Communications in Statistics—Theory and Methods, 37: 2890–2902, 2008

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920802162656

Taylor & Francis
Taylor & Francis Group

Nonparametric Inference

Relative Errors of Difference-Based Variance Estimators in Nonparametric Regression

TIEJUN TONG[1], ANNA LIU[2], AND YUEDONG WANG[3]

[1]Department of Applied Mathematics, University of Colorado, Boulder, Colorado, USA
[2]Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts, USA
[3]Department of Statistics and Applied Probability, University of California, Santa Barbara, California, USA

Difference-based estimators for the error variance are popular since they do not require the estimation of the mean function. Unlike most existing difference-based estimators, new estimators proposed by Müller et al. (2003) and Tong and Wang (2005) achieved the asymptotic optimal rate as residual-based estimators. In this article, we study the relative errors of these difference-based estimators which lead to better understanding of the differences between them and residual-based estimators. To compute the relative error of the covariate-matched U-statistic estimator proposed by Müller et al. (2003), we develop a modified version by using simpler weights. We further investigate its asymptotic property for both equidistant and random designs and show that our modified estimator is asymptotically efficient.

Keywords Asymptotically efficient; Bandwidth; Kernel estimator; Mean squared error; Nonparametric regression; Variance estimation.

Mathematics Subject Classification Primary 62G08; Secondary 62G20.

1. Introduction

We consider the following nonparametric regression model:

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

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 common

Received September 1, 2006; Accepted March 3, 2008
Address correspondence to Tiejun Tong, Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, USA; E-mail: tiejun.tong@colorado.edu

variance σ^2. Estimation of the error variance σ^2 has attracted a great deal of attention; see for example Wahba (1978), Rice (1984), Gasser et al. (1986), Buckley et al. (1988), Hall and Marron (1990), Hall et al. (1990), Carter and Eagleson (1992), Dette et al. (1998), Müller et al. (2003), Munk et al. (2005), and Tong and Wang (2005), among others. A good estimator of σ^2 is essential for inferences and choosing the amount of smoothing (Rice, 1984).

For example, suppose that $x_i \in [0, 1]$ and let

$$W_2[0, 1] = \left\{ g : g, g' \text{ absolutely continuous}, \int_0^1 (g^{(2)}(x))^2 \, dx < \infty \right\}.$$

A cubic spline estimate of g, $\hat{g}_\lambda$, is the minimizer of the following penalized least square:

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(x_i))^2 + \lambda \int_0^1 (g^{(2)}(x))^2 \, dx,$$

where λ is a smoothing parameter which controls the amount of smoothing. The performance of a spline estimate critically depends on a good choice of λ. It is very important to have a data-driven method for selecting λ. The unbiased risk (UBR) method estimates λ as the minimizer of Wahba (1990)

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{g}_\lambda(x_i))^2 + \frac{2\sigma^2}{n} \operatorname{tr} A(\lambda),$$

where $A(\lambda)$ is the hat matrix such that $(\hat{g}_\lambda(x_1), \ldots, \hat{g}_\lambda(x_n))^T = A(\lambda)(Y_1, \ldots, Y_n)^T$. Therefore, we need an estimate of σ^2 without fitting the mean function g first.

Most estimators of σ^2 in the literature are quadratic forms of the response vector $Y = (Y_1, \ldots, Y_n)^T$:

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

These estimators fall into two classes in general. The first class of estimators are based on the residual sum of squares from some nonparametric fit to g. Specifically, we first estimate g by a nonparametric method such as kernel smoothing or spline smoothing (Carter and Eagleson, 1992; Hall and Marron, 1990; Hastie and Tibshirani, 1990; Wahba, 1990). For linear smoothers we have $\widehat{Y} = AY$, where A is a smoother matrix. Then an estimator of variance has the form (2) with $D = (I - A)^T (I - A)$. We refer to estimators in the first class as residual-based estimators. Residual-based estimators depend critically on the choice of smoothing parameter, which requires knowledge of some unknown quantity such as $\int_0^1 g'(t)^2 \, dt/\sigma^2$ (Hall and Marron, 1990) or $\int_0^1 g''(t)^2 \, dt/\sigma^2$ (Buckley et al., 1988). Therefore, the practical applications of these estimators are somewhat limited.

Another class of estimators use differences to remove trend in the mean, an idea originating in time series analysis. We refer to estimators in this class as difference-based estimators; see Sec. 2 for a detailed review. Difference-based estimators do not require estimate of the mean function and thus are very popular in practice due to their ease of implementation. In addition, difference-based estimators are attractive from a practical point of view because they often have small biases for

small sample sizes (Dette et al., 1998). However, as we will see in Sec. 2, none of the fixed-order difference-based estimators achieves the following asymptotic optimal rate for residual-based estimators (Buckley et al., 1988; Eagleson, 1989; Hall and Marron, 1990):

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

Recently, Müller et al. (2003) and Tong and Wang (2005) proposed two new types of difference-based estimators for the error variance σ^2. Both of them reach the asymptotic optimal rate (3). Note that the dominant term in MSE, $n^{-1}\mathrm{var}(\epsilon^2)$, cannot be reduced. Hall and Marron (1990) showed that the relative error of MSE, the second term in (3), is of size $n^{-(4r-1)/(4r+1)}$ for their residual-based estimator, and it is the smallest possible in the minimax sense (see Sec. 4). In this article, we study the relative errors of difference-based estimators which lead to better understanding of the differences between them and residual-based estimators. We review the existing difference-based estimators in Sec. 2. In Sec. 3, we develop a modified version of the covariate-matched U-statistics estimator by Müller et al. (2003) using simpler weights. We further investigate its asymptotic property for both equidistant and random designs, and show that our modified estimator is asymptotically efficient. We compare the relative errors of difference-based estimators to residual-based estimators in Sec. 4 and present technical details and proofs in Sec. 5.

2. Existing Difference-Based Estimators

2.1. *Fixed-Order Difference-Based Estimators*

The order of a difference-based estimator is defined as the number of related observations involved in calculating a local residual. Rice (1984) proposed a first-order difference-based estimator

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

an idea originated by von Neumann (1941) as the mean square successive difference. Under proper conditions, we have $\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{R}}) = \frac{3}{2}n^{-1}\mathrm{var}(\epsilon^2)(1 + o(1))$.

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

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

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 coefficient c_i are chosen such that $Ec_i^2\hat{\epsilon}_i^2 = \sigma^2$ for all i when g is linear. For equidistant design points, $c_i \equiv \sqrt{6}/3$ for any i. Gasser et al. (1986) showed that $\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{GSJ}}) = \frac{35}{18}n^{-1}\mathrm{var}(\epsilon^2)(1 + o(1))$.

Hall et al. (1990) introduced the following optimal difference-based estimator:

$$\hat{\sigma}^2_{\mathrm{HKT}} = \frac{1}{n-r} \sum_{k=m_1+1}^{n-m_2} \left(\sum_{j=-m_1}^{m_2} d_j Y_{j+k} \right)^2, \tag{4}$$

where m_1 and m_2 are non negative integers, $r = m_1 + m_2$ is referred to as the order, and the difference sequence $\mathbf{d} \triangleq \{d_i\}_{i=-m_1,\ldots,m_2}$ satisfies $\sum d_j = 0$, $\sum d_j^2 = 1$ and $d_{-m_1} d_{m_2} \neq 0$. The optimal sequence $\mathbf{d}_{opt}$ is defined to minimize the asymptotic MSE of $\hat{\sigma}^2_{\mathrm{HKT}}$. Hall et al. (1990) showed that $\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{HKT}}(\mathbf{d}_{opt})) = n^{-1}\mathrm{var}(\epsilon^2)(1 + r^{-1} + o(1))$.

Clearly, none of the fixed-order difference-based estimators achieves the above asymptotic optimal rate for residual-based estimators. Note that $\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{HKT}}(\mathbf{d}_{opt}))$ decreases asymptotically with the order r. Theoretically, $\hat{\sigma}^2_{\mathrm{HKT}}(\mathbf{d}_{opt})$ achieves the optimal rate (3) as $r \to \infty$ and $r/n \to 0$ (Dette et al., 1998).

2.2. *Müller, Schick and Wefelmeyer's Estimator*

Müller et al. (2003) proposed the following covariate-matched U-statistic:

$$\hat{\sigma}^2_{\mathrm{MSW}} = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} (Y_i - Y_j)^2 w_{ij}, \tag{5}$$

where w_{ij} are non negative symmetric weights depending on covariates only and satisfying $\sum_{j \neq i} w_{ij}/n(n-1) = 1$. Unlike the traditional difference-based estimators, the estimator (5) uses all squared differences of paired observations to estimate σ^2. Under some conditions on the weights, Müller et al. (2003) showed that (5) is asymptotically efficient. Specifically, $\hat{\sigma}^2_{\mathrm{MSW}}$ has the following *i.i.d.* representation:

$$\hat{\sigma}^2_{\mathrm{MSW}} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 + o(n^{-1/2}), \tag{6}$$

where ϵ_i's are defined in (1). Therefore, $\hat{\sigma}^2_{\mathrm{MSW}}$ possesses (3) by noting that $\hat{\sigma}^2_{\mathrm{MSW}}$ has the asymptotic variance $n^{-1}\mathrm{var}(\epsilon^2)$ and the bias square is of order $o(n^{-1})$. Müller et al. (2003) also proposed the following kernel specific weights:

$$w_{ij} = \frac{1}{2}\left(\frac{1}{\hat{f}_i} + \frac{1}{\hat{f}_j}\right) K_h(x_i - x_j), \tag{7}$$

where $K_h(x) = K(x/h)/h$ is a symmetric kernel function with bandwidth h, and $\hat{f}_i = \sum_{j:j \neq i} K_h(x_i - x_j)/(n-1)$, $i = 1, \ldots, n$.

2.3. *Tong and Wang's Estimator*

Motivated by the fact that the Rice estimator is always positively biased, Tong and Wang (2005) introduced the lag-k Rice estimators as:

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

where $\hat{\sigma}^2_{\mathrm{R}} = \hat{\sigma}^2_{\mathrm{R}}(1)$. Suppose that the mean function g has a bounded first derivative and define $J = \int_0^1 \{g'(x)\}^2 \, dx/2$. It is easy to see that

$$E(\hat{\sigma}^2_{\mathrm{R}}(k)) \doteq \sigma^2 + Jd_k, \quad \text{for any } k = o(n), \tag{8}$$

where $d_k = k^2/n^2$. Treating (8) as a simple linear regression model with d_k as the independent variable, the Tong and Wang estimator, $\hat{\sigma}^2_{\mathrm{TW}}$, is defined as the intercept which corrects the corresponding bias. Under proper conditions, Tong and Wang (2005) showed that with an optimal bandwidth,

$$\mathrm{MSE}(\hat{\sigma}^2_{\mathrm{TW}}(h_{\mathrm{opt}})) = n^{-1}\mathrm{var}(\epsilon^2)\left(1 + O(n^{-1/2})\right), \tag{9}$$

which satisfies (3).

3. Modified MSW Estimator

Note that the weights in (7) are not well defined on the event $\{\min_i \hat{f}_i = 0\}$. We consider a modified version of the MSW estimator,

$$\tilde{\sigma}^2_{\mathrm{MSW}} = \frac{1}{2W}\sum_{i=1}^{n}\sum_{j\neq i}(Y_i - Y_j)^2 w_{ij}, \tag{10}$$

where w_{ij} are some given weights and $W = \sum_{i=1}^{n}\sum_{j\neq i} w_{ij}$. It is easy to see that this estimator is unbiased when g is a constant function. Different forms of weights can be used. For example, for the equidistant designs on $[0, 1]$ with $x_i = i/n$, $i = 1, \ldots, n$, weights

$$w_{ij} = \begin{cases} 1, & \text{if } |x_i - x_j| = 1/n, \\ 0, & \text{if } |x_i - x_j| > 1/n, \end{cases}$$

lead to the Rice estimator $\hat{\sigma}^2_{\mathrm{R}}$; and weights

$$w_{ij} = \begin{cases} 4, & \text{if } |x_i - x_j| = 1/n, \\ -1, & \text{if } |x_i - x_j| = 2/n, \\ 0, & \text{if } |x_i - x_j| > 2/n, \end{cases}$$

with boundary values $w_{12} = w_{n-1,n} = 2$ lead to the GSJ estimator $\hat{\sigma}^2_{\mathrm{GSJ}} = \sum_{i=3}^{n}(Y_i - 2Y_{i-1} + Y_{i-2})^2/6(n-2)$. In addition, the modified MSW estimator also possesses a quadratic form such that $\tilde{\sigma}^2_{\mathrm{MSW}} = Y^T\mathbf{D}Y/\mathrm{tr}(\mathbf{D})$ where

$$\mathbf{D} = \frac{1}{W}\begin{pmatrix} \sum_{j\neq 1} w_{1j} & -w_{12} & \cdots & -w_{1n} \\ -w_{21} & \sum_{j\neq 2} w_{2j} & \cdots & -w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ -w_{n1} & -w_{n2} & \cdots & \sum_{j\neq n} w_{nj} \end{pmatrix}.$$

In this section, we consider kernel weights $w_{ij} = K_h(x_j - x_i)$ with bandwidth h. Note that our weights are simpler than those in (7). We distinguish two types of designs: (i) equidistant design on $[0, 1]$ with $x_i = i/n$ for $1 \leq i \leq n$; and (ii) random design where the x_i's are an *i.i.d.* sample with a density f on $[0, 1]$. Assume that the kernel K is of order r. That is, K satisfies $\int_{-1}^{1} K(u)du = 1$, $\int_{-1}^{1} u^i K(u)du = 0$

for $i = 1, \ldots, r-1$, $\int_{-1}^{1} u^r K(u)du \neq 0$ and $\int_{-1}^{1} K^2(u)du < \infty$ (Eubank, 1999). The boundary kernel will be employed in the region such that x is close to 0 or 1.

3.1. *Equidistant Design*

For a function m, denote $\|m\|_2 = \sqrt{\int_{-\infty}^{\infty} m^2(z)dz}$ and $\mu_r(m) = \int_{-\infty}^{\infty} z^r m(z)dz$. Using the fact that $\tilde{\sigma}^2_{\mathrm{MSW}}$ has a quadratic form, we have the following formula for the MSE (Dette et al., 1998):

$$\begin{aligned}\mathrm{MSE}(\tilde{\sigma}^2_{\mathrm{MSW}}) = \big\{&(\mathbf{g}^T\mathbf{D}\mathbf{g})^2 + 4\sigma^2\mathbf{g}^T\mathbf{D}^2\mathbf{g} + 4\mathbf{g}^T\{\mathbf{D}\mathrm{diag}(\mathbf{D})\mathbf{1}\}\sigma^3\gamma_3 \\ &+ \sigma^4\mathrm{tr}\{\mathrm{diag}(\mathbf{D})^2\}(\gamma_4 - 3) + 2\sigma^4\mathrm{tr}(\mathbf{D}^2)\big\}/\mathrm{tr}(\mathbf{D})^2, \end{aligned} \tag{11}$$

where $\mathbf{g} = (g(x_1), \ldots, g(x_n))^T$, $\mathrm{diag}(\mathbf{D})$ denotes the diagonal matrix of the diagonal elements of $\mathbf{D}$, $\mathbf{1} = (1, \ldots, 1)^T$ and $\gamma_i = E[(\epsilon/\sigma)^i]$, $i = 3, 4$. The first term in (11) 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.

Theorem 3.1. *Assume that the mean function g has the rth derivative on* $[0, 1]$. *Then for equidistant designs with $h \to 0$ and $nh \to \infty$, we have*:

$$\mathrm{bias}(\tilde{\sigma}^2_{\mathrm{MSW}}) = C_1 h^r + o(h^r) + O(n^{-1}), \tag{12}$$

$$\mathrm{var}(\tilde{\sigma}^2_{\mathrm{MSW}}) = n^{-1}\mathrm{var}(\epsilon^2) + (n^2h)^{-1}C_2 + O(n^{-1}h^r) + o(n^{-2}h^{-1}), \tag{13}$$

$$\mathrm{MSE}(\tilde{\sigma}^2_{\mathrm{MSW}}) = n^{-1}\mathrm{var}(\epsilon^2) + C_1^2h^{2r} + (n^2h)^{-1}C_2 + o((n^2h)^{-1} + h^{2r}) + O(n^{-2}), \tag{14}$$

where $C_1 = \mu_r(K)\int_0^1\{(g^2(x))^{(r)} - 2g(x)g^{(r)}(x)\}dx/2r!$ *and* $C_2 = 2\sigma^4\|K\|_2^2$. *The asymptotic optimal bandwidth is*

$$h_{opt} = \left(\frac{C_2}{2rn^2C_1^2}\right)^{1/(2r+1)}, \tag{15}$$

which is of order $n^{-2/(2r+1)}$.

The proofs of (12) and (13) in Theorem 3.1 are shown in Sec. 5. (14) is an immediate result from (12) and (13) by noting that $n^{-1}h^r$ is dominated by $(n^2h)^{-1}$ and h^{2r} as $h \to 0$. Substituting (15) into (14) leads to:

$$\mathrm{MSE}(\tilde{\sigma}^2_{\mathrm{MSW}}(h_{opt})) = n^{-1}\mathrm{var}(\epsilon^2)\big(1 + O(n^{-(2r-1)/(2r+1)})\big). \tag{16}$$

This asymptotic rate reaches the optimal rate (3) for any $r \geq 1$. Furthermore, with the optimal bandwidth, the following theorem shows that $\tilde{\sigma}^2_{\mathrm{MSW}}(h_{opt})$ behaves asymptotically like the average of the squared errors.

Theorem 3.2. *Under the same conditions as in Theorem* 3.1, *we have*:

$$\tilde{\sigma}^2_{\mathrm{MSW}}(h_{opt}) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + O_p(n^{-2r/(2r+1)}). \tag{17}$$

Consequently, $n^{1/2}(\tilde{\sigma}^2_{\mathrm{MSW}}(h_{opt}) - \sigma^2)$ *converges in distribution to a normal random variable with mean zero and variance* $var(\epsilon^2)$.

The proof of Theorem 3.2 is given in Sec. 5. Note that the higher-order term in (17) is expressed more precisely than that in Theorem 3.1 of Müller et al. (2003).

3.2. *Random Design*

Suppose that the design points $\{x_i, i = 1, \ldots, n\}$ are an *i.i.d.* sample with a non zero density f on $[0, 1]$. Denote $(gf)^{(r)} = (d/dx)^{(r)}\{g(x)f(x)\}$. Similarly as the equidistant design, we have the following theorem.

Theorem 3.3. *Assume that both the mean function* g *and the design density* f *have the* r*th derivative on* $[0, 1]$ *with* $h \to 0$ *and* $nh \to \infty$. *Then*

$$\begin{aligned}
\mathrm{bias}(\tilde{\sigma}^2_{\mathrm{MSW}}) &= C_3 h^r + o(h^r) + O(n^{-1}),\\
\mathrm{var}(\tilde{\sigma}^2_{\mathrm{MSW}}) &= n^{-1}\mathrm{var}(\epsilon^2) + (n^2h)^{-1}C_4 + O(n^{-1}h^r) + o(n^{-2}h^{-1}),\\
\mathrm{MSE}(\tilde{\sigma}^2_{\mathrm{MSW}}) &= n^{-1}\mathrm{var}(\epsilon^2) + C_3^2 h^{2r} + (n^2h)^{-1}C_4 + o\left((n^2h)^{-1} + h^{2r}\right) + O(n^{-2}), \quad (18)
\end{aligned}$$

where $C_3 = (2r!\|f\|_2^2)^{-1}\mu_r(K)\int_0^1 \left\{(g^2f)^{(r)} - 2g(gf)^{(r)} + f^{(r)}(x)\right\} dx$ *and* $C_4 = 2\sigma^4\|K\|_2^2/\|f\|_2^2$. *The asymptotic optimal bandwidth is*:

$$h_{opt} = \left(\frac{C_4}{2rn^2C_3^2}\right)^{1/(2r+1)}. \quad (19)$$

Comparing (18) with (14), the only differences are in the coefficients of h^{2r} and $(n^2h)^{-1}$. This implies that the asymptotic rate for the random design is again optimal. The asymptotic normality property (17) also holds for random designs. The proofs of Theorem 3.3 and the asymptotic normality are not shown because they are similar to those for the equidistant design.

4. Comparison with Residual-Based Estimators

Hall and Marron (1990) proposed the following residual-based estimator:

$$\hat{\sigma}^2_{\mathrm{HM}} = \frac{\sum_{i=1}^n (Y_i - \hat{g}(x_i))^2}{n - 2\sum_{i=1}^n w_{ii} + \sum_{i=1}^n \sum_{j=1}^n w_{ij}^2}, \quad (20)$$

where $\hat{g}(x_i) = \sum_{j=1}^n w_{ij}Y_j$ is a kernel estimator of the mean function with weights $w_{ij} = K\{(x_i - x_j)/h\}/\sum_{k=1}^n K\{(x_i - x_k)/h\}$, with h being the bandwidth and $K(\cdot)$ being a kernel of order r. This estimator is unbiased when the mean function g is zero. The MSE with an optimal bandwidth is given as:

$$\mathrm{MSE}\left(\hat{\sigma}^2_{\mathrm{HM}}(h_{opt})\right) = n^{-1}\mathrm{var}(\epsilon^2)\left(1 + O(n^{-(4r-1)/(4r+1)})\right). \quad (21)$$

Further, $\hat{\sigma}^2_{\mathrm{HM}} - \mathrm{E}(\hat{\sigma}^2_{\mathrm{HM}})$ is asymptotically normally distributed with variance $n^{-1}\mathrm{var}(\epsilon^2)$ as $h \to 0$ and $nh \to \infty$. Although residuals presented in their article are

based on kernel estimators, it applies to other methods such as smoothing spline estimators as well.

Hall and Marron (1990) proved that the relative error of size $n^{-(4r-1)/(4r+1)}$ in (21), is the smallest possible in the minimax sense. None of the difference-based estimators attains this optimal rate. This is probably the price paid by the difference-based estimators for not estimating the mean function. Noting that $\hat{\sigma}^2_{\mathrm{TW}}$ is not kernel-based estimator and its relative error is always fixed at size $n^{-1/2}$, $\tilde{\sigma}^2_{\mathrm{MSW}}(h_{opt})$ has a smaller relative error than $\hat{\sigma}^2_{\mathrm{TW}}(h_{opt})$ for any $r \geq 2$. Furthermore, the difference-based estimator $\tilde{\sigma}^2_{\mathrm{MSW}}(h_{opt})$ with a kernel of order $2r$ attains the same relative error size as the residual-based estimator $\hat{\sigma}^2_{\mathrm{HM}}(h_{opt})$ with a kernel of order r, given that the mean function is smooth enough. We note that a smaller relative error does not imply a better finite sample performance, especially for those estimators depending on a subjective choice of smoothing parameter, a fact pointed out by various authors previously. See Dette et al. (1998) and Tong and Wang (2005) for more comparisons.

5. Proofs

This section includes the proofs of (12), (13), and (17) in Theorems 3.1 and 3.2.

Proof of (12). Note that $\mathrm{tr}(\mathbf{D}) = 1$. Then from (11):

$$\text{bias} = \mathbf{g}^T\mathbf{D}\mathbf{g} = \frac{\sum_{i=1}^n \sum_{j\neq i} w_{ij}(g(x_j) - g(x_i))^2}{2W} \triangleq \frac{B}{2W}, \tag{22}$$

where

$$B = \sum_{i=1}^n \sum_{j=1}^n w_{ij}(g(x_j) - g(x_i))^2 - \sum_{i=1}^n w_{ii}(g(x_i) - g(x_i))^2 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}(g(x_j) - g(x_i))^2.$$

Let $\delta(x) = \sum_{j=1}^n v_j(g(x_j) - g(x))^2$, where $v_j = K_h(x_j - x)$. Then $B = \sum_{i=1}^n \delta(x_i)$. Thus to calculate B, it is sufficient to figure out $\delta(x)$:

$$\delta(x) = \sum_{j=1}^n v_j(g(x_j) - g(x))^2 = g^2(x)\sum_{j=1}^n v_j - 2g(x)\sum_{j=1}^n v_j g(x_j) + \sum_{j=1}^n v_j g^2(x_j). \tag{23}$$

Let $z = (y - x)/h$. For $x \in [h, 1-h]$, we have:

$$\begin{aligned}\sum_{j=1}^n v_j &= \sum_{j=1}^n K_h(x_j - x) = nh^{-1}\int_0^1 K\left(\frac{y-x}{h}\right)dy + O(1)\\ &= n\int_{-1}^1 K(z)dz + O(1) = n + O(1),\end{aligned}$$

and

$$\begin{aligned}\sum_{j=1}^n v_j g^k(x_j) &= nh^{-1}\int_0^1 K\left(\frac{y-x}{h}\right)g^k(y)dy + O(1)\\ &= n\int_{-1}^1 K(z)g^k(x+hz)dz + O(1)\end{aligned}$$

$$= n\int_{-1}^{1} K(z)\left\{g^k(x) + hz(g^k(x))' + \cdots + \frac{h^r z^r}{r!}(g^k(x))^{(r)} + o(h^r)\right\} dz + O(1)$$

$$= ng^k(x) + \frac{nh^r}{r!}(g^k(x))^{(r)}\mu_r(K) + o(nh^r) + O(1), \tag{24}$$

where $\mu_r(K) \neq 0$ since K is of order r. For x on the boundary, we use a right-skewed boundary kernel on $[0, h]$ and a left-skewed kernel on $[1-h, 1]$, as discussed in Hall and Marron (1990), Wand and Jones (1995), and Eubank (1999). For $k = 1$ and $k = 2$, plugging (24) into (23), we get:

$$\begin{aligned}\delta(x) &= ng^2(x) - 2g(x)\left\{ng(x) + \frac{nh^r}{r!}g^{(r)}(x)\mu_r(K)\right\} + ng^2(x)\\ &\quad + \frac{nh^r}{r!}(g^2(x))^{(r)}\mu_r(K) + o(nh^r) + O(1)\\ &= \frac{nh^r}{r!}\mu_r(K)\left\{(g^2(x))^{(r)} - 2g(x)g^{(r)}(x)\right\} + o(nh^r) + O(1).\end{aligned}$$

Then:

$$\begin{aligned}B &= \sum_{i=1}^{n}\delta(x_i) = \sum_{i=1}^{n}\left\{\frac{nh^r}{r!}\mu_r(K)((g^2(x))^{(r)} - 2g(x)g^{(r)}(x)) + o(nh^r) + O(1)\right\}\\ &= \frac{n^2h^r}{r!}\mu_r(K)\int_0^1((g^2(x))^{(r)} - 2g(x)g^{(r)}(x))dx + o(n^2h^r) + O(n).\end{aligned}$$

Similarly:

$$W = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} - \sum_{i=1}^{n} w_{ii} = n^2 + O(n) - nh^{-1}K(0).$$

Thus, we have (12).

To proof (13), we present a lemma first.

Lemma 5.1. *Under the same conditions as in Theorem* 3.1, *we have*:

(a) $\mathbf{g}^T\mathbf{D}^2\mathbf{g} = O(n^{-1}h^{2r}) + O(n^{-3})$.
(b) $\mathbf{g}^T\{\mathbf{D}\,\mathrm{diag}(\mathbf{D})\mathbf{1}\} = O(n^{-1}h^r) + O(n^{-2})$.
(c) $\mathrm{tr}\{\mathrm{diag}(\mathbf{D})^2\} = n^{-1} + O(n^{-2})$.
(d) $\mathrm{tr}(\mathbf{D^2}) = n^{-1} + (n^2h)^{-1}\|K\|_2^2 + o(n^{-2}h^{-1})$.

Proof. (a) First by the symmetry of D, we have:

$$\mathbf{g}^T\mathbf{D}^2\mathbf{g} = \mathbf{g}^T\mathbf{D}^T\mathbf{D}\mathbf{g} = (\mathbf{D}\mathbf{g})^T\mathbf{D}\mathbf{g} \triangleq \mathbf{N}^T\mathbf{N}, \tag{25}$$

where

$$\mathbf{N} = \mathbf{Dg} = \frac{1}{W}\begin{pmatrix} \sum_{j=1}^n w_{1j}g(x_1) - \sum_{j=1}^n w_{1j}g(x_j) \\ \sum_{j=1}^n w_{2j}g(x_2) - \sum_{j=1}^n w_{2j}g(x_j) \\ \vdots \\ \sum_{j=1}^n w_{nj}g(x_n) - \sum_{j=1}^n w_{nj}g(x_j) \end{pmatrix}.$$

Similar calculations as those in deriving the bias lead to:

$$\begin{aligned} W\mathbf{N} &= \begin{pmatrix} \sum_{j=1}^n w_{1j}g(x_1) \\ \sum_{j=1}^n w_{2j}g(x_2) \\ \vdots \\ \sum_{j=1}^n w_{nj}g(x_n) \end{pmatrix} - \begin{pmatrix} \sum_{j=1}^n w_{1j}g(x_j) \\ \sum_{j=1}^n w_{2j}g(x_j) \\ \vdots \\ \sum_{j=1}^n w_{nj}g(x_j) \end{pmatrix} \\ &= \begin{pmatrix} ng(x_1) \\ ng(x_2) \\ \vdots \\ ng(x_n) \end{pmatrix} - \begin{pmatrix} ng(x_1) + \frac{nh^r}{r!}g^{(r)}(x_1)\mu_r(K) \\ ng(x_2) + \frac{nh^r}{r!}g^{(r)}(x_2)\mu_r(K) \\ \vdots \\ ng(x_n) + \frac{nh^r}{r!}g^{(r)}(x_n)\mu_r(K) \end{pmatrix} + \mathbf{o}(nh^r) + \mathbf{O}(1) \\ &= -\frac{nh^r}{r!}\mu_r(K)\mathbf{g^{(r)}} + \mathbf{o}(nh^r) + \mathbf{O}(1), \end{aligned}$$

where $\mathbf{g^{(r)}} = (g^{(r)}(x_1), \ldots, g^{(r)}(x_n))^T$, $\mathbf{o}(nh^r) = (o(nh^{(r)}), \ldots, o(nh^{(r)}))^T$ and $\mathbf{O}(1) = (O(1), \ldots, O(1))^T$. Using the fact that $W = n^2 + O(nh^{-1})$, we have:

$$\mathbf{N} = -\frac{h^r}{nr!}\mu_r(K)\mathbf{g^{(r)}} + \mathbf{o}(n^{-1}h^r) + \mathbf{O}(n^{-2}).$$

Therefore,

$$\begin{aligned} \mathbf{g}^T\mathbf{D}^2\mathbf{g} &= \frac{h^{2r}}{n^2(r!)^2}\mu_r^2(K)\sum_{i=1}^n g^{(r)}(x_i)^2 + o(n^{-1}h^{2r}) + O(n^{-3}) \\ &= \frac{h^{2r}}{n(r!)^2}\mu_r^2(K)\int_0^1 g^{(r)}(x)^2dx + o(n^{-1}h^{2r}) + O(n^{-3}) \\ &= O(n^{-1}h^{2r}) + O(n^{-3}). \end{aligned}$$

(b)

$$\begin{aligned} \mathbf{g}^T\{\mathbf{D}\mathrm{diag}(\mathbf{D})\mathbf{1}\} &= (\mathbf{Dg})^T\{\mathrm{diag}(\mathbf{D})\mathbf{1}\} = \mathbf{N}^T\{\mathrm{diag}(\mathbf{D})\mathbf{1}\} \\ &= \left\{-\frac{h^r}{nr!}\mu_r(K)\mathbf{g^{(r)}} + \mathbf{o}(n^{-1}h^r) + \mathbf{O}(n^{-2})\right\}\frac{1}{W}\begin{pmatrix} \sum_{j\neq 1} w_{1j} \\ \vdots \\ \sum_{j\neq n} w_{nj} \end{pmatrix} \end{aligned}$$

$$= -\frac{h^r}{nr!}\mu_r(K)(g^{(r)}(x_1), \ldots, g^{(r)}(x_n)) \begin{pmatrix} n^{-1} \\ \vdots \\ n^{-1} \end{pmatrix} + o(n^{-1}h^r) + O(n^{-2})$$

$$= -\frac{h^r}{nr!}\mu_r(K) \int_0^1 g^{(r)}(x)dx + o(n^{-1}h^r) + O(n^{-2})$$

$$= O(n^{-1}h^r) + O(n^{-2}).$$

(c) For any $i = 1, \ldots, n$, $\sum_{j \neq i} w_{ij} = \sum_{j=1}^n w_{ij} - K_h(0) = n - h^{-1}K(0) + O(1)$. Thus:

$$\begin{aligned} \mathrm{tr}\{\mathrm{diag}(\mathbf{D})^2\} &= \frac{\sum_{i=1}^n (\sum_{j \neq i} w_{ij})^2}{W^2} = \frac{\sum_{i=1}^n (\sum_{j \neq i} w_{ij})^2}{(\sum_{i=1}^n \sum_{j \neq i} w_{ij})^2} \\ &= \frac{\sum_{i=1}^n (n - h^{-1}K(0) + O(1))^2}{\left\{\sum_{i=1}^n (n - h^{-1}K(0) + O(1))\right\}^2} \\ &= \frac{n^3 - 2n^2h^{-1}K(0) + nh^{-2}K^2(0) + O(n^2)}{n^4 - 2n^3h^{-1}K(0) + n^2h^{-2}K^2(0) + O(n^3)} \\ &= n^{-1} + O(n^{-2}). \end{aligned}$$

(d) First:

$$\mathrm{tr}(\mathbf{D^2}) = \frac{\sum_{i=1}^n \left((\sum_{j \neq i} w_{ij})^2 + \sum_{j \neq i} w_{ij}^2\right)}{W^2} = \mathrm{tr}\{\mathrm{diag}(\mathbf{D})^2\} + \frac{\sum_{i=1}^n \sum_{j \neq i} w_{ij}^2}{W^2}.$$

To calculate $\mathrm{tr}(\mathbf{D^2})$, we only need to figure out the second term. Similar calculations as in the proof of (12) lead to:

$$\begin{aligned} \frac{\sum_{i=1}^n \sum_{j \neq i} w_{ij}^2}{W^2} &= \frac{1}{W^2}\left\{\sum_{i=1}^n \sum_{j=1}^n w_{ij}^2 - nh^{-2}K^2(0)\right\} \\ &= \frac{1}{W^2}\left\{\sum_{i=1}^n \left(nh^{-1}\int_{-1}^1 K^2(z)dz + o(nh^{-1})\right) - nh^{-2}K^2(0)\right\} \\ &= (n^2h)^{-1}\|K\|_2^2 + o(n^{-2}h^{-1}). \end{aligned} \tag{26}$$

Therefore,

$$\mathrm{tr}(\mathbf{D^2}) = n^{-1} + (n^2h)^{-1}\|K\|_2^2 + o(n^{-2}h^{-1}).$$

Proof of (13). The last four terms in (11) make up the variance. Using Lemma 5.1 and the fact that $\sigma^4(\gamma_4 - 3) = \mathrm{var}(\epsilon^2) - 2\sigma^4$, it is easy to see that:

$$\begin{aligned} \mathrm{var}(\tilde{\sigma}^2_{\mathrm{MSW}}) &= 4\sigma^2\mathbf{g}^T\mathbf{D}^2\mathbf{g} + 4\mathbf{g}^T\{\mathbf{D}\mathrm{diag}(\mathbf{D})\mathbf{1}\}\sigma^3\gamma_3 + \sigma^4\mathrm{tr}\{\mathrm{diag}(\mathbf{D})^2\}(\gamma_4 - 3) + 2\sigma^4\mathrm{tr}(\mathbf{D^2}) \\ &= n^{-1}\mathrm{var}(\epsilon^2) + (n^2h)^{-1}C_2 + O(n^{-1}h^r) + o(n^{-2}h^{-1}), \end{aligned}$$

where C_2 is defined in Theorem 3.1.

Proof of (17). As in Müller et al. (2003), we write $\tilde{\sigma}^2_{\mathrm{MSW}}(h_{opt})$ as the sum of three parts, $\tilde{\sigma}^2_{\mathrm{MSW}}(h_{opt}) = U_1 + U_2 + U_3$, where

$$U_1 = \frac{1}{2W}\sum_{i=1}^{n}\sum_{j\neq i}(\epsilon_i - \epsilon_j)^2 w_{ij},$$

$$U_2 = \frac{1}{W}\sum_{i=1}^{n}\sum_{j\neq i}(\epsilon_i - \epsilon_j)(g(x_i) - g(x_j))w_{ij},$$

$$U_3 = \frac{1}{2W}\sum_{i=1}^{n}\sum_{j\neq i}(g(x_i) - g(x_j))^2 w_{ij}.$$

From (22), U_3 is exactly the bias term. Thus, $U_3 = O(h^r_{opt}) = O(n^{-2r/(2r+1)})$. Since w_{ij} is symmetric:

$$U_2 = \frac{2}{W}\sum_{i=1}^{n}\epsilon_i\Delta_i,$$

where $\Delta_i = \sum_{j\neq i}(g(x_i) - g(x_j))w_{ij}$. Similar to the proof of Lemma 5.1(a), we have $\Delta_i = g(x_i)\sum_{j=1}^{n} w_{ij} - \sum_{j=1}^{n} w_{ij}g(x_j) = O(nh^r_{opt})$. Since $W = n^2 + o(n^2)$, we have:

$$E(U_2^2) = \frac{4\sigma^2}{W^2}\sum_{i=1}^{n}\Delta_i^2 = O(n^{-1}h^{2r}),$$

which implies that $U_2 = O_p(n^{-1/2}h^r_{opt})$. We can further decompose U_1 as:

$$U_1 = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + T - S,$$

where $S = \sum_{i=1}^{n}\sum_{j\neq i}\epsilon_i\epsilon_j w_{ij}/W$ and $T = \sum_{i=1}^{n}\epsilon_i^2\big(\sum_{j\neq i} w_{ij} - W/n\big)/W$. From (26) it is easy to see that:

$$E(S^2) = \frac{2\sigma^4}{W^2}\sum_{i=1}^{n}\sum_{j\neq i} w_{ij}^2 = O(n^{-2}h^{-1}_{opt}),$$

which implies that $S = O_p(n^{-1}h^{-1/2}_{opt}) = O_p(n^{-2r/(2r+1)})$. Similar to the proof of Lemma 5.1(c), we have $\sum_{j\neq i} w_{ij} - W/n = O(1)$. Thus:

$$E(T^2) = \frac{1}{W^2}O(n^2) = O(n^{-2}),$$

which implies that $T = O_p(n^{-1})$. Therefore, $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\epsilon_i^2 + O_p(n^{-2r/(2r+1)})$.

Acknowledgment

This research was supported by NIH Grant R01 GM58533.

References

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

Carter, C. K., Eagleson, G. K. (1992). A comparison of variance estimations in nonparametric regression. *J. Roy. Statist. Soc. B* 54:773–780.

Dette, H., Munk, A., Wagner, T. (1998). Estimating the variance in nonparametric regression – what is a reasonable choice? *J. Roy. Statist. Soc. B* 60:751–764.

Eagleson, G. K. (1989). Curve estimation -– whatever happened to the variance? *Proc. 47th Sess. Int. Statist. Inst.* 535–551.

Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Dekker.

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

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

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

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

Müller, U., Schick, A., Wefelmeyer, W. (2003). Estimating the error variance in nonparametric 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. *J. Roy. Statist. Soc. B* 67:19–41.

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

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

von Neumann, J. (1941). Distribution of the ratio of the mean squared successive difference to the variance. *Ann. Mathemat. Statist.* 12:367–395.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. B* 40:364–372.

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

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