Biometrika (1993), **80**, 2, *pp.* 373–83
Printed in Great Britain

Nonparametric estimation of residual variance revisited

BY BURKHARDT SEIFERT, THEO GASSER
Biostatistics Department, Institut für Sozial- und Präventivmedizin, Universität Zürich, CH-8006 Zürich, Sumatrastrasse 30, Switzerland

AND ANDREAS WOLF
Biostatistics Department, Zentralinstitut für Seelische Gesundheit, W-6800 Mannheim, J5, Germany

SUMMARY

Several difference-based estimators of residual variance are compared for finite sample size. Since the introduction of a rather simple estimator by Gasser, Sroka & Jennen-Steinmetz (1986) other proposals have been made. Here the one given by Hall, Kay & Titterington (1990) is of particular interest. It minimizes the asymptotic variance. Unfortunately it has severe problems with finite sample bias, and the estimator of Gasser et al. (1986) proves still to be a good choice. A new estimator is introduced, compromising between bias and variance.

Some key words: Divided differences; Efficiency; Nonparametric estimation; Nonparametric regression; Residual variance.

1. INTRODUCTION

When fitting a nonparametric regression function, it is natural to ask for a nonparametric estimator of residual variance σ^2 as well. It is also needed to check goodness of fit (Eubank & Spiegelman, 1990), outliers and homoscedasticity, and in bandwidth selection (Rice, 1984; Gasser, Kneip & Köhler, 1991) or signal restoration (Thompson, Kay & Titterington, 1991).

A simple difference-based estimator of σ^2 was introduced by Gasser, Sroka & Jennen-Steinmetz (1986). Several authors discussed improvements (Buckley, Eagleson & Silverman, 1988; Buckley & Eagleson, 1989; Hall & Marron, 1990; Hall, Kay & Titterington, 1990). To compare the different methods, let us consider a fixed design regression model $y = r + \varepsilon$, where $y = (y_1, \ldots, y_n)'$ is the vector of observations, $r = (r(x_1), \ldots, r(x_n))'$ is an unknown 'smooth' regression function at design points $x_1 \leqslant \ldots \leqslant x_n$, and where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ are independent random errors satisfying $E(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma^2$. We stick to finite sample properties of difference-based estimators as far as possible and try to avoid asymptotics. In this way, we can also avoid simulations and give exact results for selected examples.

In § 2 finite sample properties of difference-based estimators of residual variance are discussed. A new class of estimators is introduced, which combines the ideas of Gasser et al. (1986) and Hall, Kay & Titterington (1990). A case study in § 3 compares these estimators. In § 4 some generalizations are discussed.

2. DIFFERENCE-BASED ESTIMATORS

2·1. *Finite sample characteristics*

Naive nonparametric residuals, obtained by subtracting an appropriately smoothed curve from the observations, have been proposed for estimating σ^2 (Silverman, 1985; Wahba, 1983). Inevitably, the smoothing bias results in a substantial positive bias of the resulting estimator of residual variance. Choosing the curve estimator with respect to extracting residual variance has been studied by Buckley et al. (1988) and Hall & Marron (1990). Carter & Eagleson (1992) show the superiority of the estimator of Buckley et al. over that of Wahba. The resulting estimators are not difference-based. Hall, Kay & Titterington (1990) mentioned some of their disadvantages.

According to Anderson (1971, pp. 60–), differences were used for correlation between two series by Cave-Browne-Cave (1904), Hooker (1905) and Student (1914). O. Anderson (1929) and Tintner (1940) studied variance estimators for equally spaced designs. Gasser et al. (1986) introduced a method for general designs and Hall, Kay & Titterington (1990) found asymptotically optimal differences. It is an open question which differences to use for finite samples.

For coefficients d_{ik} define the ith pseudo-residual of order m as

$$e_i = \sum_{k=0}^{m} d_{ik} y_{i+k} \tag{1}$$

for $i = 1, \ldots, n-m$. Let

$$D = \begin{pmatrix} d_{10} & \ldots & & d_{1m} & 0 & \ldots & 0 \\ 0 & d_{20} & \ldots & d_{2m} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \ddots & & 0 \\ 0 & \ldots & 0 & d_{(n-m),0} & \ldots & & d_{(n-m),m} \end{pmatrix}$$

and $A = D'D$. Then $e = Dy$ is the vector of pseudo-residuals (1), and

$$\hat{\sigma}^2 = e'e = y'D'Dy = y'Ay \tag{2}$$

is called a difference-based estimator of the residual variance σ^2. It has expectation

$$E(\hat{\sigma}^2) = \sigma^2 \operatorname{tr}(A) + r'Ar, \tag{3}$$

where tr(.) denotes the trace of a matrix. For moments of quadratic forms, results given by Rao & Kleffe (1988, pp. 31–) are used.

Assume that the residuals ε_i have finite fourth moments, and let $E(\varepsilon_i^3) = \gamma\sigma^3$ and $E(\varepsilon_i^4) = (\kappa+3)\sigma^4$. Then, the variance and bias of $\hat{\sigma}^2$ are

$$\begin{aligned} \operatorname{var}(\hat{\sigma}^2) = {} & 2\sigma^4 \operatorname{tr}(A^2) + 4\sigma^2 r'A^2 r + 2\gamma\sigma^3[\operatorname{tr}\{A \operatorname{Diag}(Ar1')\} + r'A \operatorname{Diag}(A)1] \\ & + \kappa\sigma^4 \operatorname{tr}\{A \operatorname{Diag}(A)\}, \end{aligned} \tag{4}$$

$$\operatorname{bias}(\hat{\sigma}^2) = r'Ar + \sigma^2\{\operatorname{tr}(A) - 1\}. \tag{5}$$

Here Diag (A) denotes the diagonal matrix with the same diagonal elements as A. Assume

$$\operatorname{tr}(A) = \sum_{i=1}^{n-m} \sum_{k=0}^{m} d_{ik}^2 = 1; \tag{6}$$

that is $\hat{\sigma}^2$ is unbiased for $r = 0$. Asymptotically $\operatorname{tr}(A) \to 1$ is necessary for consistency.

Formulae (4) and (5) allow exact computation of variance, bias and mean squared error of difference-based estimators for finite samples without simulations. Further, the impact of neglecting terms when deriving an asymptotically optimal estimator can be assessed. A direct representation in terms of coefficients d_{ik} allows fast computer programs; further, for given design, r, n and m, the finite sample optimal estimator can be obtained as a reference. Let $e(r) = Dr$ and $a = (a_{11}, \ldots, a_{nn})'$. Then

$$\operatorname{tr}(A^2) = \sum_{i=1}^{n-m} \left(\sum_{k=0}^{m} d_{ik}^2 \right)^2 + 2 \sum_{l=1}^{m} \sum_{i=1}^{n-m-l} \left(\sum_{k=0}^{m-l} d_{i,(k+l)} d_{(i+l),k} \right)^2, \tag{7}$$

$$\operatorname{tr}\{A \operatorname{Diag}(A)\} = \sum_{i=1}^{n-m} \sum_{k=0}^{m} d_{ik}^2 a_{(i+k),(i+k)}, \tag{8}$$

$$r'A^2r = \sum_{i=1}^{n-m} \left(\sum_{k=0}^{m} d_{ik}^2 \right) e_i^2(r) + 2 \sum_{l=1}^{m} \sum_{i=1}^{n-m-l} \left(\sum_{k=0}^{m-l} d_{i,(k+l)} d_{(i+l),k} \right) e_i(r) e_{i+l}(r), \tag{9}$$

$$r'Ar = \sum_{i=1}^{n-m} e_i^2(r), \quad r'A \operatorname{Diag}(A)1 = \sum_{i=1}^{n-m} e_i(a) e_i(r), \tag{10}$$

$$\operatorname{tr}\{A \operatorname{Diag}(Ar1')\} = \sum_{i=1}^{n-m} \sum_{k=0}^{m} d_{ik}^2 \sum_{j=\max(1,i+k-m)}^{\min(i+k,n-m)} d_{j,(i+k-j)} e_j(r). \tag{11}$$

2·2. *The estimator of Gasser, Sroka & Jennen-Steinmetz*

For the rest of this section let $x_1 < \ldots < x_n$: see § 4·2 below for a brief discussion of multiple measurements.

The problem is to find suitable coefficients d_{ik}. The disturbing bias of nonparametric variance estimators and the fact that smooth functions can be locally well approximated by polynomials led Gasser et al. (1986) to the following procedure for $m = 2$. Consider pseudo-residuals $e_1, \ldots, e_{n-m}$ as in (1) satisfying $E(e_i) = 0$ when r is a polynomial of order less than m. The latter is equivalent to

$$\sum_{k=0}^{m} d_{ik} r_{i+k} = 0 \tag{12}$$

for all i. Using the normalizing condition

$$\sum_{k=0}^{m} d_{ik}^2 = 1/(n-m) \quad (i = 1, \ldots, n-m) \tag{13}$$

leads to equal variances $\operatorname{var}(e_i) = \sigma^2/(n-m)$ for pseudo-residuals of such polynomials. We then get an implicit definition of a Gasser-Sroka-Jennen-Steinmetz-estimator, $\hat{\sigma}^2_{\mathrm{GSJ}}$ say, for general m.

Definition 1. Let $e_1, \ldots, e_{n-m}$ be pseudo-residuals as in (1) satisfying (12) and (13). Then, a GSJ-estimator of order m is defined as

$$\hat{\sigma}^2_{\mathrm{GSJ}} = \sum_{i=1}^{n-m} e_i^2. \tag{14}$$

THEOREM 1. *The GSJ-estimator of order m is unique.*

Proof. Let $d_i = (d_{i0}, \ldots, d_{im})'$. By definition we have $e_i = (y_i, \ldots, y_{i+m}) d_i$. For every polynomial r of order less than m it follows that $(r_i, \ldots, r_{i+m})' = F_i \beta$, where F_i is an

$m \times (m+1)$ matrix of rank m. Relation (12) yields $d_i' F_i = 0$. Consequently d_i is determined up to a scalar factor. Equation (13) then determines d_i up to the sign. □

Divided differences, consult e.g. de Boor (1978, pp. 4–), provide pseudo-residuals with small bias. Divided differences $\Delta^{(m)}$ of order m reduce polynomials r of order less than m to zero: $\Delta^{(m)} r = 0$. They are constructed as follows. Denote by $\operatorname{diag}(w_i)$ the diagonal matrix with diagonal elements w_i and define

$$D^{(k)} = \operatorname{diag}\left(\frac{1}{x_{i+k} - x_i}\right)_{i=1,(n-k)}, \quad B^{(k)} = \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}$$

a weighting matrix of order $(n-k) \times (n-k)$ and a bidiagonal matrix of order $(n-k) \times (n-k+1)$. Then divided differences of y of order m are obtained as

$$\Delta^{(m)} y = D^{(m)} B^{(m)} \ldots D^{(1)} B^{(1)} y. \tag{15}$$

Relation (12) is fulfilled for divided differences of order m. As a consequence of Theorem 1 the GSJ-estimator is of that form, and (15) together with (13) give a recursive algorithm for coefficient d_{ik}, GSJ-pseudo-residuals e_i and $\hat{\sigma}^2_{\mathrm{GSJ}}$ in (14).

Because of their small bias for polynomial functions the GSJ-estimators are minimax in certain classes of 'smooth' functions. Smoothness is usually defined by

$$\int \{r^{(p)}(x)\}^2 \, dx \leq c\sigma^2 \tag{16}$$

for some smoothness order p. This assumption ensures that every regression function can be approximated with bounded error by a polynomial of order $p-1$. For finite samples we only have information at design points x_i, and the derivatives are replaced by some finite difference-version. Buckley et al. (1988) proposed a version connected with cubic spline interpolation. Consider now a general difference-version of (16):

$$r' \Omega r \leq c\sigma^2, \tag{17}$$

where Ω is of the form

$$\Omega = (\Delta^{(p)})' R' R \Delta^{(p)} \tag{18}$$

for some nonsingular matrix R of order $(n-p) \times (n-p)$.

THEOREM 2. *Consider the class of regression functions r satisfying* (17) *and* (18) *for some given p and arbitrary but fixed c and R, where R is a nonsingular matrix. Consider further the class of difference-based estimators* (2) *of order* $m \leq p$ *satisfying* (13). *Then, if* $n > p$, *the GSJ-estimator of order* $m = p$ *is the unique minimax estimator in this class with respect to mean squared error.*

Proof. From (18) we get $r' \Omega r = 0$ for every polynomial of order less than p. If $Dr \neq 0$ for such a polynomial, we get $r' A r = r' D' D r > 0$. Hence the mean squared error of $\hat{\sigma}^2 = y' A y$ is unbounded in the class of regression functions r satisfying (17) and (18). Consequently $Dr = 0$ for all polynomials of order less than p is a necessary condition for a bounded mean squared error of $\hat{\sigma}^2$. Following de Boor (1978, pp. 4–) the divided differences of order $m = p$ in (15) are the only differences of order m with this property. From Definition 1 and Theorem 1 it follows that the GSJ-estimator is the unique difference-based one of order m satisfying (13) and $Dr = 0$ for all polynomials of order less than p.

It remains to show that the mean squared error of the GSJ-estimator is bounded. Indeed,

$$\|\Delta^{(p)}r\|^2 = r'(\Delta^{(p)})'\Delta^{(p)}r \leqslant \frac{r'(\Delta^{(p)})'R'R\Delta^{(p)}r}{\lambda_{\min}(R'R)} \leqslant \frac{c\sigma^2}{\lambda_{\min}(R'R)}.$$

From (4) and (5) it is standard algebra to prove a bounded mean squared error of $\hat{\sigma}^2_{\mathrm{GSJ}} = y'(\Delta^{(p)})'C\Delta^{(p)}y$ for bounded $\|\Delta^{(p)}r\|^2$. □

The result is rather strong, since it holds for all sample sizes, all fixed designs, arbitrary error distributions, and independently of c and R. Moreover, every function r satisfies such a condition (17) and (18), even functions with jumps. On the other hand, the restriction to estimators satisfying (13) is motivated more heuristically than decision-theoretically. However, the gain from using more general weights is small (§ 4·3). Another assumption is $m \leqslant p$, which again is motivated by a heuristic argument only. In § 2·4 the increase of m for fixed p is discussed.

As a consequence of Theorem 2, these estimators are attractive candidates for initial estimation of residual variance. Of course, with additional knowledge about r and the error structure, both the minimax argument and the class of quadratic estimators lose their legitimacy.

2·3. *The estimator of Hall, Kay & Titterington*

Hall, Kay & Titterington (1990) observed that bias and certain expressions in the variance are asymptotically negligible. Consequently, based on (4) and (5), the mean squared error becomes

$$\mathrm{MSE}(\hat{\sigma}^2) \doteq 2\sigma^4 \operatorname{tr}(A^2) + \kappa\sigma^4 \operatorname{tr}\{A \operatorname{Diag}(A)\}. \tag{19}$$

Hall, Kay & Titterington (1990) minimized the asymptotic expression of $\operatorname{tr}(A^2)$ for $m \leqslant 10$ under

$$d_{ik} = d_k \quad (k = 0, \ldots, m;\ i = 1, \ldots, n-m), \tag{20}$$

$$\sum_{k=0}^{m} d_k = 0, \quad \sum_{k=0}^{m} d_k^2 = 1/(n-m). \tag{21}$$

Let us denote the resulting estimators by $\hat{\sigma}^2_{\mathrm{HKT}}$. Relations (20) and (21) reflect that these estimators are designed for smoothness order $p = 1$ in (16). For $p = 1$ every regression function can be approximated with bounded error by a constant. Independently of the design, (20) is then appropriate, and (21) is analogous to (12) and (13).

Under the restrictions (20) and (21), $n \operatorname{tr}\{A \operatorname{Diag}(A)\}$ tends to 1 independently of the choice of $d_0, \ldots, d_m$, so that the HKT-estimator is asymptotically optimal for normal and nonnormal residuals.

For $m = 2$ the HKT-estimator yields

$$\operatorname{tr}(A^2_{\mathrm{HKT}}) = 5/\{4(n-2)\} - 3/\{8(n-2)^2\},$$

which is very close to the finite minimum $\operatorname{tr}(A^2_{\mathrm{HKT}}) - 1/\{8(n-2)^2(2n-7)\}$. The corresponding value of the GSJ-estimator for equidistant design is

$$\operatorname{tr}(A^2_{\mathrm{GSJ}}) = 35/\{18(n-2)\} - 1/(n-2)^2.$$

Assuming r = constant, the finite sample gain of the HKT-estimator over the GSJ-estimator is 52% for $n = 10$ and still 36% for large n. The asymptotic gain holds for arbitrary r.

However, the finite-sample performance of the HKT-estimator depends strongly on r. Not only bias but also variance are adversely affected. The case study in § 3 below shows that it may take sample sizes of $n = 500$ or more until the asymptotic formula (19) works.

2·4. *A new estimator*

Roughly speaking, Theorem 2 says that it is impossible to find a difference-based estimator, of order not greater than m, which behaves better for a smoothness order $p = m$ than the GSJ-estimator. A way out is to increase m, in the same way that Hall, Kay & Titterington (1990) improved the estimator of Rice (1984) (see Table 1) by increasing m for a fixed smoothness order $p = 1$.

Table 1. *Coefficients for equidistant design, relative weights of pth order divided differences and asymptotic mean squared error of some difference-based estimators*

Estimator	m	p	$\sqrt{(n-m)}(d_0, \ldots, d_m)$	δ_0	δ_1	δ_2	MSE$(\hat{\sigma}^2)$
Rice	1	1	(−0·707, 0·707)	1			$3{\cdot}00 \times \sigma^4/n$
HKT	2	1	(−0·809, 0·500, 0·309)	1	0·382		$2{\cdot}50 \times \sigma^4/n$
GSJ	2	2	(−0·408, 0·816, −0·408)	1			$3{\cdot}89 \times \sigma^4/n$
HKT	3	1	(0·194, 0·281, 0·383, −0·858)	1	2·448	4·423	$2{\cdot}33 \times \sigma^4/n$
'New'	3	2	(0·535, −0·802, 0·000, 0·267)	1	0·500		$3{\cdot}00 \times \sigma^4/n$
GSJ	3	3	(0·224, −0·671, 0·671, −0·224)	1			$4{\cdot}62 \times \sigma^4/n$

Now for smoothness order $p = 2$ the question arises whether it is possible to improve the variance of the GSJ-estimator without essentially increasing the bias by going from $m = 2$ to 3. While the idea of divided differences and of the GSJ-estimators is successive differencing, that of the HKT-estimators is to smooth normalized first-order differences $y_{i+1} - y_i$ to improve the variance of the estimator. Indeed, for $m = 2$ the HKT-pseudo-residuals are

$$e_i = (n-2)^{-\frac{1}{2}}\, 0{\cdot}809(1, 0{\cdot}382)(y_{i+1} - y_i, y_{i+2} - y_{i+1})'$$

(Table 1). To generalize this idea to general p and $m = p + 1$ let us introduce

$$\Delta = \begin{pmatrix} 1 & \delta_1 & & \\ & \ddots & \ddots & \\ & & 1 & \delta_1 \end{pmatrix},$$

a bidiagonal smoothing matrix of order $(n-m) \times (n-m+1)$. Then define general differences of order $m = p + 1$ for smoothness order p as

$$\Delta^{(m,p)} y = \Delta D^{(p)} B^{(p)} \ldots D^{(1)} B^{(1)} y = \Delta \Delta^{(p)} y. \tag{22}$$

Let $\Delta_i^{(m,p)}$ be the rows of $\Delta^{(m,p)}$. Then, pseudo-residuals $e_i = w_i \Delta_i^{(m,p)} y$ can be defined as weighted general differences, such that (13) is fulfilled, and $\hat{\sigma}^2$ is as in (2).

A generalization to arbitrary $m > p$ is straightforward. Special cases are the GSJ-estimator for arbitrary p and $m = p$ with $\delta_1 = 0$ and the GSJ-estimator for $m = p + 1$ with $\delta_1 = -1$. For equidistant design the HKT-estimator for $m = 2$ with $p = 1$ and $\delta_1 = 0{\cdot}382$ is a special case (Table 1).

The question is how to specify the weight δ_1. A finite optimal δ_1 depends on the class of regression functions, sample size and design. Here, the asymptotic idea of Hall, Kay

& Titterington (1990) proves to be useful. For $p=2$, $m=3$ and equidistant design on $[0, 1]$ general differences (22) become

$$\Delta_i^{(3,2)}y=(n^2/2)(1, -2+\delta_1, 1-2\delta_1, \delta_1)(y_i, \ldots, y_{i+3})',$$

and $\mathrm{tr}\,(A^2)$ in (7) is minimal for

$$\delta_1=\frac{35(n-3)-54}{28(n-3)-36}\pm\left[\left\{\frac{35(n-3)-54}{28(n-3)-36}\right\}^2-1\right]^{\frac{1}{2}}.$$

For $n\to\infty$ this optimal value tends to $\delta_1=(5\pm3)/4$. Table 1 shows the estimator for $\delta_1=0{\cdot}5$. It is called 'new' and used in the case study in § 3. The other solution gives just the reflected difference.

Consider the class of difference-based estimators of order $m=p+1$ satisfying (12) for polynomials of order less than p. Arguments similar to Hall, Kay & Titterington (1990) show that the 'new' estimator is asymptotically optimal in this class under standard assumptions for regular designs, general error distributions and general regression functions. For an equidistant design we obtain

$$\mathrm{tr}\,(A_{\mathrm{new}}^2)=3/\{2(n-3)\}-33/\{49(n-3)^2\},$$

which leads to a relative gain in asymptotic variance of 23% relative to the GSJ-estimator and a loss of 20% relative to the HKT-estimator for $m=2$; see Table 1.

3. A CASE STUDY

3·1. *The design of the case study*

The finite sample properties of the GSJ-, HKT- and 'new' estimators have been investigated using the formulae (4) and (5) and the explicit expressions (6)–(11). All results for fixed designs are exact, and no simulations were necessary. After describing the situations considered the results are illustrated for some typical and interesting ones.

The regression functions considered were (i) linear: $r(x)=2x$, (ii) exponential: $r(x)=2\exp(-x/0{\cdot}3)$, (iii) sine: $r(x)=2\sin(4\pi x)$, and (iv) linear with Gaussian peak: $r(x)=2-5x+\exp\{-100(x-0{\cdot}5)^2\}$. The results were compared for two residual variances (i) $\sigma^2=0{\cdot}1$ and (ii) $\sigma^2=1$. The sample size varied between 15 and 500. Four types of design on $[0, 1]$ were studied: (i) equidistant fixed design: $x_i=(i-0{\cdot}5)/n$, (ii) nonequidistant fixed design: x_i are quantiles of a Beta $(2, 2)$ distribution, (iii) random design: x_i are uniformly distributed on $[0, 1]$, and (iv) random design: x_i are distributed Beta $(2, 2)$. The error distributions were: (i) normal $(\gamma=0, \kappa=0)$, (ii) skewed $(\gamma=1{\cdot}155, \kappa=2)$: χ^2-distribution with 6 degrees of freedom, and (iii) platykurtic $(\gamma=0, \kappa=3)$: t-distribution with 6 degrees of freedom.

3·2. *Equidistant design and normal errors*

For every situation, $r(x)$, σ^2, n and design given, the finite optimal coefficient δ_1 for general differences in (22) of order $m=2$ for $p=1$ was computed by grid search, and the relative inefficiencies of the HKT-, GSJ- and 'new' estimators relative to the resulting 'ideal' one were plotted: $\mathrm{ineff}\,(\hat\sigma^2)=\mathrm{MSE}(\hat\sigma^2)/\mathrm{MSE}(\hat\sigma^2_{\mathrm{ideal}})$.

For exponential and linear regression functions, asymptotics begin to work already at sample size $n=30$. The HKT-estimator becomes then practically the optimal estimator of order $m=2$. The GSJ-estimator achieves its asymptotic relative inefficiency of $\frac{14}{9}$, and the

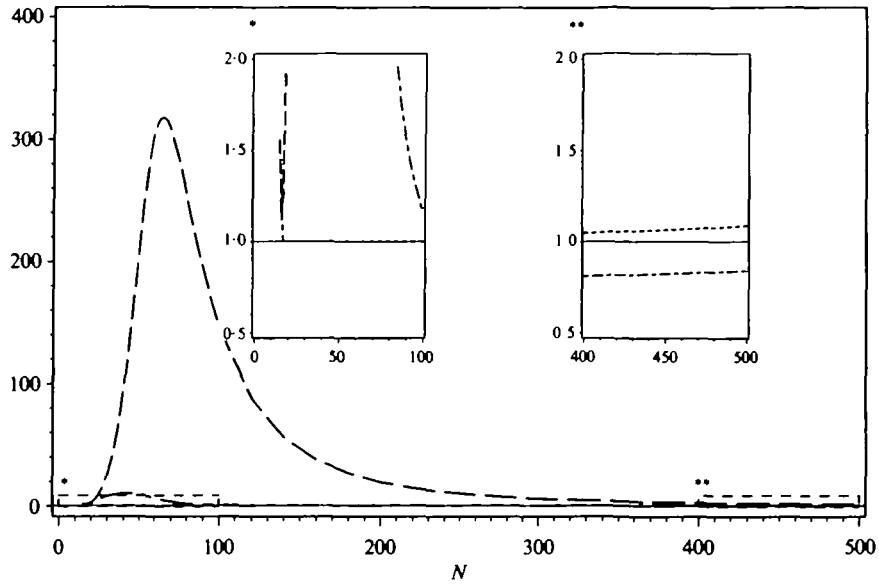

Fig. 1. Relative inefficiencies of the HKT- (long dashes), GSJ- (short dashes), and 'new' (dash-dot) estimators for sample size $n = 15$–500, sine regression function and $\sigma^2 = 0{\cdot}1$. Above are magnified windows for small (left) and large (right) sample size.

'new' estimator is a compromise, with relative inefficiency $\frac{6}{5}$. The situation changes dramatically for the sine function and $\sigma^2 = 0{\cdot}1$; see Fig. 1. The relative inefficiency of the HKT-estimator achieves a value of 317 for $n = 65$, while the other estimators behave well over the whole range of sample sizes (compare the magnified windows in Fig. 1). For $n = 200$ observations the HKT-estimator has a relative inefficiency of 20. Even for $n = 500$ the asymptotic formula (19) does not correctly reflect the situation. There the HKT-estimator still has a relative inefficiency of more than 2, while the 'new' estimator is superefficient. The situation eases for higher noise to signal ratio. The regression function with a Gaussian peak (Fig. 2) gives similar results.

3·3. *Other cases*

The situation is comparable for regular nonequidistant fixed designs. The asymptotic bias, variance and mean squared error remain, but for small sample size the values may change. In the example of a nonequidistant design with $n = 25$, sine regression and normal errors with $\sigma^2 = 0{\cdot}1$, the mean squared error of the estimators is reduced by factors 3·5 (HKT), 9 (GSJ) and 11 ('new'). Table 2 shows the mean squared error for a moderate example. There changes are small for all error distributions.

The shape of the error distribution has no influence on the expectation of a difference-based estimator (compare (3)). The variance, however, changes; compare (4). The influence of skewness asymptotically is negligible and small for finite samples. As to the influence of kurtosis let us assume an equidistant design. Then, standard calculations using (8) yield

$$(n-2m)/(n-m)^2 \leqslant \operatorname{tr}\{A \operatorname{Diag}(A)\} \leqslant 1/(n-m).$$

Consequently, the kurtosis of the error distribution heavily influences the variances of estimators, but for all estimators by nearly the same amount. A similar observation holds for nonequidistant designs (Table 2).

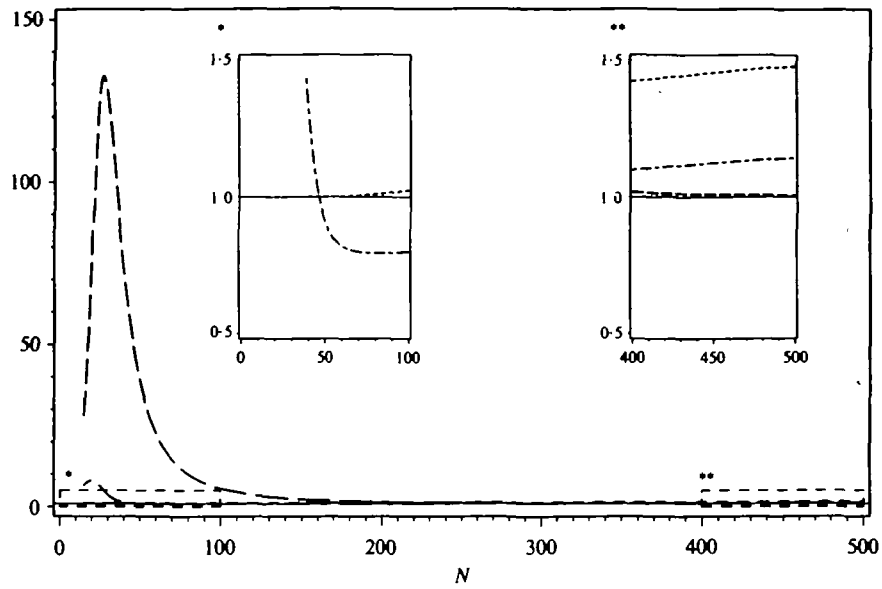

Fig. 2. Relative inefficiencies of the HKT- (long dashes), GSJ- (short dashes), and 'new' (dash-dot) estimators for sample size $n = 15$–500, linear regression function with Gaussian peak and $\sigma^2 = 0{\cdot}1$. Above are magnified windows for small (left) and large (right) sample size.

Table 2. *Mean squared error of estimators for sample size* $n = 100$, *sine regression function and* $\sigma^2 = 1$

Design	Error distribution	MSE($\hat{\sigma}^2$) HKT	GSJ	New
Equidistant	Normal	0·086	0·039	0·031
	Skewed	0·106	0·060	0·051
	Platykurtic	0·116	0·070	0·062
Nonequidistant	Normal	0·051	0·040	0·032
	Skewed	0·072	0·060	0·053
	Platykurtic	0·081	0·071	0·063
Random $B(2, 2)$	Normal	0·031	0·041	0·050
Random $U[0, 1]$	Normal	0·032	0·040	0·049

For random designs, the explicit formulae for finite sample mean squared error no longer work. In the study each case was simulated 400 times, and formulae

$$E(\hat{\sigma}^2) = E\{E(\hat{\sigma}^2 | x_1, \ldots, x_n)\},$$

$$\text{var}\,(\hat{\sigma}^2) = E\{\text{var}\,(\hat{\sigma}^2 | x_1, \ldots, x_n)\} + E[\{E(\hat{\sigma}^2 | x_1, \ldots, x_n)\}^2] - \{E(\hat{\sigma}^2)\}^2$$

together with (4) and (5) were used to improve efficiency of simulations. The asymptotic mean squared error was the same for HKT- and GSJ-estimators as in the equidistant case. The asymptotic mean squared error of the 'new' estimator, however, increased by a factor of about 1·5; see Table 2.

3·4. *Conclusions*

The HKT-estimator should be used only for large sample sizes and flat regression functions. But many typical applications, for example biostatistical ones, have sample

sizes $n = 15$ to 100. The problems with bias for the HKT-estimator are in qualitative accordance with the smoothness assumption $p = 1$. The 'new' estimator behaves well over a wide range of situations, but may be somewhat inefficient for small sample size and for irregular and random designs. The GSJ-estimator behaves well in all situations. Consequently, the GSJ-estimator is a reasonable compromise.

4. SOME GENERALIZATIONS

4·1. *Random designs*

The finite sample minimax property of the GSJ-estimator in Theorem 2 essentially remains for random designs. Let us discuss convenient assumptions. The class of regression functions now is restricted by (16) to 'smooth' ones. The $n > p$ design points should be distinct with probability 1. Otherwise we can do better (§ 4·2). Some additional assumption on $r(x)$ and/or the distribution of design points is needed to ensure a bounded mean squared error of the GSJ-estimator of order $m = p$. If we assume for simplicity that the pth derivative of the regression function is uniformly Lipschitz continuous and the design is on [0, 1], no additional assumption is necessary.

THEOREM 3. *Under the above assumptions, the GSJ-estimator of order $m = p$ is the essentially unique minimax estimator in the class of difference-based estimators* (2) *of order $m \leqslant p$ satisfying* (13) *for almost all realizations of the design.*

The proof goes along the lines of that of Theorem 2. As a consequence, the mean squared error of the HKT-estimator is unbounded, while that of the GSJ- and 'new' estimators is bounded.

4·2. *Multiple measurements*

Multiple measurements are a chance for estimation of variance, for only in this situation is there an unbiased estimator. Gasser et al. (1986) and Hall, Kay & Titterington (1990) proceed as usual; others, e.g. Buckley et al. (1988), even exclude this situation. Assume observations y_{ij} for $j = 1, \ldots, n_i \geqslant 1$ at different design points $x_1 < \ldots < x_n$. Let $\bar{y}_i$ denote a cell mean and s^2 the unbiased analysis-of-variance-estimator of σ^2. For normal errors, s^2 is independently distributed of any nonparametric estimator $\hat{\sigma}^2$ based on $\bar{y}_i$. Pseudo-residuals for $\bar{y}_i$. can be computed as before. Condition (13) is replaced by $\Sigma\, d_{ik}^2 / n_{i+k} = 1/(n-m)$. The mean squared error of $\hat{\sigma}^2_{\text{mix}} = as^2 + (1-a)\hat{\sigma}^2$ is minimized for

$$a = \text{MSE}(\hat{\sigma}^2)/\{\text{MSE}(s^2) + \text{MSE}(\hat{\sigma}^2)\}.$$

The gain over s^2 and $\hat{\sigma}^2$ can be considerable.

4·3. *General weights*

The choice of equal weights in (13) is the simplest but not necessarily the natural and optimal one. Several authors (Kendall, 1946, Problem 30.8; Quenouille, 1953; Anderson, 1971, pp. 73–) discussed corrections. One possibility is to use general weights $\Sigma\, d_{ik}^2 = c_i$ for $i = 1, \ldots, n-m$ instead of (13) and find the optimal ones. For an equidistant design, GSJ-pseudo-residuals for $m = 2$, and $n = 10$; e.g. we get the weights $c_{\text{opt}} = (0{\cdot}182, 0{\cdot}080, 0{\cdot}129, 0{\cdot}109, 0{\cdot}109, 0{\cdot}129, 0{\cdot}080, 0{\cdot}182)$. The gain of variance is 3% only. Since the gain is relatively small for the additional amount of work, we recommend the classical weights, at least for sample size $n \geqslant 10$.

4·4. *Multidimensional designs*

Difference-based methods can easily be generalized to multidimensional designs. Important differences between the one- and higher-dimensional case are the very rich variety of configurations and the increasing portion of the boundary for growing dimension. In an as yet unpublished paper, E. Herrmann, M. P. Wand, J. Engel and T. Gasser generalized the GSJ-estimator for $m = 2$ to the bivariate case. Hall, Kay & Titterington (1991) generalized the HKT-estimator to bivariate lattice designs and discussed the problem of different configurations in detail. Further research has to be done to find optimal configurations.

REFERENCES

ANDERSON, O. (1929). *Die Korrelationsrechnung in der Konjunkturforschung.* Bonn: Schroeder.

ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series.* New York: Wiley.

DE BOOR, C. (1978). *A Practical Guide to Splines.* New York: Springer.

BUCKLEY, M. J. & EAGLESON, G. K. (1989). A graphical method for estimating the residual variance in nonparametric regression. *Biometrika* **76**, 203–10.

BUCKLEY, M. J., EAGLESON, G. K. & SILVERMAN, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75**, 189–99.

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

CAVE-BROWNE-CAVE, F. E. (1904). On the influence of the time factor on the correlation between the barometric heights at stations more than 1000 miles apart. *Proc. R. Soc. London* **74**, 403–13.

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

HALL, P., KAY, J. W. & TITTERINGTON, D. M. (1991). On estimation of noise variance in two-dimensional signal processing. *Adv. Appl. Prob.* **23**, 476–95.

HALL, P. & MARRON, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77**, 415–19.

HOOKER, R. H. (1905). On the correlation of successive observations, illustrated by corn prices. *J. R. Statist. Soc.* **68**, 696–703.

KENDALL, M. G. (1946). *The Advanced Theory of Statistics*, **2**. London: Griffin.

QUENOUILLE, M. H. (1953). Modifications to the variate-difference method. *Biometrika* **40**, 383–408.

RAO, C. R. & KLEFFE, J. (1988). *Estimation of Variance Components and Applications.* Amsterdam: North-Holland.

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

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc.* B **47**, 1–52.

"STUDENT" (1914). The elemination of spurious correlation due to position in time and space. *Biometrika* **10**, 179–80.

THOMPSON, A. M., KAY, J. W. & TITTERINGTON, D. M. (1991). Noise estimation in signal restoration using regularisation. *Biometrika* **78**, 475–88.

TINTNER, G. (1940). *The Variate Difference Method.* Bloomington, Ind.: Principia Press.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B **45**, 133–50.

[*Received October* 1991. *Revised November* 1992]