Estimating the error variance in nonparametric regression by a covariate-matched U-statistic

Ursula U. Müller, Anton Schick, and Wolfgang Wefelmeyer

Abstract. For nonparametric regression models with fixed and random design, two classes of estimators for the error variance have been introduced: second sample moments based on residuals from a nonparametric fit, and difference-based estimators. The former are asymptotically optimal but require estimating the regression function; the latter are simple but have larger asymptotic variance. For nonparametric regression models with random covariates, we introduce a class of estimators for the error variance that are related to difference-based estimators: covariate-matched U-statistics. We give conditions on the random weights involved that lead to asymptotically optimal estimators of the error variance. Our explicit construction of the weights uses a kernel estimator for the covariate density.

AMS 2000 subject classifications. Primary 62G08; secondary 62G07, 62G20.

Key words and Phrases. Empirical estimator, i.i.d. representation, efficient estimator, kernel estimator, relative mean square errors, cross validation.

1. Introduction

Let X and ε be independent random variables, and $Y = r(X) + \varepsilon$ for some unknown smooth regression function r. Assume that ε has mean zero and finite fourth moment. Denote the distribution function of ε by F. We observe independent copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y) and want to estimate the error variance $\sigma^2 = \int x^2 \, dF(x)$.

If the regression function r were known, then the errors $\varepsilon_i = Y_i - r(X_i)$ were observable, and we could estimate the error variance σ^2 with the second sample moment

$$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2. \tag{1.1}$$

Address for correspondence: Wolfgang Wefelmeyer, Universität Siegen, Fachbereich 6 Mathematik, Walter-Flex-Str. 3, 57068 Siegen, Germany.
E-mail: wefelmeyer@mathematik.uni-siegen.de.

Alternatively, we could use the sample variance based on the errors. The sample variance can be written as the U-statistic

$$\frac{1}{n(n-1)}\sum\sum_{i\neq j}\frac{1}{2}(\varepsilon_i-\varepsilon_j)^2 \tag{1.2}$$

and is asymptotically equivalent to the second sample moment $\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$.

We do not know the regression function. But the above two estimators of σ^2 for known r each suggest an estimator of σ^2 for unknown r. The estimator for unknown r suggested by (1.1) is the second sample moment based on residuals from a nonparametric fit,

$$\frac{1}{n}\sum_{i=1}^n (Y_i-\hat r(X_i))^2,$$

where $\hat r$ is an estimator of the regression function. For fixed design see Wahba (1978), Carter and Eagleson (1992) and Carter, Eagleson and Silverman (1992), who use a spline estimator for r, and Hall and Carroll (1989) and Hall and Marron (1990), who use a kernel estimator for r and also treat random designs. Generalizations to heteroscedastic regression with random covariates are studied in Neumann (1994), Stadtmüller and Tsybakov (1995) and Ruppert, Wand, Holst and Hössjer (1997).

For random covariates, Hall and Marron (1990) prove that $\frac{1}{n}\sum_{i=1}^n (Y_i-\hat r(X_i))^2$, with $\hat r$ a kernel estimator, has asymptotic variance

$$\tau^2=\int x^4\,dF(x)-\sigma^4. \tag{1.3}$$

This is the asymptotic variance of $\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$. Hence their estimator has minimal asymptotic variance. Müller, Schick and Wefelmeyer (2001) show that, more precisely, Hall and Marron's estimator is stochastically equivalent to $\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$. This implies that the estimator is asymptotically normal and efficient in the sense of Hájek and Le Cam: it is a least dispersed regular estimator and has minimal asymptotic risk for all bounded bowl-shaped and symmetric loss functions.

Motivated by the U-statistic representation (1.2) of the sample variance, we introduce the covariate-matched U-statistic

$$U=\frac{1}{n(n-1)}\sum\sum_{i\neq j}\frac{1}{2}(Y_i-Y_j)^2\,W_{ij}, \tag{1.4}$$

where the random weights W_{ij} will be based on the covariates only and will be chosen small or zero if X_i and X_j are not close. Our estimator U is related to difference-based estimators for fixed design $X_i=i/n$, for which X_i and X_j are close if the indices i and j are close. Rice (1984) and Gasser, Sroka and Jennen-Steinmetz (1986) have introduced the estimators

$$\hat\sigma_R^2=\frac{1}{2(n-1)}\sum_{i=2}^n (Y_i-Y_{i-1})^2 \quad\text{and}\quad \hat\sigma_{GSJ}^2=\frac{1}{6(n-2)}\sum_{i=3}^n (Y_i+Y_{i-2}-2Y_{i-1})^2.$$

See also Buckley, Eagleson and Silverman (1988), Buckley and Eagleson (1989) and Ullah and Zinde-Walsh (1992). Hall, Kay and Titterington (1990) consider higher-order estimators

$$\frac{1}{n-r}\sum_{i=m_1+1}^{n-m_2}\Big(\sum_{j=-m_1}^{m_2} d_j Y_{i+j}\Big)^2.$$

These estimators have asymptotic variances larger than τ^2. For comparisons see Seifert, Gasser and Wolf (1993) and Dette, Munk and Wagner (1998, 1999).

In Section 2 we determine properties of the weights W_{ij} which guarantee that the covariate-matched U-statistic U behaves asymptotically like the sample second moment based on the errors, i.e., that it has the i.i.d. representation

$$U = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + o_p(n^{-1/2}). \tag{1.5}$$

In particular, unlike the difference-based estimators mentioned above, our estimator U is efficient for σ^2.

In Section 3 we construct explicit weights W_{ij}. They require kernel estimators for the covariate density. This is the price we pay for efficiency. However, we do not need to estimate the regression function, as required for the traditional residual-based efficient estimator. In particular, we get by with weaker assumptions on the regression function: r is assumed Hölder with exponent larger than $1/4$; the corresponding result for $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat r(X_i))^2$ in Müller, Schick and Wefelmeyer (2001) requires an exponent larger than $1/2$.

In Section 4 we compare the small-sample behavior of our estimator with that of the estimators $\hat\sigma_R^2$ of Rice (1984) and $\hat\sigma_{GSJ}^2$ of Gasser, Sroka and Jennen-Steinmetz (1986). Our results carry over to fixed designs under appropriate conditions on the asymptotic behavior of the design. In this case we have independent, but not identically distributed observations, and the proof of efficiency must be rewritten. We refer to McNeney and Wellner (2000) for general results in this setting.

2. The asymptotic behavior of the covariate-matched U-statistic

In this section we study the asymptotic behavior of the covariate-matched U-statistic U introduced in (1.4). We make the following assumptions on the error distribution and the weights.

Assumption 1. The error variable ε is centered and possesses a finite fourth moment:

$$\int x\,dF(x) = 0 \quad\text{and}\quad \int x^4\,dF(x) < \infty.$$

ASSUMPTION 2. The weights W_{ij} depend on the covariates but not on the errors, and they are non-negative, symmetric, and average to one:

$$W_{ij} \geq 0, \quad i,j = 1,\dots,n, \quad i \neq j; \tag{2.1}$$

$$W_{ij} = W_{ji}, \quad i,j = 1,\dots,n, \quad i \neq j; \tag{2.2}$$

$$\frac{1}{n(n-1)} \sum\sum_{i \neq j} W_{ij} = 1. \tag{2.3}$$

For later use we set

$$\begin{aligned} \overline{W}_i &= \frac{1}{n-1} \sum_{j:j\neq i} W_{ij}, \quad i = 1,\dots,n; \\ \Delta_i &= \frac{1}{n-1} \sum_{j:j\neq i} (r(X_i) - r(X_j)) W_{ij}, \quad i = 1,\dots,n. \end{aligned}$$

The following theorem gives conditions under which U behaves asymptotically like the average of the squared errors.

THEOREM 1. *Suppose that Assumptions 1 and 2 hold and that*

$$\frac{1}{n(n-1)} \sum\sum_{i \neq j} W_{ij}^2 = o_p(n); \tag{2.4}$$

$$\frac{1}{n} \sum_{i=1}^{n} (\overline{W}_i - 1)^2 = o_p(1); \tag{2.5}$$

$$\frac{1}{n} \sum_{i=1}^{n} \Delta_i^2 = o_p(1); \tag{2.6}$$

$$\frac{1}{n(n-1)} \sum\sum_{i \neq j} (r(X_i) - r(X_j))^2 W_{ij} = o_p(n^{-1/2}). \tag{2.7}$$

Then

$$U = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 + o_p(n^{-1/2}).$$

In particular, $n^{1/2}(U - \sigma^2)$ converges in distribution to a normal random variable with mean zero and variance $\tau^2 = \int x^4 \, dF(x) - \sigma^4$.

PROOF. We can write U as the sum $U_1 + U_2 + U_3$ of the three U-statistics

$$\begin{aligned} U_1 &= \frac{1}{n(n-1)} \sum\sum_{i \neq j} \frac{1}{2}(\varepsilon_i - \varepsilon_j)^2 W_{ij}, \\ U_2 &= \frac{1}{n(n-1)} \sum\sum_{i \neq j} (\varepsilon_i - \varepsilon_j)(r(X_i) - r(X_j)) W_{ij}, \\ U_3 &= \frac{1}{n(n-1)} \sum\sum_{i \neq j} \frac{1}{2}(r(X_i) - r(X_j))^2 W_{ij}. \end{aligned}$$

We have $U_3 = o_p(n^{-1/2})$ by (2.7). Let us now show that $U_2 = o_p(n^{-1/2})$. Using the symmetry (2.2) of the weights we can write

$$U_2 = \frac{2}{n} \sum_{i=1}^{n} \varepsilon_i \Delta_i.$$

Thus, by assumption (2.6),

$$nE[U_2^2 \mid X_1, \ldots, X_n] = 4\sigma^2 \frac{1}{n} \sum_{i=1}^{n} \Delta_i^2 = o_p(1).$$

This implies the desired $U_2 = o_p(n^{-1/2})$.

Using again the symmetry of the weights, we obtain that

$$U_1 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \overline{W}_i - S = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 + T - S,$$

where

$$\begin{aligned} S &= \frac{1}{n(n-1)} \sum\sum_{i \neq j} \varepsilon_i \varepsilon_j W_{ij}, \\ T &= \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 (\overline{W}_i - 1) = \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i^2 - \sigma^2)(\overline{W}_i - 1), \end{aligned}$$

with the last identity a consequence of assumption (2.3). By assumption (2.5) we have

$$nE[T^2 \mid X_1, \ldots, X_n] \leq \int x^4 \, dF(x) \, \frac{1}{n} \sum_{i=1}^{n} (\overline{W}_i - 1)^2 = o_p(1).$$

This yields $T = o_p(n^{-1/2})$. By assumption (2.4) we have

$$nE[S^2 \mid X_1, \ldots, X_n] = \frac{2\sigma^4}{n-1} \frac{1}{n(n-1)} \sum\sum_{i \neq j} W_{ij}^2 = o_p(1),$$

which yields $S = o_p(n^{-1/2})$. Thus we have $U_1 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 + o_p(n^{-1/2})$. This completes the proof. □

REMARK 1. We did not use requirement (2.1) in the above proof. However, it is natural to impose this requirement as it guarantees that our estimator is non-negative. Inspection of the above proof also shows that assumption (2.3) is only used to conclude the second identity for T. But it is enough to have this identity only up to a term of order $o_p(n^{-1/2})$. Thus we can relax (2.3) to

$$\frac{1}{n(n-1)}\sum\sum_{i\neq j} W_{ij} = 1 + o_p(n^{-1/2}). \tag{2.8}$$

The latter is implied by

$$P\Big(\frac{1}{n(n-1)}\sum\sum_{i\neq j} W_{ij} = 1\Big) \to 1. \tag{2.9}$$

REMARK 2. In view of assumptions (2.1), a sufficient condition for (2.4) is (2.8) and

$$\max_{i,j} W_{ij} = o_p(n). \tag{2.10}$$

Indeed, we can bound the left-hand side of (2.4) by the product of the left-hand sides of (2.8) and (2.10). A sufficient condition for (2.5) is

$$\max_i |\overline{W}_i - 1| = o_p(1). \tag{2.11}$$

An application of the Cauchy–Schwarz inequality shows that (2.11) and (2.7) imply (2.6).

REMARK 3. Suppose now that the regression function r is Hölder with constant H and exponent β:

$$|r(s) - r(t)| \le H|s-t|^\beta, \quad s,t \in \mathbb{R}.$$

In addition to Assumption 2 let the weights satisfy

$$W_{ij} = 0 \quad \text{if} \quad |X_i - X_j| > b_n, \quad i,j = 1,\dots,n, \quad i \neq j, \tag{2.12}$$

for a bandwidth b_n tending to zero. Then $|\Delta_i| \le H b_n^\beta \overline{W}_i$, and (2.6) follows from this and (2.5). We can also bound the left-hand side of (2.7) by $H^2 b_n^{2\beta}$ and obtain (2.7) if $n^{1/2} b_n^{2\beta} \to 0$. However, to satisfy the other properties such as (2.4) and (2.5), the bandwidth will need to satisfy $nb_n \to \infty$. Thus, we will need at least $\beta > 1/4$ to obtain (2.7).

3. Construction of weights

In this section we construct weights W_{ij} explicitly. We impose the following additional assumptions.

ASSUMPTION 3. The covariate X takes values in the interval $[0,1]$ and possesses a density g whose restriction to $[0,1]$ is continuous and positive.

ASSUMPTION 4. The regression function r satisfies the Hölder condition

$$|r(s) - r(t)| \le H|s-t|^\beta, \quad s,t \in [0,1],$$

for some finite constant H and some positive β with $\beta \le 1$.

We shall now construct non-negative symmetric weights which fulfill the requirements (2.4) to (2.7) and (2.8). For this let K be a bounded symmetric density with compact support $[-1,1]$, and let b_n be a bandwidth. Then our choice is

$$W_{ij} = \frac{1}{2}\left(\frac{1}{\hat{g}_i} + \frac{1}{\hat{g}_j}\right) K_n(X_i - X_j), \tag{3.1}$$

where $K_n(x) = K(x/b_n)/b_n$ and

$$\hat{g}_i = \frac{1}{n-1}\sum_{j:j\neq i} K_n(X_i - X_j), \quad i = 1, \ldots, n.$$

Actually, these weights are not well defined on the event $\{\min_i \hat{g}_i = 0\}$. We shall see later that this event has probability going to zero for our choice of bandwidth. Thus we can redefine the weights on this event without affecting the asymptotics. For example, we may take the weights on this event to correspond to a larger bandwidth for which all $\hat{g}_i$ are positive.

We denote the estimator corresponding to the above weights by $\hat{\sigma}^2$, so that

$$\hat{\sigma}^2 = \frac{1}{n(n-1)}\sum\sum_{i\neq j} \frac{1}{2}(Y_i - Y_j)^2 \frac{1}{2}\left(\frac{1}{\hat{g}_i} + \frac{1}{\hat{g}_j}\right) K_n(X_i - X_j).$$

THEOREM 2. *Suppose Assumptions 1, 3 and 4 hold. Let the bandwidth b_n satisfy $n^{1/2}b_n^{2\beta} \to 0$ and $nb_n/\log n \to \infty$. Then*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 + o_p(n^{-1/2}).$$

PROOF. Clearly, our weights are non-negative and symmetric. Moreover, they average to one except on the event $\{\min_j \hat{g}_j = 0\}$. But the probability of this event tends to zero, see (3.2) below. Since we have (2.1), (2.2) and (2.9), we need only verify (2.4) to (2.7) for our weights (3.1). It follows from Remark 3 and the choice of bandwidth that (2.6) and (2.7) hold. We are left to verify (2.4) and (2.5). To this end let us define

$$\overline{g}_n(x) = \int K_n(x-t)g(t)\,dt = \int g(x - b_n t)K(t)\,dt, \quad x \in \mathbb{R}.$$

The key to verifying the remaining two conditions (2.4) and (2.5) will be the fact that

$$V = \max_i |\hat{g}_i - \overline{g}_n(X_i)| = o_p(1).$$

But this can be established by applications of the Bernstein inequality (Serfling, 1980, p. 95, Lemma A).

It follows from Assumption 3 that the density g is bounded and bounded away from zero on $[0,1]$. Thus there are constants $0 < a < A < \infty$ such that $a \le g(x) \le A$ for all $x \in [0,1]$. This lets us conclude that $\frac{1}{2}a \le \overline{g}_n(x) \le A$ for all $x \in [0,1]$ if $b_n < 1/2$. Thus we immediately obtain that

$$N = \max_i \hat{g}_i = O_p(1) \quad \text{and} \quad M = \max_i \frac{1}{\hat{g}_i} = O_p(1). \tag{3.2}$$

From the latter and the properties of K and b_n we obtain (2.10) and hence (2.4) as shown in Remark 2. Easy calculations show that on the event $\{\min_j \hat{g}_j > 0\}$,

$$\begin{aligned} |\overline{W}_i - 1| &\le \frac{1}{n-1}\sum_{j:j\neq i}\Big|\frac{1}{\hat{g}_i}-\frac{1}{\hat{g}_j}\Big| K_n(X_i - X_j) \\ &\le M^2\frac{1}{n-1}\sum_{j:j\neq i}|\hat{g}_i - \hat{g}_j|\, K_n(X_i - X_j) \\ &\le M^2\frac{1}{n-1}\sum_{j:j\neq i}|\overline{g}_n(X_i) - \overline{g}_n(X_j)|\, K_n(X_i - X_j) + 2M^2VN. \end{aligned}$$

Thus (2.5) follows if we show that

$$B = \frac{1}{n}\sum_{i=1}^n \Big(\frac{1}{n-1}\sum_{j:j\neq i}|\overline{g}_n(X_i) - \overline{g}_n(X_j)|\, K_n(X_i - X_j)\Big)^2 = o_p(1).$$

Let w_g denote the modulus of continuity of g restricted to $[0,1]$:

$$w_g(b) = \sup_{x,y\in[0,1],|x-y|\le b}|g(x) - g(y)|, \quad b > 0.$$

For $x, y \in (b_n, 1-b_n)$ with $|y-x| < b_n$, we find that

$$|\overline{g}_n(x) - \overline{g}_n(y)| \le \int |g(x - b_n t) - g(y - b_n t)|\, K(t)\, dt \le w_g(b_n).$$

By Assumption 3 we have $w_g(b_n) \to 0$. We can bound B by

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}[X_i \in (2b_n, 1-2b_n)](w_g(b_n)N)^2 + \frac{1}{n}\sum_{i=1}^n \mathbf{1}[X_i \notin (2b_n, 1-2b_n)](2AN)^2.$$

It is now easy to see that $B = o_p(1)$. □

4. Small sample behavior

In this section we shall take a brief look at the small sample behavior of our estimator and an automatic bandwidth selection based on cross-validation. Our estimator $\hat{\sigma}^2$ is a U-statistics version of the Rice estimator $\hat{\sigma}_R^2$. Gasser, Sroka and Jennen-Steinmetz (1986) show that their estimator $\hat{\sigma}_{GSJ}^2$ has less of a bias problem than $\hat{\sigma}_R^2$ when the signal to noise ratio is large. We have performed a small simulation study in which we compare our estimator with $\hat{\sigma}_R^2$ and $\hat{\sigma}_{GSJ}^2$. To keep it simple, we only look at the case of an equidistant design on $[0,1]$, with design points $i/(n-1)$, $i = 0, \dots, n$, and consider only normally distributed errors. We take $n = 25$ and choose the following three regression functions.

$$r_1(x) = 0.3\exp(-4(4x-1)^2) + 0.7\exp(-16(4x-3)^2), \tag{4.1}$$

$$r_2(x) = \sin(5\pi x), \tag{4.2}$$

$$r_3(x) = \sqrt{|(x-.2)(x-.7)|}. \tag{4.3}$$

For our estimator we take the weights proposed in the previous section with $K(x) = 3(1-x^2)_+/4$ and several choices of bandwidths, namely $b = .05$, .10, .15, .20 and .25. Because of the equidistant design our estimator with bandwidth $b = .05$ equals the Rice estimator $\hat{\sigma}_R^2$. We also look at a version of our estimator with a data-driven bandwidth using a cross-validation principle. More precisely, we choose a bandwidth from the five choices above that minimizes

$$\sum_{i=1}^{n} (\hat{\sigma}^2 - \hat{\sigma}_i^2)^2,$$

where $\hat{\sigma}_i^2$ is our the estimator constructed without the i-th pair (X_i, Y_i).

TABLE 1. Relative MSE's for $r = r_1$ and selected values of σ.

σ	$b = .05$	$b = .10$	$b = .15$	$b = .20$	$b = .25$	$\hat{\sigma}_d^2$	$\hat{\sigma}_{GSJ}^2$
0.20	2.29	3.05	4.85	7.09	9.55	3.87	2.21
0.40	1.63	1.53	1.61	1.78	1.98	1.56	2.07
0.70	1.70	1.49	1.41	1.38	1.38	1.33	2.23
1.00	1.56	1.37	1.29	1.25	1.23	1.22	2.04
2.00	1.53	1.32	1.21	1.16	1.13	1.13	2.08
3.00	1.49	1.30	1.19	1.14	1.11	1.13	2.01
5.00	1.55	1.35	1.24	1.18	1.15	1.17	2.10
10.00	1.48	1.29	1.18	1.13	1.11	1.12	1.95

TABLE 2. Relative MSE's for $r = r_2$ and selected values of σ.

σ	$b = .05$	$b = .10$	$b = .15$	$b = .20$	$b = .25$	$\hat{\sigma}_d^2$	$\hat{\sigma}_{GSJ}^2$
0.20	93.41	238.79	588.09	1071.00	1600.35	381.73	4.07
0.40	7.85	17.26	40.14	71.69	106.22	26.32	2.27
0.70	2.33	3.28	5.87	9.50	13.49	5.07	2.11
1.00	1.77	1.89	2.56	3.54	4.61	2.47	2.05
1.25	1.65	1.60	1.83	2.23	2.69	1.82	2.06
1.50	1.55	1.44	1.53	1.73	1.98	1.52	1.97
2.00	1.62	1.44	1.40	1.45	1.54	1.35	2.12
3.00	1.50	1.31	1.24	1.23	1.25	1.18	1.97
5.00	1.52	1.32	1.21	1.18	1.17	1.16	2.06
10.00	1.57	1.37	1.26	1.21	1.18	1.19	2.13

Tables 1 to 3 report the relative mean square errors (RMSE) of these six versions of our estimator and of the Gasser et al. (1986) estimator for selected values of σ. The RMSE is the (simulated) MSE divided by τ^2/n, the variance based on the asymptotic considerations, which is

TABLE 3. Relative MSE's for $r = r_3$ and selected values of σ.

σ	$b = .05$	$b = .10$	$b = .15$	$b = .20$	$b = .25$	$\hat{\sigma}_d^2$	$\hat{\sigma}_{GSJ}^2$
0.10	1.87	2.10	3.12	4.95	7.41	3.08	2.16
0.20	1.64	1.45	1.44	1.59	1.80	1.46	2.21
0.40	1.63	1.42	1.31	1.28	1.29	1.24	2.16
0.70	1.55	1.36	1.25	1.20	1.18	1.18	2.10
1.00	1.53	1.33	1.22	1.16	1.13	1.13	2.04
2.00	1.59	1.37	1.24	1.19	1.16	1.17	2.12
5.00	1.52	1.31	1.20	1.15	1.12	1.13	2.06
10.00	1.57	1.37	1.26	1.21	1.17	1.18	2.13

$2\sigma^4/n$ for our normal errors. The first column in each table lists the selected values of σ. The next five columns give the RMSE's for our estimator with the indicated bandwidth. Since our estimator with bandwidth $b = .05$ equals the Rice estimator, the second column in each table also provides the RMSE's of $\hat{\sigma}_R^2$. The seventh column (labeled $\hat{\sigma}_d^2$) in each table gives the RMSE's of our estimator with the data-driven bandwidth described above. The last column in each table gives the RMSE's of the estimator $\hat{\sigma}_{GSJ}^2$.

When the error variance is very small, our estimator has bias problems similar to the Rice estimator $\hat{\sigma}_R^2$. In the more interesting case of an error variance that is not small, our estimator is better than $\hat{\sigma}_{GSJ}^2$. This holds for quite small sample sizes, even though our estimator requires estimating the covariate density.

Acknowledgments

The research of Anton Schick was partially supported by NSF Grant DMS 0072174.

References

Buckley, M. J. and Eagleson, G. K. (1989). A graphical method for estimating the residual variance in non-parametric regression. *Biometrika* **76**, 203–210.

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

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

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

Dette, H., Munk, A. and Wagner, T. (1999). A review of variance estimators with extensions to multivariate nonparametric regression models. In: *Multivariate Analysis, Design of Experiments, and Survey Sampling* (S. Ghosh, ed.), 469–498, Statistics: Textbooks and Monographs 159, Dekker, New York.

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

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *J. Roy. Statist. Soc. Ser. B. (Methodological)* **51**, 3–14.

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

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

McNeney, B. and Wellner, J. A. (2000). Application of convolution theorems in semiparametric models with non-i.i.d. data. *J. Statist. Plann. Inference* **91**, 441–480.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001). Estimating linear functionals of the error distribution in nonparametric regression. Technical Report, Department of Mathematical Sciences, Binghamton University.
http://math.binghamton.edu/anton/index.html

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

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

Ruppert, D., Wand, M. P., Holst, U. and Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics* **39**, 262–273.

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

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Stadtmüller, U. and Tsybakov, A. B. (1995). Nonparametric recursive variance estimation. *Statistics* **27**, 55–63.

Ullah, A. and Zinde-Walsh, V. (1992). On the estimation of the residual variance in nonparametric regression. *J. Nonparametr. Statist.* **1**, 263–265.

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