{t}_2 = (t_{21}, t_{22})$ where

t_{21}=latitude and t_{22}=longitude). For the purpose of illustration, we consider nonparametric regression model (1) with three different cases of x: $x = t_1$, $x = \mathbf{t}_2$ and $x = (t_1, \mathbf{t}_2)$. These three cases correspond to three different domains of one, two and three dimensions respectively. For the first two cases, we use simple Euclidean norms. For the third case, we transform t_1 and $\mathbf{t}_2$ to the same scale before estimating the variance. Estimates are listed in Table 4. For comparison, we fit a cubic spline model, a thin-plate spline model and an smoothing spline ANOVA model for three cases of x respectively using the *ssr* function in the *ASSIST* package (Wang and Ke 2002). We use the generalized cross-validation (GCV) and generalized maximum likelihood method (GML) to estimate the smoothing parameters (Wahba 1990, Gu 2002, Wang and Ke 2002). The resulting residual-based estimates of σ^2 are also listed in Table 4.

Table 4: Estimated variances.

	$\hat{\sigma}^2(M_S)$	$\hat{\sigma}^2(M_T)$	$\hat{\sigma}^2_{GCV}$	$\hat{\sigma}^2_{GML}$
t_1	0.0821	0.0708	0.0791	0.0848
$\mathbf{t}_2$	0.0884	0.0899	0.0912	0.0917
$(t_1, \mathbf{t}_2)$	0.0544	0.0666	0.0655	0.0656

For the third case, there is no clear definition of the distance. For example, we may define $\|x_i - x_j\|^2 = \gamma(t_{1i} - t_{1j})^2 + (t_{21i} - t_{21j})^2 + (t_{22i} - t_{22j})^2$, where γ is a scale parameter. It is clear that γ also acts as a tuning parameter. Our simulations (not shown here) indicate that the estimate depends on both the bandwidth M and the scale parameter γ. Future research is required on the choices of these two parameters.

6. Conclusion

In this article we propose a new method for estimating the error variance in nonparametric regression. We show, in both theory and simulations, that our method compares favorably with some of the existing methods. The biggest advantage of our method is its generality: it applies to general domains such as Euclidean d-space, circles and spheres, where no method exists in the literature. Our method does not require dense design points in the whole domain, thus avoiding the curse of dimensionality problem in high dimensional space and allow potential gaps between design points.

The theoretical optimal bandwidth is usually too large for finite sample sizes. A good choice of the bandwidth depends on many factors such as the mean function, signal-to-noise ratio and the sample size. Our simulations indicate that two simple choices, $m_s = n^{1/2}$ and $m_t = n^{1/3}$, work well in practice. More research is required on the choice of the bandwidth, especially for general domains with unequally spaced design points.

Acknowledgement

This research was supported by NIH Grant R01 GM58533.

Appendix 1

Proof of Theorem 1

(a) Suppose that $g(x) = \mu + \delta x$ and denote $g_i = g(x_i)$. Then

$$\begin{aligned} Es_k &= \sigma^2 + \frac{1}{2(n-k)} \sum_{i=k+1}^{n} (g_i - g_{i-k})^2 = \sigma^2 + \frac{1}{2(n-k)} \sum_{i=k+1}^{n} \delta^2 (x_i - x_{i-k})^2 \\ &= \sigma^2 + \frac{\delta^2}{2(n-k)} \sum_{i=k+1}^{n} \frac{k^2}{n^2} = \sigma^2 + \frac{k^2}{2n^2}\delta^2, \qquad k = 1, 2, \cdots, m, \end{aligned}$$

and

$$\begin{aligned} E(\bar{s}_w) &= \frac{1}{N} \sum_{k=1}^{m} (n-k) Es_k = \frac{1}{N} \sum_{k=1}^{m} (n-k) \left(\sigma^2 + \frac{k^2}{2n^2}\delta^2 \right) \\ &= \sigma^2 + \frac{\delta^2}{2Nn^2} \sum_{k=1}^{m} k^2 (n-k) = \sigma^2 + \frac{1}{2}\delta^2 \bar{d}_w. \end{aligned}$$

Let

$$I_t = \sum_{k=1}^{m} k^t, \qquad t = 1, 2, \cdots. \tag{20}$$

We have

$$\begin{aligned} \sum_{k=1}^{m} w_k (d_k - \bar{d}_w) Es_k &= \sum_{k=1}^{m} w_k d_k Es_k - \bar{d}_w E(\bar{s}_w) \\ &= \frac{1}{Nn^2} \sum_{k=1}^{m} k^2 (n-k) \left(\sigma^2 + \frac{k^2}{2n^2}\delta^2 \right) - \bar{d}_w \left(\sigma^2 + \frac{1}{2}\delta^2 \bar{d}_w \right) \\ &= \frac{\delta^2}{2Nn^4} \sum_{k=1}^{m} k^4 (n-k) - \frac{1}{2}\delta^2 \bar{d}_w^2 \\ &= \frac{1}{2}\delta^2 \left(\frac{I_4}{Nn^3} - \frac{I_5}{Nn^4} - \bar{d}_w^2 \right), \end{aligned}$$

and

$$\sum_{k=1}^{m} w_k (d_k - \bar{d}_w)^2 = \sum_{k=1}^{m} w_k d_k^2 - \bar{d}_w^2 = \frac{I_4}{Nn^3} - \frac{I_5}{Nn^4} - \bar{d}_w^2.$$

Finally,

$$\begin{aligned}
E(\hat{\sigma}^2) &= E(\bar{s}_w) - E(\hat{\beta}\bar{d}_w) \\
&= E(\bar{s}_w) - \frac{\bar{d}_w}{\sum_{k=1}^m w_k(d_k-\bar{d}_w)^2}\sum_{k=1}^m w_k(d_k-\bar{d}_w)Es_k \\
&= \sigma^2 + \frac{1}{2}\delta^2\bar{d}_w - \frac{\bar{d}_w}{\frac{I_4}{Nn^3}-\frac{I_5}{Nn^4}-\bar{d}_w^2}\frac{1}{2}\delta^2\left(\frac{I_4}{Nn^3}-\frac{I_5}{Nn^4}-\bar{d}_w^2\right) \\
&= \sigma^2.
\end{aligned} \tag{21}$$

(b) It is straightforward to check that

$$\hat{\sigma}^2 = \sum_{k=1}^m b_k w_k s_k = \frac{1}{2N}\sum_{k=1}^m\left(b_k\sum_{i=k+1}^n (y_i-y_{i-k})^2\right) = \frac{1}{2N}\mathbf{y}^T\mathbf{D}\mathbf{y},$$

where the last equality can be checked directly by expanding both sides and comparing corresponding terms. Thus to prove that $\hat{\sigma}^2 = \mathbf{y}^T\mathbf{D}\mathbf{y}/tr(\mathbf{D})$, we only need to show that $tr(\mathbf{D}) = 2N$. Note that $\mathbf{D}$ does not depend on g. Setting $g \equiv 0$, we have

$$E(\hat{\sigma}^2) = \frac{1}{2N}E\left(\mathbf{y}^T\mathbf{D}\mathbf{y}\right) = \frac{1}{2N}E\left(\boldsymbol{\epsilon}^T\mathbf{D}\boldsymbol{\epsilon}\right) = \frac{\sigma^2}{2N}tr(\mathbf{D}),$$

where $\boldsymbol{\epsilon} = (\epsilon_1,\cdots,\epsilon_n)^T$. Now since $\hat{\sigma}^2$ is unbiased for any linear function g, we have $tr(\mathbf{D}) = 2N$.

Appendix 2

Proof of Theorem 2

1. *Proof of* (14)

Instead of using the formula $Bias(\hat{\sigma}^2) = \mathbf{g^T D g}/tr(\mathbf{D})$, we calculate this quantity directly from (12) which gives a more accurate approximation. Note that $N = nm - m(m+1)/2$. Similar to Appendix A, it is easy to check that

$$\bar{d}_w = \frac{I_2}{Nn} - \frac{I_3}{Nn^2} = \frac{m^2}{3n^2} + o(\frac{m^2}{n^2}), \tag{22}$$

$$\sum_{k=1}^m w_k(d_k-\bar{d}_w)^2 = \frac{I_4}{Nn^3} - \frac{I_5}{Nn^4} - \left(\frac{I_2}{Nn}-\frac{I_3}{Nn^2}\right)^2 = \frac{4m^4}{45n^4} + o(\frac{m^4}{n^4}). \tag{23}$$

Thus

$$\eta \triangleq \frac{\bar{d}_w}{\sum_{k=1}^m w_k(d_k-\bar{d}_w)^2} = \frac{\frac{m^2}{3n^2}+o(\frac{m^2}{n^2})}{\frac{4m^4}{45n^4}+o(\frac{m^4}{n^4})} = \frac{15n^2}{4m^2} + o(\frac{n^2}{m^2}). \tag{24}$$

Let $g_i = g(x_i)$, $g_i^{'} = g^{'}(x_i)$ and $g_i^{''} = g^{''}(x_i)$, $i = 1, \cdots, n$. Then

$$\begin{aligned}
Es_k &= \sigma^2 + \frac{1}{2(n-k)} \sum_{i=k+1}^{n} (g_i - g_{i-k})^2 \\
&= \sigma^2 + \frac{1}{2(n-k)} \sum_{i=k+1}^{n} \left(\frac{k}{n} g_i^{'} + O(\frac{k^2}{n^2}) \right)^2 \\
&= \sigma^2 + \frac{1}{2(n-k)} \sum_{i=k+1}^{n} \left(\frac{k^2}{n^2} (g_i^{'})^2 + O(\frac{k^3}{n^3}) \right) \\
&= \sigma^2 + \frac{k^2}{2n^2} \frac{n}{n-k} \left(\frac{1}{n} \sum_{i=1}^{n} (g_i^{'})^2 - \frac{1}{n} \sum_{i=1}^{k} (g_i^{'})^2 \right) + O(\frac{k^3}{n^3}) \\
&= \sigma^2 + \frac{k^2}{2n^2} \frac{n}{n-k} \left(\int_0^1 (g^{'}(x))^2 dx + O(\frac{k}{n}) \right) + O(\frac{k^3}{n^3}) \\
&= \sigma^2 + \frac{k^2}{n^2} J + O(\frac{k^3}{n^3}),
\end{aligned}$$

where $J = \frac{1}{2} \int_0^1 (g^{'}(x))^2 dx$. Consequently,

$$\begin{aligned}
E(\bar{s}_w) &= \frac{1}{N} \sum_{k=1}^{m} (n-k)(\sigma^2 + \frac{k^2}{n^2} J + O(\frac{k^3}{n^3})) \\
&= \sigma^2 + \frac{J}{Nn^2} \sum_{k=1}^{m} (n-k) k^2 + O(\frac{m^3}{n^3}) \\
&= \sigma^2 + \frac{I_2}{Nn} J - \frac{I_3}{Nn^2} J + O(\frac{m^3}{n^3}), \qquad (25)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{k=1}^{m} w_k (d_k - \bar{d}_w) Es_k &= \frac{1}{Nn^2} \sum_{k=1}^{m} k^2 (n-k) \left(\sigma^2 + \frac{k^2}{n^2} J + O(\frac{k^3}{n^3}) \right) - \bar{d}_w E(\bar{s}_w) \\
&= \left(\frac{I_2}{Nn} - \frac{I_3}{Nn^2} \right) \sigma^2 + \frac{J}{Nn^4} (nI_4 - I_5) + O(\frac{m^5}{n^5}) \\
&\quad - \left(\frac{I_2}{Nn} - \frac{I_3}{Nn^2} \right) \left(\sigma^2 + \frac{I_2}{Nn} J - \frac{I_3}{Nn^2} J + O(\frac{m^3}{n^3}) \right) \\
&= J \left\{ \frac{I_4}{Nn^3} - \frac{I_5}{Nn^4} - \left(\frac{I_2}{Nn} - \frac{I_3}{Nn^2} \right)^2 \right\} + O(\frac{m^5}{n^5}). \qquad (26)
\end{aligned}$$

Plugging (22)–(26) into (21) gives

$$\begin{aligned}
E(\hat{\sigma}^2) &= E(\bar{s}_w) - \frac{\frac{I_2}{Nn} - \frac{I_3}{Nn^2}}{\frac{I_4}{Nn^3} - \frac{I_5}{Nn^4} - \left(\frac{I_2}{Nn} - \frac{I_3}{Nn^2}\right)^2} \left[J \left\{ \frac{I_4}{Nn^3} - \frac{I_5}{Nn^4} - \left(\frac{I_2}{Nn} - \frac{I_3}{Nn^2} \right)^2 \right\} + O(\frac{m^5}{n^5}) \right] \\
&= \left(\sigma^2 + \frac{I_2}{Nn} J - \frac{I_3}{Nn^2} J + O(\frac{m^3}{n^3}) \right) - \left\{ \frac{I_2}{Nn} J - \frac{I_3}{Nn^2} J + \left(\frac{15n^2}{4m^2} + o(\frac{n^2}{m^2}) \right) O(\frac{m^5}{n^5}) \right\} \\
&= \sigma^2 + O(\frac{m^3}{n^3}).
\end{aligned}$$

2. *Proof of* (15)

We prove the following two lemmas first.

Lemma 1. *Assume that $m \to \infty$ and $m/n \to 0$. Then*

(a) $\sum_{k=1}^m b_k = m - \frac{5m^2}{16n} + o(\frac{m^2}{n})$.

(b) $\sum_{k=1}^{l-1} b_k = \frac{9}{4} l - \frac{5l^3}{4m^2} + o(l) + O(1), \quad 1 \le l \le m$.

(c) $\sum_{k=1}^m b_k^2 = \frac{9}{4} m + o(m)$.

(d) $\sum_{k=l}^m k b_k = O(m^2), \quad 1 \le l \le m$.

(e) $\sum_{k=1}^{l-1} k^2 b_k = O(l^3), \quad 1 \le l \le m$.

(f) $\sum_{k=1}^m k^2 b_k = o(m^3)$.

Proof. (*a*)

$$\begin{aligned}
\sum_{k=1}^m (d_k - \bar{d}_w) &= \frac{1}{n^2} \sum_{k=1}^m k^2 - \frac{m}{Nn^2} \sum_{k=1}^m (nk^2 - k^3) \\
&= \frac{m}{Nn^2} \sum_{k=1}^m k^3 + \frac{1}{n^2} \left(1 - \frac{mn}{N} \right) \sum_{k=1}^m k^2 \\
&= \frac{m^4}{4n^3} - \frac{m^4}{6n^3} + o(\frac{m^4}{n^3}) = \frac{m^4}{12n^3} + o(\frac{m^4}{n^3}).
\end{aligned}$$

Thus using (24),

$$\begin{aligned}
\sum_{k=1}^m b_k &= \sum_{k=1}^m \left(1 - \eta(d_k - \bar{d}_w) \right) = m - \eta \sum_{k=1}^m \left(d_k - \bar{d}_w \right) \\
&= m - \left(\frac{15n^2}{4m^2} + o(\frac{n^2}{m^2}) \right) \left(\frac{m^4}{12n^3} + o(\frac{m^4}{n^3}) \right) = m - \frac{5m^2}{16n} + o(\frac{m^2}{n}).
\end{aligned}$$

(b) It is easy to check that $\eta \bar{d}_w = 5/4 + o(1)$. Thus for $1 \le l \le m$,

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

(c) It is easy to check that $\sum_{k=1}^m (d_k - \bar{d}_w)^2 = \frac{4m^5}{45n^4} + o(\frac{m^5}{n^4})$. Then

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

(d) For $1 \le l \le m$,

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

(e) For $1 \le l \le m$,

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

(f) Similar to part (e), we have

$$\sum_{k=1}^{m} k^2 b_k = \left(1 + \eta\bar{d}_w\right) \sum_{k=1}^{m} k^2 - \eta \sum_{k=1}^{m} k^2 d_k = \frac{3}{4}m^3 - \frac{3}{4}m^3 + o(m^3) = o(m^3).$$

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

(a) $\mathbf{g}^T \mathbf{D}^2 \mathbf{g} = O(\frac{m^5}{n^2})$.

(b) $tr(\mathbf{D^2}) = 4nm^2 - \frac{103}{28}m^3 + \frac{9}{2}nm + o(m^3) + o(nm)$.

(c) $\mathbf{g^T}(\mathbf{D}diag(\mathbf{D})\mathbf{1}) = O(\frac{m^4}{n})$.

(d) $tr\{diag(\mathbf{D})^2\} = 4nm^2 - \frac{103}{28}m^3 + o(m^3)$.

Proof. (a) Since $\mathbf{D}$ is symmetric,

$$\mathbf{g^T D^2 g} = \mathbf{g^T D^T D g} = (\mathbf{Dg})^T\mathbf{Dg} \triangleq \mathbf{p}^T\mathbf{p}, \tag{27}$$

where $\mathbf{p} = \mathbf{Dg} = (p_1, p_2, \cdots, p_n)^T$. For $i \in [m+1, n-m]$, using Lemma 1(f), we have

$$\begin{aligned} p_i &= 2g_i\sum_{k=1}^{m} b_k - \sum_{k=1}^{m} b_k g_{i-k} - \sum_{k=1}^{m} b_k g_{i+k} \\ &= \sum_{k=1}^{m} b_k(g_i - g_{i-k}) - \sum_{k=1}^{m} b_k(g_{i+k} - g_i) \\ &= \sum_{k=1}^{m} b_k\left(\frac{k}{n}g_i' - \frac{k^2}{2n^2}g_i^{''} + o(\frac{k^2}{n^2})\right) - \sum_{k=1}^{m} b_k\left(\frac{k}{n}g_i' + \frac{k^2}{2n^2}g_i^{''} + o(\frac{k^2}{n^2})\right) \\ &= -\frac{1}{n^2}g_i^{''}\sum_{k=1}^{m} k^2 b_k + o(\frac{m^3}{n^2}) = o(\frac{m^3}{n^2}), \qquad\qquad m+1 \le i \le n-m. \end{aligned}$$

For $i \in [1, m]$, using Lemma 1(d), (e) and (f), we have

$$\begin{aligned} p_i &= \sum_{k=1}^{i-1} b_k(g_i - g_{i-k}) - \sum_{k=1}^{m} b_k(g_{i+k} - g_i) \\ &= \sum_{k=1}^{i-1} b_k\left(\frac{k}{n}g_i' - \frac{k^2}{2n^2}g_i^{''} + o(\frac{k^2}{n^2})\right) - \sum_{k=1}^{m} b_k\left(\frac{k}{n}g_i' + \frac{k^2}{2n^2}g_i^{''} + o(\frac{k^2}{n^2})\right) \\ &= -\frac{g_i^{'}}{n}\sum_{k=i}^{m} k b_k - \left(\sum_{k=1}^{m} b_k\frac{k^2}{2n^2}g_i^{''} + \sum_{k=1}^{i-1} b_k\frac{k^2}{2n^2}g_i^{''}\right) + o(\frac{m^3}{n^2}) \\ &= O(\frac{m^2}{n}) + O(\frac{i^3}{n^2}) + o(\frac{m^3}{n^2}) = O(\frac{m^2}{n}), \qquad\qquad 1 \le i \le m. \end{aligned}$$

Similar arguments show that $p_i = O(\frac{m^2}{n})$ for $i \in [n-m+1, n]$. Consequently,

$$\mathbf{g}^T\mathbf{D}^2\mathbf{g} = \mathbf{p}^T\mathbf{p} = \sum_{i=1}^{m} p_i^2 + \sum_{i=m+1}^{n-m} p_i^2 + \sum_{i=n-m+1}^{n} p_i^2 = O(\frac{m^5}{n^2}).$$

(b) Using Lemma 1, we have

$$\begin{aligned}
tr(\mathbf{D}^2) &= (n-2m)\left\{\left(2\sum_{k=1}^m b_k\right)^2 + 2\sum_{k=1}^m b_k^2\right\} + 2\sum_{l=1}^m \left\{\left(\sum_{k=1}^m b_k + \sum_{k=1}^{l-1} b_k\right)^2 + \sum_{k=1}^m b_k^2 + \sum_{k=1}^{l-1} b_k^2\right\} \\
&= (n-2m)\left\{4\left(m - \frac{5m^2}{16n} + o(\frac{m^2}{n})\right)^2 + 2\left(\frac{9}{4}m + o(m)\right)\right\} \\
&\quad +2\sum_{l=1}^m \left(m - \frac{5m^2}{16n} + o(\frac{m^2}{n}) + \frac{9}{4}l - \frac{5l^3}{4m^2} + o(l) + O(1)\right)^2 + O(m^2) \\
&= 4nm^2 - \frac{21}{2}m^3 + \frac{9}{2}nm + o(m^3) + o(nm) + \frac{191}{28}m^3 + o(m^3) \\
&= 4nm^2 - \frac{103}{28}m^3 + \frac{9}{2}nm + o(m^3) + o(nm).
\end{aligned}$$

(c) Using Lemma $1(a)$, (b) and Lemma $2(a)$, we have

$$\begin{aligned}
\mathbf{g^T}(\mathbf{D}diag(\mathbf{D})\mathbf{1}) &= (\mathbf{Dg})^T \cdot diag(\mathbf{D})\mathbf{1} = \mathbf{p}^T \cdot diag(\mathbf{D})\mathbf{1} \\
&= \sum_{l=1}^m p_l \left(\sum_{k=1}^m b_k + \sum_{k=1}^{l-1} b_k\right) + \sum_{l=m+1}^{n-m} p_l \left(2\sum_{k=1}^m b_k\right) + \sum_{l=n-m+1}^{n} p_l \left(\sum_{k=1}^m b_k + \sum_{k=1}^{n-l} b_k\right) \\
&= \sum_{l=1}^m p_l \cdot O(m) + \sum_{l=m+1}^{n-m} p_l \cdot O(m) + \sum_{l=n-m+1}^{n} p_l \cdot O(m) \\
&= O(\frac{m^4}{n}) + o(\frac{m^4}{n}) + O(\frac{m^4}{n}) = O(\frac{m^4}{n}).
\end{aligned}$$

(d) Using Lemma $1(a)$ and (b), we have

$$\begin{aligned}
tr\{diag(\mathbf{D})^2\} &= 2\sum_{l=1}^m \left(\sum_{k=1}^m b_k + \sum_{k=1}^{l-1} b_k\right)^2 + \sum_{l=m+1}^{n-m} \left(2\sum_{k=1}^m b_k\right)^2 \\
&= 2\sum_{l=1}^m \left(m + \frac{9}{4}l - \frac{5l^3}{4m^2} + o(m)\right)^2 + 4\sum_{l=m+1}^{n-m} \left(m - \frac{5m^2}{16n} + o(\frac{m^2}{n})\right)^2 \\
&= 4nm^2 - \frac{103}{28}m^3 + o(m^3).
\end{aligned}$$

Now we are ready to prove (15). As mentioned in Section 2.3, the last four terms in (13) make up the variance. Using Lemma 1, Lemma 2 and the fact that $\sigma^4(\gamma_4 - 3) = var(\epsilon^2) - 2\sigma^4$, we have

$$\begin{aligned}
var(\hat{\sigma}^2) &= \left\{4\sigma^2\mathbf{g^T D^2 g} + 4\mathbf{g^T}(\mathbf{D}diag(\mathbf{D})\mathbf{1})\sigma^3\gamma_3 + \sigma^4 tr\{diag(\mathbf{D})^2\}(\gamma_4 - 3) + 2\sigma^4 tr(\mathbf{D^2})\right\}/tr(\mathbf{D})^2 \\
&= \frac{1}{4N^2}\left\{O(\frac{m^5}{n^2}) + O(\frac{m^4}{n}) + (var(\epsilon^2) - 2\sigma^4)\left(4nm^2 - \frac{103}{28}m^3 + o(m^3)\right)\right. \\
&\qquad \left. + 2\sigma^4\left(4nm^2 - \frac{103}{28}m^3 + \frac{9}{2}nm + o(m^3) + o(nm)\right)\right\} \\
&= \frac{1}{4N^2}\left\{var(\epsilon^2)\left(4nm^2 - \frac{103}{28}m^3\right) + 9nm\sigma^4 + o(m^3) + o(nm)\right\} \\
&= \frac{1}{n}var(\epsilon^2) + \frac{9}{4nm}\sigma^4 + \frac{9m}{112n^2}var(\epsilon^2) + o(\frac{1}{nm}) + o(\frac{m}{n^2}).
\end{aligned}$$

3. *Proof of* (16)

The proof of (16) can be obtained immediately from (14) and (15).

References

Buckley, M. J., Eagleson, G. K. and Silverman, B. W. (1988). The estimation of residual variance in non-parametric regression, *Biometrika* **75**: 189–199.

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

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

Carter, C. K. and Eagleson, G. K. (1992). A comparison of variance estimations in nonparametric regression, *Journal of the Royal Statistical Society B* **54**: 773–780.

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 B* **60**: 751–764.

Douglas, A. and Delampady, M. (1990). Eastern lake survey - phase I: documentation for the data base and the derived data sets, *SIMS Technical Report 160*. Department of Statistics, University of British Columbia, Vancouver.

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

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

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

Gu, C. (2002). *Smoothing Spline ANOVA Models*, Springer-Verlag, New York.

Gu, C. and Wahba, G. (1993). Semiparametric ANOVA with tensor product thin plate spline, *Journal of the Royal Statistical Society B* **55**: 353–368.

Hall, P. and Carroll, R. J. (1989). Variance functin estimation in regression: the effect of estimating the mean, *Journal of the Royal Statistical Society B* **51**: 3–14.

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

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

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall.

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

Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis, *Annals of Statistics* **15**: 610–635.

Neumann, M. H. (1994). Fully data-driven nonparametric variance estimators, *Statistics* **25**: 189–212.

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

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

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wang, Y. and Ke, C. (2002). ASSIST: A suite of s-plus functions implementing spline smoothing techniques, Manual for the ASSIST package. Available at `http://www.pstat.ucsb.edu/faculty/yuedong/software`.