Biometrika (2005), **92,** 4, *pp.* 821–830

Printed in Great Britain

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, California 93106, U.S.A.

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

SUMMARY

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. For the special case of a 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. Our method can be extended to nonparametric regression models with multivariate functions defined on arbitrary subsets of normed spaces, possibly observed on unequally spaced or clustered designed points.

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

1. INTRODUCTION

Consider a nonparametric regression model

$$y_i = g(x_i) + \varepsilon_i \quad (1 \leqslant i \leqslant n),$$

where the y_i's are observations, g is an unknown mean function, and the 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 & Stadtmüller, 1987; Hall & Carroll, 1989; Carter & Eagleson, 1992; Neumann, 1994). However, it is often desirable to have an accurate estimator 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 & Spiegelman, 1990; Rice, 1984; Gasser et al., 1991; Kulasekera & Gallagher, 2002). An accurate estimator of σ^2 can also be used to estimate the detection limits of immunoassay (Carroll, 1987; Carroll & Ruppert, 1988).

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

$$\hat{\sigma}^2_D = y^{\mathrm{T}} D y / \mathrm{tr}(D). \tag{1}$$

These estimators usually fall into two classes. The first class of estimators are based on the residual sum of squares from some nonparametric fit to g (Wahba, 1990; Hastie & Tibshirani, 1990). For linear smoothers the fitted values are $\hat{y} = Ay$, where A is a smoother matrix. Then an estimator of variance has the form (1) with $D = (I - A)^{\mathrm{T}}(I - A)$ (Hastie & Tibshirani, 1990). We refer to estimators in the first class as residual-based estimators. Residual-based estimators depend critically on the amount of smoothing (Dette et al., 1998). Some methods require knowledge about some unknown quantity such as $\int_0^1 \{g'(t)\}^2 dt$ (Hall & Marron, 1990) or $\int_0^1 \{g''(t)\}^2 dt$ (Buckley et al., 1988).

The second class of estimators use differences that aim to remove trend in the mean function, an idea originating in time series analysis. Such methods do not require an estimator of the mean function and are often called difference-based estimators. Assume that x is univariate and $0 \leqslant x_1 \leqslant \ldots \leqslant x_n \leqslant 1$. Rice (1984) proposed the first-order difference-based estimator

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

Gasser et al. (1986) proposed the second-order difference-based estimator

$$\hat{\sigma}^2_{\mathrm{GSJ}} = \frac{1}{n-2} \sum_{i=2}^{n-1} c_i^2 \hat{\varepsilon}_i^2,$$

where $\hat{\varepsilon}_i$ 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 $Ec_i^2 \hat{\varepsilon}_i^2 = \sigma^2$ for all i when g is linear. For equidistant design points, $\hat{\sigma}^2_{\mathrm{GSJ}}$ reduces to

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

Hall et al. (1990) introduced the estimator

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

where m_1 and m_2 are nonnegative integers, $m = m_1 + m_2$ is referred to as the order, and the difference sequence $\{d_i\}_{i=-m_1,\ldots,m_2}$ 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 above difference-based estimators achieves the asymptotic optimal rate for the mean squared error (Dette et al., 1998), namely

$$\mathrm{MSE}(\hat{\sigma}^2) = E(\hat{\sigma}^2 - \sigma^2)^2 = n^{-1} \operatorname{var}(\varepsilon^2) + o(n^{-1}). \tag{3}$$

In practice, the choice of the order m and an appropriate difference sequence which minimises the finite-sample mean squared error 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.

Müller et al. (2003) proposed the class of difference-based estimators

$$\hat{\sigma}^2_{\mathrm{MSW}} = \frac{1}{2 \sum_{i \neq j} W_{ij}} \sum_{i \neq j} W_{ij} (y_i - y_j)^2,$$

where W_{ij} are weights depending on x only. For the random design, $\hat{\sigma}^2_{\mathrm{MSW}}$ achieves the asymptotic optimal rate (3) under certain assumptions for the weights; Müller et al. (2003) constructed weights based on a kernel density estimate for the variable x.

In this paper we propose a new estimator which is the estimated intercept of a linear model. When the design points are equally spaced in $[0, 1]$, using the optimal bandwidth, we can reduce the asymptotic rate of mean squared error to

$$\mathrm{MSE}(\hat{\sigma}^2) = n^{-1}\,\mathrm{var}(\varepsilon^2) + O(n^{-3/2}). \tag{4}$$

2. Main results

2·1. *Motivation*

We assume that $x_i = i/n$ for $1 \leqslant i \leqslant n$. Taking expectation of the Rice estimator, we have

$$E(\hat{\sigma}^2_{\mathrm{R}}) = \frac{1}{2(n-1)} \sum_{i=2}^{n} E(y_i - y_{i-1})^2 = \sigma^2 + \frac{1}{2(n-1)} \sum_{i=2}^{n} \{g(x_i) - g(x_{i-1})\}^2, \tag{5}$$

indicating that Rice's estimator is always positively biased. Suppose that g has a bounded first derivative. Then, from (5), we have

$$E(\hat{\sigma}^2_{\mathrm{R}}) = \sigma^2 + \frac{1}{n^2} J + o\left(\frac{1}{n^2}\right), \tag{6}$$

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

$$\hat{\sigma}^2_{\mathrm{R}}(k) = \frac{1}{2(n-k)} \sum_{i=k+1}^{n} (y_i - y_{i-k})^2 \quad (k = 1, \ldots, n-1).$$

Similar calculations give

$$E\{\hat{\sigma}^2_{\mathrm{R}}(k)\} = \sigma^2 + \frac{k^2}{n^2} J + O\left\{\frac{k^3}{n^2(n-k)}\right\} + o\left(\frac{1}{n^2}\right).$$

Thus, for any fixed $m = o(n)$, we have

$$E\{\hat{\sigma}^2_{\mathrm{R}}(k)\} \backsimeq \sigma^2 + J d_k \quad (1 \leqslant k \leqslant m), \tag{7}$$

where $d_k = k^2/n^2$. Throughout this paper, we take the integer part of m whenever necessary. Treating (7) as a simple linear regression model with d_k as the independent variable, we can estimate σ^2 as the intercept.

In being based on a simple parametric model, our method is similar in spirit to the simulation-extrapolation method of Cook & Stefanski (1994) and the empirical-bias bandwidth selection method of Ruppert (1997). For measurement error models, the simulation-extrapolation method adds additional measurement error to establish a trend of measurement-error-induced bias versus the variance of the added measurement error, and extrapolates this trend back to the case of no measurement error. For local polynomial nonparametric regression, Ruppert's method uses estimates at several bandwidths to fit a polynomial model and estimate the bias at a particular bandwidth as the intercept.

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 $1 \leqslant k \leqslant m$. As discussed in § 2·1, we regress s_k on d_k to estimate σ^2 as the intercept. We will discuss the choice of m in §§ 2·3 and 3·1. 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 $N = (n-1) + (n-2) + \ldots + (n-m) = nm - m(m+1)/2$. To be specific, we fit the linear model

$$s_k = \alpha + \beta d_k + e_k \quad (k = 1, \ldots, m)$$

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

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

$$\hat{\sigma}^2 = \hat{\alpha} = \bar{s}_w - \hat{\beta}\bar{d}_w, \tag{8}$$

where

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

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

THEOREM 1. *For the equally spaced design, we have the following*:

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

(ii) $\hat{\sigma}^2$ *can be written as a quadratic form* $\hat{\sigma}^2 = y^{\mathrm{T}} D y / \mathrm{tr}\,(D)$ *as in* (1), *where* D *is an* $n \times n$ *matrix with elements*

$$d_{ij} = \begin{cases} \sum_{k=1}^{m} b_k + \sum_{k=0}^{\min\{i-1, n-i, m\}} b_k & (1 \leqslant i = j \leqslant n), \\ -b_{|i-j|} & (0 < |i-j| \leqslant m), \\ 0 & \textit{otherwise}, \end{cases}$$

where

$$b_0 = 0, \quad b_k = 1 - \frac{\bar{d}_w (d_k - \bar{d}_w)}{\sum_{k=1}^{m} w_k (d_k - \bar{d}_w)^2} \quad (k = 1, \ldots, m).$$

The proof of Theorem 1 is omitted as it is straightforward. Note that D is a symmetric matrix with both row and column sums equal to zero. Since D is not guaranteed to be positive definite, $\hat{\sigma}^2$ may take a negative value and we recommend replacing such a value by zero. Our experience indicates that a negative estimate occurs very rarely; it never occurred in our extensive simulations. However, negative estimates could happen for other functions.

2·3. *Asymptotic results*

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

$$\begin{aligned} \mathrm{MSE}(\hat{\sigma}^2) = [&(g^{\mathrm{T}} D g)^2 + 4\sigma^2 g^{\mathrm{T}} D^2 g + 4 g^{\mathrm{T}} \{D\,\mathrm{diag}(D) u\} \sigma^3 \gamma_3 \\ &+ \sigma^4\,\mathrm{tr}\,\{\mathrm{diag}(D)^2\}(\gamma_4 - 3) + 2\sigma^4\,\mathrm{tr}(D^2)]/\mathrm{tr}(D)^2, \end{aligned} \tag{9}$$

where $g = (g(x_1), \ldots, g(x_n))^{\mathrm{T}}$, $\mathrm{diag}(D)$ denotes the diagonal matrix of the diagonal elements of D, $u = (1, \ldots, 1)^{\mathrm{T}}$ and $\gamma_i = E\{(\varepsilon/\sigma)^i\}$, for $i = 3, 4$. The first term in (9) 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 the Appendix we will prove the following theorem.

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 obtain*

$$\text{MSE}(\hat{\sigma}^2) = \frac{1}{n}\text{var}(\varepsilon^2) + \frac{9}{4nm}\sigma^4 + \frac{9m}{112n^2}\text{var}(\varepsilon^2) + o\left(\frac{1}{nm}\right) + o\left(\frac{m}{n^2}\right) + O\left(\frac{m^6}{n^6}\right). \quad (10)$$

The last term in (10) comes from the bias and the remaining terms come from the variance. Theorem 2 indicates that $\hat{\sigma}^2$ is a consistent estimator of σ^2. The asymptotically optimal bandwidth is $m_{\text{opt}} = \{28n\sigma^4/\text{var}(\varepsilon^2)\}^{\frac{1}{2}}$. It is obvious that $\text{MSE}\{\hat{\sigma}^2(m_{\text{opt}})\}$ satisfies (4).

3. Simulations and comparisons with other estimators

3·1. *Finite-sample choice of the bandwidth*

In our comparisons, we evaluate the performance of the Rice, Gasser et al., Hall et al., Müller et al. and our estimators. For simplicity of notation, we assume that random errors are normally distributed with mean zero and variance σ^2. Then $\text{var}(\varepsilon^2) = 2\sigma^4$ and $m_{\text{opt}} = (14n)^{\frac{1}{2}}$. 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 terms of slightly higher order ignored in the mean squared error (10) do depend on the smoothness of the function, and therefore the asymptotic optimal bandwidth applies for very large n only. For small to moderate n, we find that $m_{\text{opt}} = (14n)^{\frac{1}{2}}$ is too large, and we now discuss three strategies for selecting m in these situations.

Note that the dominant term in (10), $\text{var}(\varepsilon^2)/n$, cannot be reduced. Let $h(m) = 9\sigma^4/(4nm) + 9m\sigma^4/(56n^2)$ be the two higher-order terms. Our first strategy is to select the smallest $m = cn^{\frac{1}{2}}$ such that $h(m)/h(m_{\text{opt}}) \leqslant 1 + \lambda$, where $100\lambda\%$ is the percentage of increase in the higher-order terms. It is easy to check that $m = \{1 + \lambda - (\lambda^2 + 2\lambda)^{\frac{1}{2}}\}(14n)^{\frac{1}{2}}$. Note that the convergence rate of mean squared error remains the same. Our simulations in § 3·2 indicate that $m = n^{\frac{1}{2}}$ with $\lambda \doteq 1$ works very well. Let $m_{\text{s}} = n^{\frac{1}{2}}$. The increases in mean squared error are in the higher-order terms, and therefore the overall increase is usually not large.

Simulations in § 3·2 indicate that $m_{\text{s}} = n^{\frac{1}{2}}$ 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 $\text{bias}(\hat{\sigma}^2) = O(n^{-2})$. If we take $m = cn^{\tau}$, then $\text{bias}(\hat{\sigma}^2) = O(n^{-3+3\tau})$. It is easy to see that the largest τ such that $\text{bias}(\hat{\sigma}^2) = O(n^{-2})$ is $\tau = \frac{1}{3}$. This suggests choosing $m = m_{\text{t}} = n^{1/3}$. Simulations in § 3·2 indicate that m_{t} performs well when n is small and g is rough. For $m_{\text{t}} = n^{1/3}$, $\text{MSE}\{\hat{\sigma}^2(m_{\text{t}})\} = 2\sigma^4/n + 9n^{-4/3}\sigma^4/4 + o(n^{-4/3})$, which still satisfies (3).

An alternative approach is to use a data-driven method such as V-fold crossvalidation (Müller et al., 2003). We split the whole dataset into V disjoint subsamples, $S_1, \ldots, S_V$, we let $\hat{\sigma}_v^2(m)$ be the estimate of σ^2 based on the subsample $\cup_{i \neq v} S_i$ with bandwidth m, and we select $m = m_{\text{CV}}$ to minimise

$$\text{CV}(m) = \sum_{v=1}^{V} \{\hat{\sigma}^2(m) - \hat{\sigma}_v^2(m)\}^2. \quad (11)$$

Note that the design points in $\cup_{i \neq v} S_i$ are usually not equally spaced, and the $\hat{\sigma}_v^2(m)$ are computed using (12), see § 4, for the general design.

3·2. *Simulation results*

As in Seifert et al. (1993) and Dette et al. (1998), our simulations are based on $g(x) = 5\sin(\omega\pi x)$, with design points $x_i = i/n$ and independent and identically distributed Gaussian random errors ε_i with mean zero and variance σ^2. We consider $\omega = 1$, 2 and 4, corresponding to low, moderate and high oscillations respectively, three standard deviations, $\sigma = 0{\cdot}5$, 1·5, 4, and three sample sizes, $n = 15$, 100, 500.

For each simulation setting, we generate observations and compute the estimators $\hat{\sigma}^2_{\mathrm{R}}$, $\hat{\sigma}^2_{\mathrm{GSJ}}$, $\hat{\sigma}^2_{\mathrm{HKT}}(m)$, $\hat{\sigma}^2_{\mathrm{MSW}}$, $\hat{\sigma}^2(m_{\mathrm{s}})$, $\hat{\sigma}^2(m_{\mathrm{t}})$ and $\hat{\sigma}^2(m_{\mathrm{CV}})$. We repeat this process 10 000 times and compute mean squared errors for each method. The order m in $\hat{\sigma}^2_{\mathrm{HKT}}(m)$ acts as a tuning parameter which depends on the unknown function g. We set $m = 2$ in our simulations, so that

$$\hat{\sigma}^2_{\mathrm{HKT}}(2) = \frac{1}{n-2}\sum_{i=1}^{n-2}(0{\cdot}8090y_i - 0{\cdot}5y_{i+1} - 0{\cdot}3090y_{i+2})^2.$$

As in Müller et al. (2003), we compute the weights in $\hat{\sigma}^2_{\mathrm{MSW}}$ using a kernel density estimate for the variable x and select the bandwidth b for the kernel density estimate by cross-validation using (11). We also compute m_{CV} for our estimator, based on leave-one-out crossvalidation for $n = 15$ and $n = 100$, and 10-fold crossvalidation for $n = 500$. For $\hat{\sigma}^2_{\mathrm{MSW}}$, we take $b \in \{0{\cdot}08k, k = 1, \ldots, 5\}$ for $n = 15$, $b \in \{0{\cdot}012k, k = 1, \ldots, 10\}$ for $n = 100$ and $b \in \{0{\cdot}0025k, k = 1, \ldots, 30\}$ for $n = 500$. For our method, we take $m \in \{1, \ldots, 5\}$ for $n = 15$, $m \in \{1, \ldots, 10\}$ for $n = 100$ and $m \in \{1, \ldots, 30\}$ for $n = 500$.

Table 1 lists relative mean squared errors, $n\mathrm{MSE}/(2\sigma^4)$, for all methods for $n = 100$ and $n = 500$. In general, $\mathrm{MSE}_{\hat{\sigma}^2(m_{\mathrm{CV}})} \doteq \mathrm{MSE}_{\hat{\sigma}^2(m_{\mathrm{s}})} < \mathrm{MSE}_{\hat{\sigma}^2_{\mathrm{HKT}}(2)} < \mathrm{MSE}_{\hat{\sigma}^2_{\mathrm{R}}} < \mathrm{MSE}_{\hat{\sigma}^2_{\mathrm{GSJ}}}$ for most cases. We find that $\hat{\sigma}^2(m_{\mathrm{CV}})$ and $\hat{\sigma}^2(m_{\mathrm{s}})$ have smaller mean squared errors than $\hat{\sigma}^2_{\mathrm{MSW}}$ in most settings. The comparative performance of $\hat{\sigma}^2(m_{\mathrm{s}})$, $\hat{\sigma}^2(m_{\mathrm{t}})$ and $\hat{\sigma}^2(m_{\mathrm{CV}})$ depends on the

Table 1: *Simulation study. Relative mean squared errors of various estimators.*

n	ω	σ	$\hat{\sigma}^2_{\mathrm{R}}$	$\hat{\sigma}^2_{\mathrm{GSJ}}$	$\hat{\sigma}^2_{\mathrm{HKT}}(2)$	$\hat{\sigma}^2_{\mathrm{MSW}}$	$\hat{\sigma}^2(m_{\mathrm{s}})$	$\hat{\sigma}^2(m_{\mathrm{t}})$	$\hat{\sigma}^2(m_{\mathrm{CV}})$
100	1	0·5	1·56	1·98	1·43	6·06	1·19	1·36	1·18
		1·5	1·51	2·03	1·22	1·31	1·12	1·35	1·14
		4	1·53	2·00	1·24	1·09	1·14	1·34	1·19
	2	0·5	2·00	2·01	4·09	15·11	1·81	1·38	1·30
		1·5	1·52	2·02	1·30	2·95	1·14	1·36	1·15
		4	1·45	1·95	1·31	1·15	1·15	1·36	1·19
	4	0·5	9·06	1·96	48·24	50·63	26·77	1·52	5·51
		1·5	1·54	2·01	1·84	6·67	1·46	1·36	1·28
		4	1·58	1·98	1·24	1·68	1·15	1·35	1·20
500	1	0·5	1·48	1·94	1·24	1·96	1·06	1·21	1·14
		1·5	1·48	1·94	1·23	1·14	1·05	1·18	1·15
		4	1·52	1·96	1·28	1·09	1·05	1·18	1·12
	2	0·5	1·49	1·93	1·27	2·48	1·06	1·16	1·16
		1·5	1·53	1·98	1·29	1·63	1·05	1·17	1·17
		4	1·48	1·94	1·23	1·12	1·06	1·18	1·11
	4	0·5	1·58	1·97	1·65	4·04	1·51	1·18	1·24
		1·5	1·50	1·95	1·26	2·43	1·07	1·17	1·18
		4	1·51	1·96	1·26	1·36	1·04	1·16	1·11

smoothness of g, the sample size and the signal-to-noise ratio: $\hat{\sigma}^2(m_{\mathrm{CV}})$ and $\hat{\sigma}^2(m_{\mathrm{s}})$ have slightly smaller mean squared errors than that of $\hat{\sigma}^2(m_{\mathrm{t}})$ except for the cases $(n, \omega, \sigma) = (100, 2, 0{\cdot}5), (100, 4, 0{\cdot}5), (500, 4, 0{\cdot}5)$, in which g is rough and σ is small.

Performance when n is small is illustrated in Table 2 for $n = 15$, with squared biases and variances rescaled by multiplying by $n/(2\sigma^4)$. Mean squared errors of $\hat{\sigma}^2(m_{\mathrm{s}})$ and $\hat{\sigma}^2_{\mathrm{HKT}}(2)$ are dominated by biases when g is rough, and $\hat{\sigma}^2_{\mathrm{GSJ}}$ has much smaller biases, and thus much smaller mean squared errors, in these situations. Similar comparative results were reported in Seifert et al. (1993) and Dette et al. (1998): $\hat{\sigma}^2_{\mathrm{HKT}}(2)$ performs better when g is flat and/or n is large, while $\hat{\sigma}^2_{\mathrm{GSJ}}$ performs better when the opposite is true. As discussed in § 3·1, the approximate optimal rate of m, $n^{\frac{1}{2}}$, requires a large n or a smooth g. When g is rough and n is small, $m_{\mathrm{s}} = n^{\frac{1}{2}}$ is too large and this leads to large biases. One option is to control bias using $m_{\mathrm{t}} = n^{1/3}$, as discussed in § 3·1. As expected, $\hat{\sigma}^2(m_{\mathrm{t}})$ reduces the bias with small increase in the variance. Though the performance of $\hat{\sigma}^2(m_{\mathrm{t}})$ is a little worse than that of $\hat{\sigma}^2(m_{\mathrm{s}})$ for other cases, it performs well when $\hat{\sigma}^2(m_{\mathrm{s}})$ fails, and we therefore recommend $\hat{\sigma}^2(m_{\mathrm{t}})$ when the sample size is small and either g is rough or little is known about g.

Table 2: *Simulation study. Relative mean squared errors*, MSE, *squared biases*, Bias^2, *and variances*, Var, *for* $n = 15$.

ω	σ		$\hat{\sigma}^2_{\mathrm{R}}$	$\hat{\sigma}^2_{\mathrm{GSJ}}$	$\hat{\sigma}^2_{\mathrm{HKT}}(2)$	$\hat{\sigma}^2_{\mathrm{MSW}}$	$\hat{\sigma}^2(m_{\mathrm{s}})$	$\hat{\sigma}^2(m_{\mathrm{t}})$	$\hat{\sigma}^2(m_{\mathrm{CV}})$
1	0·5	MSE	9·68	2·15	42·68	32·16	2·93	2·44	7·51
		Bias^2	7·73	0·00	39·67	25·33	0·87	0·11	4·12
		Var	1·95	2·15	3·01	6·83	2·06	2·33	3·39
	1·5	MSE	1·73	2·17	2·08	2·53	1·67	2·13	1·38
		Bias^2	0·10	0·00	0·50	0·66	0·01	0·00	0·01
		Var	1·63	2·17	1·58	1·87	1·66	2·13	1·37
	4	MSE	1·57	2·16	1·43	1·25	1·60	2·07	1·29
		Bias^2	0·00	0·00	0·01	0·00	0·00	0·00	0·02
		Var	1·57	2·16	1·42	1·25	1·60	2·07	1·27
2	0·5	MSE	125·88	2·78	616·81	230·15	20·09	5·67	66·83
		Bias^2	122·39	0·62	607·19	206·58	16·67	2·51	33·72
		Var	3·49	2·16	9·62	23·57	3·42	3·16	33·11
	1·5	MSE	3·34	2·20	9·84	9·48	1·98	2·21	2·65
		Bias^2	1·55	0·01	7·54	5·75	0·20	0·03	0·93
		Var	1·79	2·19	2·30	3·73	1·78	2·18	1·72
	4	MSE	1·66	2·18	1·71	1·66	1·63	2·11	1·32
		Bias^2	0·03	0·00	0·15	0·16	0·01	0·00	0·00
		Var	1·63	2·18	1·56	1·50	1·62	2·11	1·32
4	0·5	MSE	1876·20	117·49	8006·20	5222·90	848·50	118·42	1890·30
		Bias^2	1860·60	114·01	7949·80	4928·50	836·00	111·90	1868·00
		Var	15·60	3·48	56·40	294·40	12·50	6·52	22·30
	1·5	MSE	26·35	3·72	105·72	70·88	13·08	3·82	35·64
		Bias^2	23·22	1·41	98·16	60·10	10·25	1·33	28·32
		Var	3·13	2·31	7·56	10·78	2·83	2·49	7·32
	4	MSE	2·25	2·18	4·24	3·62	1·96	2·16	2·45
		Bias^2	0·47	0·03	1·97	1·37	0·22	0·03	0·53
		Var	1·78	2·15	2·27	2·25	1·74	2·13	1·92

For the equally spaced design, it is clear that $\hat{\sigma}^2(1) = \hat{\sigma}^2_{\mathrm{R}}$; that is, the Rice estimator is a special case of our estimator with $m = 1$. One interesting observation from simulations is that $\hat{\sigma}^2(2) \doteq \hat{\sigma}^2_{\mathrm{GSJ}}$ when σ^2 is not very small. In theory it is easy to show that the dominant term of $\mathrm{MSE}\{\hat{\sigma}^2(2)\}$ is $35\sigma^4/(9n)$, which is exactly the same as that of $\hat{\sigma}^2_{\mathrm{GSJ}}$. We have performed many more simulations with different mean functions, signal-to-noise ratios and sample sizes; the comparative conclusions remain the same.

4. DISCUSSION

Most difference-based methods in the literature focus on univariate x, although some extensions to multivariate x have recently been developed (Kulasekera & Gallagher, 2002; Müller et al., 2003): Kulasekera & Gallagher (2002) required an artificial ordering of the design points in $x \in [0, 1]^d$; Müller et al. (2003) required weights to satisfy certain conditions and specific weights based on kernel density estimation were constructed for univariate x only; and Spokoiny (2002) proposed a residual-based estimator for multivariate x.

Our estimator is different from existing residual-based and difference-based estimators. Most existing difference-based estimators require the design points to be ordered. It is thus difficult to extend these methods to high-dimensional or general domains since there is no clear ordering in these scenarios. In addition, for asymptotic theory, design points are assumed to satisfy $\max|x_i - x_{i-1}| = O(n^{-1+\delta})$, where $0 < \delta < \frac{1}{2}$ for $\hat{\sigma}^2_{\mathrm{HKT}}$ and $\delta = 0$ for $\hat{\sigma}^2_{\mathrm{R}}$ and $\hat{\sigma}^2_{\mathrm{GSJ}}$. In practice, unequally spaced designs may have clusters, tied design points and/or large gaps between neighbouring design points, so that it may not hold that $\max|x_i - x_{i-1}| = O(n^{-1+\delta})$.

We now provide an alternative derivation of our estimator which extends naturally to a general domain. The basic idea is to collect squared distances, $d_{ij} = (x_i - x_j)^2$, for all pairs $\{x_i, x_j\}$, and half squared differences, $s_{ij} = (y_i - y_j)^2/2$, for all pairs $\{y_i, y_j\}$, and then regress s_{ij} on d_{ij} using paired design points which are close to each other. To be specific, we fit the following simple linear model,

$$s_{ij} = \alpha + \beta d_{ij} + e_{ij} \quad (d_{ij} \leqslant M), \tag{12}$$

by least squares, where $M > 0$ is the bandwidth. The estimator of σ^2 is $\hat{\sigma}^2 = \hat{\alpha}$. For univariate x with $x_i = i/n$ and $M = (m/n)^2$, it is not difficult to check that $\hat{\sigma}^2$ reduces to the weighted least squares estimator proposed in § 2·2. Note that the above derivation only needs the distances between design points; no ordering is required. Therefore, to extend our method to a general domain $\mathscr{T}$, where $\mathscr{T}$ is an arbitrary subset of a normed space, we let $d_{ij} = \|x_i - x_j\|^2$ and proceed as above. Interesting examples of $\mathscr{T}$ are $\mathscr{R}^d$, the unit circle and the unit sphere. Since we use pairs of points which are close together, our method does not require dense design points in the whole domain, thus avoiding the curse of dimensionality problem in high-dimensional space and allowing sizeable gaps between some design points. Further research is necessary to investigate the properties of our estimator for a general domain and to develop methods for selecting the bandwidth M. One simple approach is to plot d_{ij} against s_{ij} and determine an appropriate range for the approximate linear relationship, or to suggest a different parametric relationship. Another future research topic is to fit (12) using weighted least squares with weights that decrease as distances increase. This might improve the performance and make our estimator less sensitive to the choice of the bandwidth.

Acknowledgement

This research was supported by a grant from the U.S. National Institutes of Health. The authors thank the editor and three referees for their constructive comments and suggestions that have led to a substantial improvement in the paper.

Appendix

Proof of Theorem 2

We provide a sketch of the proof only. Details of the proof and the proof of Theorem 1 can be found in a technical report available at http://www.pstat.ucsb.edu/faculty/yuedong/research.

Asymptotic bias. Instead of using the formula $\text{bias}(\hat{\sigma}^2) = g^{\mathrm{T}}Dg/\text{tr}(D)$, we calculate this quantity directly from (8), thereby obtaining a more accurate approximation. Let $I_t = \sum_{k=1}^{m} k^t$, for $t = 1, 2, \ldots$. It is not difficult to check that

$$E(\bar{s}_w) = \sigma^2 + \{I_2/(Nn) - I_3/(Nn^2)\}J + O(m^3/n^3), \quad \bar{d}_w = I_2/(Nn) - I_3/(Nn^2) = m^2/(3n^2) + o(m^2/n^2),$$

$$\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2 = I_4/(Nn^3) - I_5/(Nn^4) - \{I_2/(Nn) - I_3/(Nn^2)\}^2 = 4m^4/(45n^4) + o(m^4/n^4),$$

$$\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)Es_k = J[I_4/(Nn^3) - I_5/(Nn^4) - \{I_2/(Nn) - I_3/(Nn^2)\}^2] + O(m^5/n^5).$$

Therefore,

$$E(\hat{\sigma}^2) = E(\bar{s}_w) - \frac{\bar{d}_w}{\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2} \sum_{k=1}^{m} w_k(d_k - \bar{d}_w)Es_k = \sigma^2 + O\left(\frac{m^3}{n^3}\right).$$

Asymptotic variance. It can be shown that $g^{\mathrm{T}}D^2g = O(m^5/n^2)$, $g^{\mathrm{T}}\{D\,\text{diag}(D)u\} = O(m^4/n)$, $\text{tr}\{\text{diag}(D)^2\} = 4nm^2 - 103m^3/28 + o(m^3)$ and $\text{tr}(D^2) = 4nm^2 - 103m^3/28 + 9nm/2 + o(m^3) + o(nm)$. Together with the fact that $\sigma^4(\gamma_4 - 3) = \text{var}(\varepsilon^2) - 2\sigma^4$, we have

$$\begin{aligned}\text{var}(\hat{\sigma}^2) &= \frac{1}{\{\text{tr}(D)\}^2}[4\sigma^2 g^{\mathrm{T}}D^2g + 4g^{\mathrm{T}}\{D\,\text{diag}(D)u\}\sigma^3\gamma_3 + \sigma^4\text{tr}\{\text{diag}(D)^2\}(\gamma_4 - 3) + 2\sigma^4\text{tr}(D^2)] \\ &= \frac{1}{n}\text{var}(\varepsilon^2) + \frac{9}{4nm}\sigma^4 + \frac{9m}{112n^2}\text{var}(\varepsilon^2) + o\left(\frac{1}{nm}\right) + o\left(\frac{m}{n^2}\right).\end{aligned}$$

Asymptotic mean squared error. The proof of (10) can be obtained immediately from the asymptotic bias and variance.

References

Buckley, M. J., Eagleson, G. K. & Silverman, B. W. (1988). The estimation of residual variance in non-parametric regression. *Biometrika* **75**, 189–99.

Carroll, R. J. (1987). The effects of variance function estimation on prediction and calibration: an example. In *Statistical Decision Theory and Related Topics, IV*, Ed. S. S. Gupta and J. O. Berger, **2**, pp. 273–80. New York: Springer-Verlag.

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

Carter, C. K. & Eagleson, G. K. (1992). A comparison of variance estimators in nonparametric regression. *J. R. Statist. Soc.* B **54**, 773–80.

Cook, J. R. & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Am. Statist. Assoc.* **89**, 1314–28.

Dette, H., Munk, A. & Wagner, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. R. Statist. Soc.* B **60**, 751–64.

Eubank, R. L. & Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Am. Statist. Assoc.* **85**, 387–92.

GASSER, T., KNEIP, A. & KOHLER, W. (1991). A flexible and fast method for automatic smoothing. *J. Am. Statist. Assoc.* **86**, 643–52.

GASSER, T., SROKA, L. & JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–33.

HALL, P. & CARROLL, R. J. (1989). Variance function estimation in regression: the effect of estimating the mean. *J. R. Statist. Soc.* B **51**, 3–14.

HALL, P. & MARRON, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77**, 415–9.

HALL, P., KAY, J. W. & TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–8.

HASTIE, T. & TIBSHIRANI, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

KULASEKERA, K. B. & GALLAGHER, C. (2002). Variance estimation in nonparametric multiple regression. *Commun. Statist.* A **31**, 1373–83.

MÜLLER, H. G. & STADTMÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610–35.

MÜLLER, U. U., SCHICK, A. & WEFELMEYER, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched *U*-statistic. *Statistics* **37**, 179–88.

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

RICE, J. A. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215–30.

RUPPERT, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Assoc.* **92**, 1049–62.

SEIFERT, B., GASSER, T. & WOLF, A. (1993). Nonparametric estimation of residual variance revisited. *Biometrika* **80**, 373–83.

SPOKOINY, V. (2002). Variance estimation for high-dimensional regression models. *J. Mult. Anal.* **82**, 111–33.

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

[*Received August* 2004. *Revised March* 2005]