Journal of the Korean Statistical Society (2006), 35: 1, pp 105–114

A SIMPLE VARIANCE ESTIMATOR IN NONPARAMETRIC REGRESSION MODELS WITH MULTIVARIATE PREDICTORS†

YOUNG KYUNG LEE[1], TAE YOON KIM[2] AND BYEONG U. PARK[3]

ABSTRACT

In this paper we propose a simple and computationally attractive difference-based variance estimator in nonparametric regression models with multivariate predictors. We show that the estimator achieves $n^{-1/2}$ rate of convergence for regression functions with only a first derivative when d, the dimension of the predictor, is less than or equal to 4. When $d > 4$, the rate turns out to be $n^{-4/(d+4)}$ under the first derivative condition for the regression functions. A numerical study suggests that the proposed estimator has a good finite sample performance.

AMS 2000 subject classifications. Primary 62G08; Secondary 62G20.
Keywords. Variance estimation, multivariate regression, rate of convergence.

1. INTRODUCTION

A homoscedastic regression problem of the form

$$Y_i = m(X_i) + \epsilon_i \quad (1 \le i \le n) \tag{1.1}$$

is considered where m is an unknown regression function, the errors are independent and identically distributed random variables with mean zero and variance τ, and the random design points X_i are assumed to arise from independent realizations of a distribution having a density f on $\mathbb{R}^d$. Recently, estimation of

Received January 2006; accepted March 2006.
†Young Kyung Lee was supported by the Brain Korea 21 Project in 2004. Tae Yoon Kim was supported by KOSEF R01-2003-000-10589-0. Byeong U. Park was supported by KOSEF through Statistical Research Center for Complex Systems at Seoul National University.
[1]Department of Statistics, Seoul National University, Seoul 151-747, Korea
[2]Department of Statistics, Keimyung University, Taegu 704-701, Korea
[3]Corresponding author. Department of Statistics, Seoul National University, Seoul 151-747, Korea (e-mail : bupark@stats.snu.ac.kr)

the variance τ has been getting attention because it is an essential part in constructing confidence intervals for m as well as for many other applications, such as prediction, estimation of detection limits or immunoassay (see *e.g.*, Carroll, 1988; Carroll and Ruppert, 1988). In this paper, we propose a simple variance estimator which has a quite good rate of convergence for regression functions with only a first derivative.

Various estimators of τ have been proposed. Most of variance estimators are quadratic in the vector of responses $Y = (Y_1, \ldots, Y_n)^t$ and have the form:

$$\hat{\tau}_D = \frac{Y^t D Y}{tr(D)}, \tag{1.2}$$

for some symmetric matrix D that depends on X_i. There are two classes of variance estimators. The first contains kernel based estimators (KBE) (*e.g.*, Müller and Stadtmüller, 1987; Hall and Carroll, 1989; Hall and Marron, 1990; Neumann, 1994). They are based on a sum of squared residuals from a nonparametric fit of the regression function, and thus depend on a smoothing parameter. Typically, kernel based regression estimators are linear fits of the form $\hat{Y} = HY$, which leads to the form (1.2) with $D = (I - H)^t(I - H)$.

Difference-based estimators (DBE) are members of the second class. Here, D is a symmetric $n \times n$ non-negative definite matrix which may depends on the predictors X_i but not on the responses Y_i. Let $X_i \in \mathbb{R}$ be ordered so that $X_1 \leq \cdots \leq X_n$. A simple DBE suggested by Rice (1984) is given by

$$\hat{\tau}_{D,1} = \frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_i - Y_{i-1})^2.$$

For equally spaced data, Gasser *et al.* (1986) proposed another DBE

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

which is based on second order differencing. Hall *et al.* (1990) considered a class of DBE which includes $\hat{\tau}_{D,1}$ and $\hat{\tau}_{D,2}$. Let $\{d_j\}$ be a normalized contrast such that $\sum d_j = 0$ and $\sum d_j^2 = 1$ where d_{-m_1} and d_{m_2} are not zero for some nonnegative integers m_1 and m_2 and $d_j = 0$ for all $j < -m_1$ and $j > m_2$. Define $r = m_1 + m_2$ which denotes the order of differencing. A general form of DBE considered by Hall *et al.* (1990) is

$$\hat{\tau}_{D,r} = (n-r)^{-1} \sum_{i=m_2+1}^{n-m_1} \left(\sum_{j=-m_1}^{m_2} d_j Y_{i-j} \right)^2 = (n-r)^{-1} Y^t D Y,$$

where the $(i,j)^{th}$ entry of the matrix D is given by $D_{ij} = \sum_{k=m_2+1}^{n-m_1} d_{k-i}d_{k-j}$. Hall *et al.* (1990) determined the optimal difference sequences $\{d_j\}$ for each r which achieves the minimal mean squared error within the class. They also argued that as r increases the optimal difference sequence converges to a 'spike' of unit mass at one of the entries d_j, and converges to zero everywhere else. Dette *et al.* (1998) provided a higher order analysis of the mean squared error for the class of DBE.

All of the aforementioned works focus on variance estimation in univariate regression model. There are a few works dealing with estimating the variance in multiple regression. Those include Hall *et al.* (1991), Kulasekera and Gallagher (2002) and Spokoiny (2002). Hall *et al.* (1991) developed methods for estimating the variance of white noise in a two-dimensional degraded signal. Kulasekera and Gallagher (2002) extended DBE to the multivariate setting by introducing an algorithm that orders the multivariate predictors. Spokoiny (2002) established minimax-type results for a class of regression functions with second derivatives.

The main advantage of DBE over KBE is that they are computationally much cheaper. Our aim in this paper is to provide a simple and computationally attractive difference-based variance estimator which are applicable for multivariate predictors. Our estimator $\hat{\tau}$ uses differences of paired responses whose concomitant predictors are close to each other. It does not need a complicated procedure of ordering multivariate predictors which the method of Kulasekera and Gallagher (2002) is based on. When the regression function has continuous first partial derivatives, we show $\hat{\tau}$ may be constructed so that as n grows

$$E\left(\hat{\tau} - \tau\right)^2 \simeq c_1 n^{-1} + c_2 n^{-8/(d+4)}, \tag{1.3}$$

for some positive constants c_1 and c_2. Thus, when $d \leq 4$, the estimator achieves root-n rate. When $d > 4$, it estimates τ at the rate $n^{-4/(d+4)}$. Although not presented in this paper, it may be shown, following the arguments of Spokoiny (2002), that one cannot obtain $n^{-1/2}$ consistency under some conditions on the first derivative when $d > 4$.

The finite sample performance of the proposed estimator is also investigated. In a simulation study, the proposed estimator is compared with a KBE. It is observed that the minimal mean squared errors of the proposed estimator at the optimal tuning parameter values are better than those of the KBE for moderate sample sizes. Furthermore, the mean squared error properties of the proposed estimator turn out to be less sensitive to the choice of the tuning parameter than those of the KBE.

2. RESULTS

For a constant $h > 0$, let N_h denote the number of paired indices (i, j) such that $|X_i - X_j| \leq h$, *i.e.*

$$N_h \equiv \sum\sum_{i \neq j} I(|X_i - X_j| \leq h),$$

where $I(A)$ denotes the indicator. We consider the following variance estimator:

$$\hat{\tau}_h = (2N_h)^{-1} \sum\sum_{i \neq j} (Y_i - Y_j)^2 I(|X_i - X_j| \leq h).$$

We assume that m has a continuous first derivative m', that the density f of X_i is continuous, and that $\int f^3(x)\,dx < \infty$ and $\int |m'(x)|^2 f^2(x)\,dx < \infty$. Call these conditions (A). Below, we present conditional mean squared error properties of $\hat{\tau}_h$ under these assumptions.

To state a theorem, define $\nu_0 = \int_{|t| \leq 1} dt$ and $\nu_2 = \int_{|t| \leq 1} t_1^2\,dt = d^{-1} \int_{|t| \leq 1} |t|^2\,dt$ Also, define

$$\begin{aligned}
c_1 &= 2^{-1} \left(\nu_0 \int f^2(x)\,dx \right)^{-1} \nu_2 \int |m'(x)|^2 f^2(x)\,dx, \\
c_2 &= \left(\int f^2(x)\,dx \right)^{-2} \left(\int f^3(x)\,dx \right) \mathrm{Var}(\epsilon^2), \\
c_3 &= \left(\nu_0 \int f^2(x)\,dx \right)^{-1} (E\epsilon^4 + \tau^2).
\end{aligned}$$

THEOREM 1. *Under the conditions (A), it follows that if* $h \to 0$ *and* $n^2 h^d \to \infty$ *as* $n \to \infty$ *then*

$$\begin{aligned}
E(\hat{\tau}_h - \tau | X_1, \ldots, X_n) &= c_1 h^2 + o_p(h^2), \\
\mathrm{Var}(\hat{\tau}_h - \tau | X_1, \ldots, X_n) &= c_2 n^{-1} + c_3 n^{-2} h^{-d} + o_p(n^{-1} + n^{-2} h^{-d}).
\end{aligned}$$

From Theorem 1, it follows that our estimator has conditional mean squared error given by

$$E\{(\hat{\tau}_h - \tau)^2 | X_1, \ldots, X_n\} = c_1^2 h^4 + c_2 n^{-1} + c_3 n^{-2} h^{-d} + o_p(n^{-1} + h^4 + n^{-2} h^{-d}). \quad (2.1)$$

If X_i's have a uniform distribution over a finite interval $[a, b]$, then $c_2 = \mathrm{Var}(\epsilon^2)$. In general, $c_2 \geq \mathrm{Var}(\epsilon^2)$ by Hölder inequality.

The asymptotically optimal h is

$$h_0 = \left[d\,\nu_0\,(E\epsilon^4 + \tau^2) \int f^2(x)\,dx \left\{ \nu_2^2 \left(\int |m'(x)|^2 f^2(x)\,dx \right)^2 n^2 \right\}^{-1} \right]^{1/(d+4)}.$$

With this optimal h_0, the conditional mean squared error at (2.1) equals, for some positive constant c_2',

$$E\{(\hat{\tau}_h - \tau)^2 | X_1, \ldots, X_n\} = c_1' n^{-8/(d+4)} + c_2 n^{-1} + o_p(n^{-1} + n^{-8/(d+4)}).$$

Thus, when $d \leq 4$, our estimator $\hat{\tau}_h$ achieves root-n rate. The rate deteriorates for $d > 4$, in which case the accuracy is $n^{-4/(d+4)}$.

It may be proved by a standard technique (*e.g.*, Brown and Kildea, 1978) that $\hat{\tau}_h$ is asymptotically normally distributed. In fact, if $h \to 0$ and $n^2 h^d \to \infty$ as $n \to \infty$ then

$$\left(c_2 n^{-1} + c_3 n^{-2} h^{-d} \right)^{-1/2} \left(\hat{\tau}_h - \tau - c_1 h^2 + o_p(h^2) \right) \longrightarrow N(0,1).$$

The condition that $h \to 0$ and $n^2 h^d \to \infty$ as $n \to \infty$ is required for consistency of the estimator. Note that our estimator is an average of the squared differences of Y_i's for N_h different pairs of indices. The size of N_h turns out to be $O_p(n^2 h^d)$. This should tend to infinity as sample size grows. Differences of $m(X_i)$'s for pairs of X_i's (within distance h) produces bias, thus h should tends to zero to make the bias vanish as the sample size grows.

As for kernel based methods, we are not aware of any theoretical results in the case of multivariate predictors. For the univariate case where $d = 1$, Hall and Marron (1990) showed that the conditional mean squared error of their KBE with bandwidth b is asymptotic to $b^{4k} + n^{-1} + n^{-2} b^{-1}$ if m has k derivatives and a k^{th} order kernel is used. Under the same condition of Theorem 1, *i.e.*, when $k = 1$, this coincides with the result (2.1) up to constant factors. By an extension of the arguments in Hall and Marron (1990) or those in Park *et al.* (2006), one may show that under the condition of Theorem 1 the conditional mean squared error of the KBE for general d admits the same asymptotic expansion as given at (2.1).

A simulation study is conducted to assess the finite sample performance of the proposed estimator. We compare the mean squared errors of the proposed estimator and a KBE based on the Nadaraya-Watson smoother. For the Nadaraya-Watson regression estimator we used the Gaussian product kernel with single bandwidth. Two regression functions were considered. One is

$m_1(x) = x_1 + x_2^2 + x_3^3$, and the other is $m_2(x) = x_2^2 + x_2^2 x_3$. Other regression functions were also considered, but we found the results were similar to those with the two functions. The three-dimensional predictors were generated from the uniform distribution on the unit cube $[0, 1]^3$. We took 500 pseudo samples of size $n = 100$ and 400.

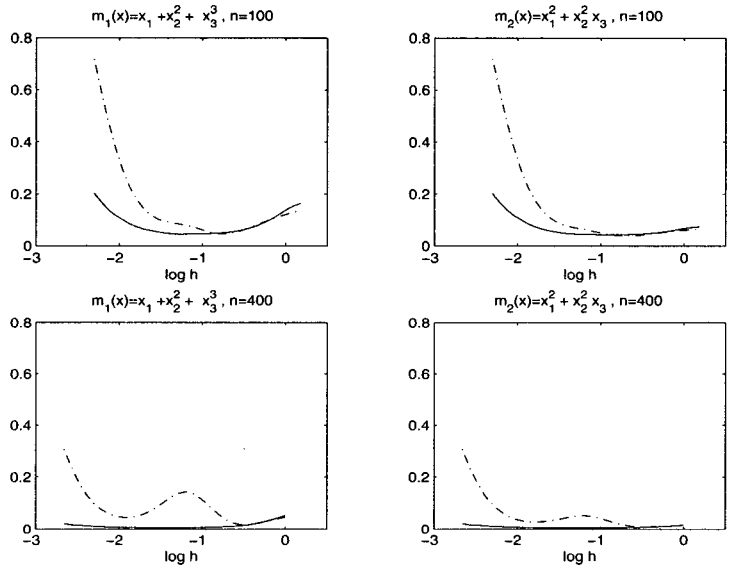

FIGURE 1 *The mean squared errors of the variance estimators as functions of the bandwidth, based on 500 pseudo samples of size 100 and 400. Solid curves correspond to the proposed estimator and the dot-dashed are for the KDE.*

TABLE 1 *Minimal mean squared errors of the variance estimators with their respective optimal bandwidths, based on 500 pseudo samples of size 100 and 400.*

Regression function	*Sample size*	*Proposed*	*KDE*
$m_1(x)$	$n = 100$	.0448	.0453
$= x_1 + x_2^2 + x_3^3$	$n = 400$	.0066	.0187
$m_2(x)$	$n = 100$	.0429	.0382
$= x_2^2 + x_2^2 x_3$	$n = 400$	.0063	.0102

Figure 1 depicts the mean squared errors of the proposed estimator and the KBE as functions of the bandwidth. The minimal mean squared errors of the two estimators at their respective optimal bandwidths, as shown in Table 1, are very close to each other when $n = 100$. However, when $n = 400$ the proposed estimator has better mean squared error properties than the KBE. Furthermore, Figure 1

suggests that the mean squared error properties of the proposed estimator are less dependent on the choice of the bandwidth than those of the KDE.

3. PROOF OF THEOREM 1

Define

$$B_n \equiv (2N_h)^{-1}\sum\sum_{i\neq j}\{m(X_i)-m(X_j)\}^2 I(|X_i-X_j|\leq h),$$
$$S_n \equiv (2N_h)^{-1}\sum\sum_{i\neq j}(\epsilon_i-\epsilon_j)^2 I(|X_i-X_j|\leq h),$$
$$R_n \equiv (N_h)^{-1}\sum\sum_{i\neq j}\{m(X_i)-m(X_j)\}(\epsilon_i-\epsilon_j) I(|X_i-X_j|\leq h).$$

Then, we obtain

$$\hat{\tau}_h - \tau = B_n + (S_n - \tau) + R_n.$$

Since $E(\epsilon_1-\epsilon_2)^2 = 2\tau$, it follows that

$$E(\hat{\tau}_h - \tau | X_1,\ldots,X_n) = B_n, \tag{3.1}$$
$$\begin{aligned}\mathrm{Var}(\hat{\tau}_h - \tau | X_1,\ldots,X_n) &= \mathrm{Var}(S_n|X_1,\ldots,X_n) + \mathrm{Var}(R_n|X_1,\ldots,X_n)\\ &\quad +2\,\mathrm{Cov}(S_n,R_n|X_1,\ldots,X_n).\end{aligned} \tag{3.2}$$

We show that if $h\to 0$ and $n^2h^d\to\infty$ as $n\to\infty$ then

$$B_n = 2^{-1}\left(\nu_0\int f^2(x)\,dx\right)^{-1}\left(\nu_2\int |m'(x)|^2 f^2(x)\,dx\right)h^2 + o_p(h^2), \tag{3.3}$$
$$\begin{aligned}\mathrm{Var}(S_n|X_1,\ldots,X_n) &= \left(\int f^2(x)\,dx\right)^{-2}\left(\int f^3(x)\,dx\right)\mathrm{Var}(\epsilon^2)\,n^{-1}\\ &\quad +\left(\nu_0\int f^2(x)\,dx\right)^{-1}(E\epsilon^4+\tau^2)\,n^{-2}h^{-d}\\ &\quad +o_p(n^{-1}+n^{-2}h^{-d}),\end{aligned} \tag{3.4}$$
$$\mathrm{Var}(R_n|X_1,\ldots,X_n) = o_p(n^{-1}+n^{-2}h^{-d}). \tag{3.5}$$

Define

$$I_{ij} = I(|X_i-X_j|\leq h), \quad M_{ij} = \{m(X_i)-m(X_j)\}I(|X_i-X_j|\leq h).$$

Then, we can show

$$EI_{12} = h^d \nu_0 \int f^2(x)\,dx + o(h^d), \tag{3.6}$$

$$EI_{12}I_{13} = h^{2d} \nu_0^2 \int f^3(x)\,dx + o(h^{2d}), \tag{3.7}$$

$$EI_{12}I_{13}I_{14} = O(h^{3d}), \tag{3.8}$$

$$EI_{12}I_{13}I_{14}I_{15} = O(h^{4d}). \tag{3.9}$$

Also, we obtain

$$EM_{12}M_{13} = O(h^{2d+2}), \tag{3.10}$$

$$EM_{12}^2 = h^{d+2} \nu_2 \int |m'(x)|^2 f^2(x)\,dx + o(h^{d+2}), \tag{3.11}$$

$$EM_{12}^4 = O(h^{d+4}), \tag{3.12}$$

$$EM_{12}^2 M_{13}^2 = O(h^{2d+4}), \tag{3.13}$$

$$EM_{12}^2 M_{13}M_{14} = O(h^{3d+4}), \tag{3.14}$$

$$EM_{12}M_{13}M_{14}M_{15} = O(h^{4d+4}). \tag{3.15}$$

Since $EN_h = n(n-1)E_{12}$ and $\mathrm{Var}(N_h) = 2n(n-1)\{EI_{12} - (EI_{12})^2\} + 4n(n-1)(n-2)\{E(I_{12}I_{13}) - (EI_{12})^2\}$, we have from (3.6)–(3.9)

$$N_h = n^2 h^d \nu_0 \int f^2(x)\,dx + o_p(n^2 h^d). \tag{3.16}$$

Now, since $E(2N_hB_n) = n(n-1)EM_{12}^2$ and $\mathrm{Var}(2N_hB_n) = 2n(n-1)\{EM_{12}^4 - (EM_{12}^2)^2\} + 4n(n-1)(n-2)\{E(M_{12}^2M_{13}^2) - (EM_{12}^2)^2\}$, we have from (3.10)–(3.15)

$$2\,N_hB_n = n^2 h^{d+2} \nu_2 \int |m'(x)|^2 f^2(x)\,dx + o_p(n^2 h^{d+2}). \tag{3.17}$$

The proof of (3.3) is now completed by (3.16) and (3.17).

Next, we prove (3.4). Note that for $i \neq j \neq k$

$$\mathrm{Var}\{(\epsilon_i - \epsilon_j)^2\} = 2(E\epsilon_1^4 + \tau^2), \quad \mathrm{Cov}\{(\epsilon_i - \epsilon_j)^2, (\epsilon_i - \epsilon_k)^2\} = \mathrm{Var}(\epsilon_1^2).$$

Thus,

$$\mathrm{Var}(S_n | X_1, \ldots, X_n) = (E\epsilon_1^4 + \tau^2)N_h^{-1} + \mathrm{Var}(\epsilon_1^2)N_h^{-2} \sum\sum_{i \neq j \neq k} \sum I_{ij}I_{ik}. \tag{3.18}$$

From (3.6)–(3.9) it can be seen that

$$\sum\sum\sum_{i\neq j\neq k} I_{ij}I_{ik} = n^3h^{2d}\nu_0^2\int f^3(x)\,dx + o_p(n^3h^{2d}) + O_p\left\{(n^3h^{2d} + n^4h^{3d} + n^5h^{4d})^{1/2}\right\}. \tag{3.19}$$

The formula (3.4) follows from (3.16), (3.18) and (3.19).

Finally, we prove (3.5). Since $\mathrm{Var}(\epsilon_i - \epsilon_j) = 2\tau$ and $\mathrm{Cov}(\epsilon_i - \epsilon_j, \epsilon_i - \epsilon_k) = \tau$ for $i \neq j \neq k$, we have

$$\mathrm{Var}(R_n|X_1,\ldots,X_n) = 4\tau N_h^{-2}\Big\{\sum\sum_{i\neq j} M_{ij}^2 + \sum\sum\sum_{i\neq j\neq k} M_{ij}M_{ik}\Big\}. \tag{3.20}$$

It follows from (3.17) that $\sum\sum_{i\neq j} M_{ij}^2 = 2N_hB_n = O(n^2h^{d+2})$. Also, from (3.10)–(3.15) we obtain

$$\begin{aligned}\sum\sum\sum_{i\neq j\neq k} M_{ij}M_{ik} &= E\Big(\sum\sum\sum_{i\neq j\neq k} M_{ij}M_{ik}\Big) + O_p\Big\{\mathrm{Var}\Big(\sum\sum\sum_{i\neq j\neq k} M_{ij}M_{ik}\Big)^{1/2}\Big\}\\ &= O_p\Big(n^3h^{2d+2}\Big) + O_p\Big\{\Big(n^3h^{2d+4} + n^4h^{3d+4} + n^5h^{4d+4}\Big)^{1/2}\Big\}.\end{aligned}$$

This with (3.16) and (3.20) shows

$$\mathrm{Var}(R_n|X_1,\ldots,X_n) = O_p\Big(n^{-2}h^{-d+2} + n^{-1}h^2\Big).$$

REFERENCES

BROWN, B. M. AND KILDEA, D. G. (1978). "Reduced U-statistics and Hodges-Lehmann estimator", *The Annals of Statistics*, **6**, 828–835.

CARROLL, R. J. (1988). "The effects of variance function estimation on prediction and calibration: An example", In *Statistical Decision Theory and Related Topics IV, Vol. 2* (J. O. Berger and S. S. Gupta, eds.), Springer-Verlag, New York.

CARROLL, R. J. AND RUPPERT, D. (1988). *Transformation and Weighting in Regression*, Chapman and Hall, New York.

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*, Ser. B, **60**, 751–764.

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", *Journal of the Royal Statistical Society*, Ser. 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.

HALL, P., KAY, J. W. AND TITTERINGTON, D. M. (1991). "On estimation of noise variance in two-dimensional signal processing", *Advances in Applied Probability*, **23**, 476–495.

KULASEKERA, K. B. AND GALLAGHER, C. (2002). "Variance estimation in nonparametric multiple regression", *Communications in Statistics-Theory and Methods*, **31**, 1373–1383.

MÜLLER, H. G. AND STADTMÜLLER U. (1987). "Estimation of heteroscedascity in regression analysis", *The Annals of Statistics*, **15**, 610–625.

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

PARK, B. U., KIM, T. Y., LEE, Y. K. AND PARK, C. (2006). "A simple estimator of error correlation in nonparametric regression models", *Scandinavian Journal of Statistics-Theory and Applications*, in print.

RICE, J. (1984). "Bandwidth choice for nonparametric regression", *The Annals of Statistics*, **12**, 1215–1230.

SPOKOINY, V. (2002). "Variance estimation for high-dimensional regression models", *Journal of Multivariate Analysis*, **82**, 111–133.