Biometrika (1990), 77, 3, *pp.* 521–8
Printed in Great Britain

Asymptotically optimal difference-based estimation of variance in nonparametric regression

BY PETER HALL
Department of Statistics, Australian National University, Canberra, ACT 2601, Australia

J. W. KAY AND D. M. TITTERINGTON
Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland, U.K.

SUMMARY

We define and compute asymptotically optimal difference sequences for estimating error variance in homoscedastic nonparametric regression. Our optimal difference sequences do not depend on unknowns, such as the mean function, and provide substantial improvements over the suboptimal sequences commonly used in practice. For example, in the case of normal data the usual variance estimator based on symmetric second-order differences is only 64% efficient relative to the estimator based on optimal second-order differences. The efficiency of an optimal mth-order difference estimator relative to the error sample variance is $2m/(2m+1)$. Again this is for normal data, and increases as the tails of the error distribution become heavier.

Some key words: Difference order; Efficiency; Error variance; Nonparametric regression; Variate difference method.

1. INTRODUCTION

In this paper we describe general difference-sequence methods for estimating variance in homoscedastic nonparametric regression. We show that, asymptotically, estimators based on symmetric differences, which are commonly used in practice, are strikingly inefficient relative to optimal difference-based estimators. For example, symmetric second-order differences are only 64% efficient relative to optimal second-order differences, when the data have normal errors. Techniques based on optimal differences are no more difficult to apply than methods which use nonoptimal differences, provided the optimal difference sequence has been computed. We list optimal difference sequences of all orders up to the 10th, and show how to compute optimal sequences of arbitrary order.

Our regression model is

$$Y_j = f(x_j) + e_j \quad (j = 1, \ldots, n),$$

where f is an unknown function and the errors e_j are independent and identically distributed random variables with zero mean and variance σ^2. We assume that the x_j's have been ordered, so that $x_1 \leqslant \ldots \leqslant x_n$. A difference sequence $\{d_j\}$ is a sequence of real numbers such that

$$\sum d_j = 0, \quad \sum d_j^2 = 1. \tag{1·1}$$

Assume that $d_j = 0$ for $j < -m_1$ and $j > m_2$, and $d_{-m_1} d_{m_2} \neq 0$, where m_1, $m_2 \geqslant 0$. Then $m = m_1 + m_2$ is called the order of the sequence. It is usually convenient to take $m_1 = 0$ and $m_2 = m$. Our estimator of σ^2 based on this difference sequence is

$$\hat{\sigma}^2 = (n-m)^{-1} \sum_{k=m_1+1}^{n-m_2} \left(\sum_j d_j Y_{j+k} \right)^2.$$

Estimators of this type have a long history in a time series context; see, for example, Anderson (1971, p. 66). They were first considered in the case of nonparametric regression by Rice (1984), Gasser, Sroka & Jennen-Steinmetz (1986), Müller & Stadtmüller (1987, 1989) and Müller (1988, p. 99ff). If the function f is smooth and if adjacent-indexed design points x_j get closer together as sample size increases, then the effect of f on the asymptotic mean squared error of $\hat{\sigma}^2$ is negligible. In finite-sample cases in which the bias component is also important, f does have an effect, as described in as yet unpublished work by A. M. Thompson, J. W. Kay and D. M. Titterington. The assumption that f has a bounded derivative confers more than enough smoothness, and the condition that the x_j's are regularly spaced on an interval, or are drawn randomly from a population whose density is bounded away from zero on an interval, is more than sufficient to guarantee that the design points are sufficiently close. Under these conditions the asymptotic variance of $\hat{\sigma}^2$ depends only on the error distribution and on choice of $\{d_j\}$. Judicious selection of $\{d_j\}$ yields minimization of asymptotic variance. Müller & Stadtmüller (1987) have taken a somewhat similar approach, although in the context of heteroscedastic regression. There the asymptotic variance formula, and the recommendation for choice of $\{d_j\}$, are quite different from our own.

For first-order differences the only available choice of (d_0, d_1) is $(2^{-\frac{1}{2}}, -2^{-\frac{1}{2}})$, or the sign reversal of this vector. The most commonly used second-order difference sequence is $(d_0, d_1, d_2) = (-6^{-\frac{1}{2}}, (\frac{2}{3})^{\frac{1}{2}}, -6^{-\frac{1}{2}})$, although as we shall show the sequence $(\frac{1}{4}(5^{\frac{1}{2}}+1), -\frac{1}{2}, -\frac{1}{4}(5^{\frac{1}{2}}-1))$ performs substantially better. As m increases, the optimal difference sequence becomes concentrated in a single spike. Curiously, the spike is at an extremity of the sequence when m is odd, but in the centre when m is even.

Related work on variance estimation includes contributions by Buckley, Eagleson & Silverman (1988), Eagleson (1990) and Hall & Marron (1990). This work focusses on variance estimators $\check{\sigma}^2$ which have the property

$$E\{(\check{\sigma}^2 - \sigma^2)^2\} = n^{-1} \operatorname{var}(e^2) + \varepsilon(n),$$

where $\varepsilon(n)$ is of smaller order than n^{-1} and depends on a smoothing parameter involved in the construction of $\check{\sigma}^2$. The necessity of choosing a smoothing parameter is a practical drawback to the use of such estimators. By way of comparison, the difference-based estimators considered in this paper do not require the selection of an extraneous parameter other than the order of the difference sequence, and have the property

$$E\{(\hat{\sigma}^2 - \sigma^2)^2\} \sim n^{-1} c \operatorname{var}(e^2),$$

where $c > 1$. We suggest using a low-order, optimal difference sequence, such as the optimal second-order sequence $(\frac{1}{4}(5^{\frac{1}{2}}+1), -\frac{1}{2}, -\frac{1}{4}(5^{\frac{1}{2}}-1))$. If an optimal mth order sequence is employed then the efficiency of $\hat{\sigma}^2$ relative to $\check{\sigma}^2$, for normal errors, is $2m/(2m+1)$. This ratio exceeds 0·9 for $m \geqslant 5$ and equals 0·8 for $m = 2$. The efficiency is even greater when the error distribution has heavier tails than the normal.

It would be possible to treat m as a smoothing parameter, and choose $m = m(n)$ to diverge to infinity according to a formula which depends on properties of the unknown

function f. However, there are obvious practical difficulties in specifying an appropriate way for varying m. We should stress that allowing m to vary does not improve the convergence rate of $\hat{\sigma}^2$, only its efficiency; the convergence rate is $n^{-\frac{1}{2}}$ even for $m = 1$.

2. Methodology

The asymptotic variance and mean squared error of the estimator $\hat{\sigma}^2$ are both equal to $n^{-1}\tau^2$, where

$$\tau^2 = \text{var}\,(e^2) + 2\sigma^4 \sum_{k \neq 0} \left(\sum_j d_j d_{j+k} \right)^2 = \sigma^4 \left\{ \kappa + 2 \sum_k \left(\sum_j d_j d_{j+k} \right)^2 \right\} \tag{2·1}$$

and κ denotes the kurtosis of e/σ. A formal theorem describing this property is stated in Appendix 1. Recall the notation given just after (1·1), in particular that $d_{-m_1} d_{m_2} \neq 0$ and that $m = m_1 + m_2$ is the order of the difference sequence $\{d_j\}$. Subject to condition (1·1), first-order differences are unique. However, there is a wide latitude of choice for higher-order differences.

Assume for simplicity that $m_1 = 0$ and $m_2 = m$. In view of (2·1) it is optimal to choose $d_0, \ldots, d_m$ to minimize

$$\delta = \sum_{k \neq 0} \left(\sum_j d_j d_{j+k} \right)^2 \tag{2·2}$$

subject to the constraint (1·1). Table 1 lists the optimal mth order difference sequences for $1 \leqslant m \leqslant 10$. These sequences are unique up to reversal of order and reversal of sign. For the optimal mth order difference sequence, and with δ defined by (2·2), we have $\delta = (2m)^{-1}$ and

$$\sum_{j=1}^{m} d_j d_{j+k} = -(2m)^{-1} \quad (1 \leqslant |k| \leqslant m).$$

Therefore the minimum asymptotic variance using an mth order difference sequence is $n^{-1}\tau_1^2$, where

$$\tau_1^2 = \text{var}\,(e^2) + m^{-1}\sigma^4. \tag{2·3}$$

Appendix 2 proves these results, and Appendix 3 discusses the computation of Table 1.

The trend in Table 1, as m increases, is for the difference sequence to converge to a 'spike' of unit mass at one of the entries, and to converge to zero everywhere else. To

Table 1. *Optimal difference sequences for* $1 \leqslant m \leqslant 10$. *Entries are rounded to four decimal places*

m	$(d_0, \ldots, d_m)$
1	(0·7071, −0·7071)
2	(0·8090, −0·5, −0·3090)
3	(0·1942, 0·2809, 0·3832, −0·8582)
4	(0·2708, −0·0142, 0·6909, −0·4858, −0·4617)
5	(0·9064, −0·2600, −0·2167, −0·1774, −0·1420, −0·1103)
6	(0·2400, 0·0300, −0·0342, 0·7738, −0·3587, −0·3038, −0·3472)
7	(0·9302, −0·1965, −0·1728, −0·1506, −0·1299, −0·1107, −0·0930, −0·0768)
8	(0·2171, 0·0467, −0·0046, −0·0348, 0·8207, −0·2860, −0·2453, −0·2260, −0·2879)
9	(0·9443, −0·1578, −0·1429, −0·1287, −0·1152, −0·1025, −0·0905, −0·0792, −0·0687, −0·0588)
10	(0·1995, 0·0539, 0·0104, −0·0140, −0·0325, 0·8510, −0·2384, −0·2079, −0·1882, −0·1830, −0·2507)

appreciate why, note that if we had the opportunity of observing the errors $e_1, \ldots, e_n$ then we would doubtlessly use $\hat{\sigma}_0^2 = n^{-1} \Sigma_j e_j^2$ to estimate σ^2. This estimator has variance $n^{-1}\tau_0^2$, where $\tau_0^2 = \text{var}(e^2)$; compare this with (2·1) and (2·3). It has only e_j, and none of the other errors, in the position of the jth summand. For large m, our optimal difference-based estimator $\hat{\sigma}^2$ is trying to emulate the performance of $\hat{\sigma}_0^2$, and so each summand is dominated by the contribution from just one data value. However, it is curious that the value chosen for emphasis is in the middle of the moving average when m is even, and at the very end of the sequence for odd m.

We might artificially construct an mth order 'spike' difference sequence by forcing a d_j towards the middle of the sequence to assume a value close to unity, and demanding that all the others be close to zero. For example, if $m = 2\nu$ then we might define

$$d_j = \begin{cases} \{2\nu/(2\nu+1)\}^{\frac{1}{2}} & (j = \nu), \\ -\{2\nu(2\nu+1)\}^{-\frac{1}{2}} & (0 \leqslant j \leqslant \nu-1 \text{ or } \nu+1 \leqslant j \leqslant 2\nu), \\ 0 & \text{otherwise;} \end{cases} \tag{2·4}$$

and, if $m = 2\nu - 1$,

$$d_j = \begin{cases} \{(2\nu-1)/(2\nu)\}^{\frac{1}{2}} & (j = \nu), \\ -\{2\nu(2\nu-1)\}^{-\frac{1}{2}} & (0 \leqslant j \leqslant \nu-1 \text{ or } \nu+1 \leqslant j \leqslant 2\nu-1), \\ 0 & \text{otherwise.} \end{cases} \tag{2·5}$$

It may be proved after some tedious algebra that for this difference sequence, the value of δ, defined at (2·2), is given by

$$\delta = \begin{cases} \{6\nu(2\nu+1)\}^{-1}(10\nu+7) & (m = 2\nu), \\ \{6\nu(2\nu-1)^2\}^{-1}(20\nu^2-18\nu+1) & (m = 2\nu-1). \end{cases}$$

Note particularly that $\delta \to 0$ as $m \to \infty$, indeed $\delta = O(m^{-1})$, just as in the case of the optimal difference sequence. Therefore the spike sequence and the optimal sequence have similar properties for large m. However, for small values of m the optimal sequence performs substantially better than the spike sequence, as we shall shortly show.

One might be tempted to use the 'ordinary' difference sequence commonly employed for numerical differentiation,

$$d_j = \begin{cases} \binom{2m}{m}^{-\frac{1}{2}} \binom{m}{j} (-1)^j & (0 \leqslant j \leqslant m), \\ 0 & \text{otherwise.} \end{cases} \tag{2·6}$$

Here the square root factor serves to ensure that $\Sigma\, d_j^2 = 1$. Unfortunately, this sequence performs very badly, particularly for large m. The reason is that it does not enjoy the 'spike' property. It is easily checked that the d_j's defined by (2·6) converge uniformly to zero, without any trace of a spike:

$$\max_{-\infty < j < \infty} |d_j| \to 0$$

as $m \to \infty$. It may be proved that, for the ordinary difference sequence defined at (2·6),

$$\delta = \binom{2m}{m}^{-2} \binom{4m}{2m} - 1 \sim (\tfrac{1}{2}\pi m)^{\frac{1}{2}}.$$

Table 2. *Comparison of optimal, spike and ordinary difference sequences*

m	Type of difference	δ	Eff. (%)
1	Optimal	$\frac{1}{2}$	67
	Spike	$\frac{1}{2}$	67
	Ordinary	$\frac{1}{2}$	67
2	Optimal	$\frac{1}{4}$	80
	Spike	$\frac{17}{18}$	51
	Ordinary	$\frac{17}{18}$	51
3	Optimal	$\frac{1}{6}$	86
	Spike	$\frac{5}{12}$	71
	Ordinary	$\frac{131}{100}$	43
4	Optimal	$\frac{1}{8}$	89
	Spike	$\frac{9}{20}$	69
	Ordinary	$\frac{797}{490}$	38
5	Optimal	$\frac{1}{10}$	91
	Spike	$\frac{127}{430}$	70
	Ordinary	$\frac{30313}{15876}$	34

Spike differences defined by (2·4) and (2·5), ordinary differences by (2·6). Fourth column lists efficiency of $\hat{\sigma}^2$ for normally distributed errors.

Therefore, far from converging to zero as $m \to \infty$, δ diverges to $+\infty$. This means that, if the ordinary difference sequence is used, asymptotic performance of the estimator becomes increasingly poor as m increases, a most undesirable property.

Table 2 lists values of δ for optimal, spike and ordinary difference sequences over the range $1 \leqslant m \leqslant 5$. It also gives the efficiency of $\hat{\sigma}^2$ relative to $\hat{\sigma}_0^2$ in the case of normally distributed errors. This tabulation demonstrates the strikingly good performance of the optimal difference sequence relative to both the others, for $2 \leqslant m \leqslant 5$.

Note that in the case of normal data, and for an estimator computed using optimal mth order differences, the efficiency of $\hat{\sigma}^2$ relative to $\hat{\sigma}_0^2$ is

$$\frac{\tau_0^2}{\tau_1^2} = \frac{2}{2+m^{-1}} = \frac{2m}{2m+1},$$

using (2·3) and the fact that $\text{var}\,(e^2) = 2\sigma^4$.

In conclusion we should note the role played by kurtosis κ in formula (2·1). Of course $\kappa = 0$ in the case of normal errors, and $\kappa > 0$ for error distributions which have heavier tails than the normal. Since the efficiency of $\hat{\sigma}^2$ relative to $\hat{\sigma}_0^2$ is $(\kappa+2)/(\kappa+2+2\delta)$, which is an increasing function of κ, then the efficiency actually improves as the tails of the error distribution become heavier. Thus, for practical purposes the ratio $2m/(2m+1)$ may be regarded as a lower bound to the efficiency of the optimal mth-order difference sequence.

Acknowledgement

This research was supported by a Visiting Fellowship Research Grant for P. Hall from the UK Science and Engineering Research Council and was conducted in the Department of Statistics at the University of Glasgow.

APPENDIX 1

Asymptotic formula for var $(\hat{\sigma}^2)$

Assume the conditions

$$E(e^4)<\infty, \quad E(e)=0, \quad E(e^2)=\sigma^2, \tag{A·1}$$

and that, for some $0<\varepsilon<\frac{1}{2}$, $C>0$, and all x, y,

$$|f(x)-f(y)| \leq C|x-y|^{\frac{1}{2}+\varepsilon}, \tag{A·2}$$

$$\max_{1\leq i\leq n-1} |x_{i+1}-x_i| = O(n^{-1+\varepsilon}). \tag{A·3}$$

Condition (A·2) is weaker than the assumption that f have a bounded derivative. Condition (A·3) holds for each $\varepsilon>0$ if the design is regularly spaced, e.g. if $x_i = i/n$ for $1\leq i\leq n$, or if the design is random on an interval I and the design density is bounded away from zero on I. The following theorem may be proved by routine methods.

THEOREM. *If* (A·1)-(A·3) *hold, and* τ^2 *is given by* (2·1), *then*

$$\text{var}\,(\hat{\sigma}^2) \sim E(\hat{\sigma}^2-\sigma^2)^2 \sim n^{-1}\tau^2$$

as $n\to\infty$.

APPENDIX 2

Selection of optimal mth order difference sequence

We begin by considering the related problem of constructing a moving average process with specified covariance structure. Assume that $d_0^2+\ldots+d_m^2=1$. Consider the moving average process

$$Y_k = \sum_{j=0}^{m} d_j e_{j+k} \quad (-\infty<k<\infty),$$

which has correlation function

$$\rho_k = E(Y_0 Y_k)(EY_0^2)^{-1} = \sum_{j=0}^{m} d_j d_{j+k}.$$

It may be shown by techniques which are standard in time series analysis that the correlation sequence defined by $\rho_0=1$, $\rho_k=-(2m)^{-1}$ for $1\leq|k|\leq m$, and $\rho_k=0$ for $|k|>m$, is allowable in the sense that there exists a sequence $\{d_j\}$ which produces these correlations. The argument uses the fact that $\Sigma\,\rho_k e^{ik\theta}$ is real-valued and nonnegative, and the important details are given by Anderson (1971, pp. 224–5).

Thus, we may produce a sequence $d_0,\ldots,d_m$ such that $d_0^2+\ldots+d_m^2=1$ and

$$\sum_{j=0}^{m} d_j d_{j+k} = -(2m)^{-1} \quad (1\leq|k|\leq m). \tag{A·4}$$

These two conditions imply that $0=\Sigma_j\,\Sigma_k\,d_j d_{j+k}=(\Sigma_j\,d_j)^2$, that is $d_0+\ldots+d_m=0$. Therefore our sequence $\{d_j\}$ has, in addition to (A·4), the properties

$$\sum_{j=0}^{m} d_j = 0, \quad \sum_{j=0}^{m} d_j^2 = 1. \tag{A·5}$$

Put

$$a_k = \sum_{j=0}^{m} d_j d_{j+k},$$

this time for a general sequence $d_0,\ldots,d_m$. Condition (A·5) is equivalent to

$$\sum_{k=1}^{m} a_k = -\tfrac{1}{2}, \quad a_0 = 1. \tag{A·6}$$

An optimal mth order difference sequence minimizes

$$\delta = \sum_{k \neq 0} \left(\sum_{j=0}^{m} d_j d_{j+k} \right)^2 = 2 \sum_{k=1}^{m} a_k^2$$

subject to (A·6). The minimum of δ subject to (A·6) occurs when $a_k = -(2m)^{-1}$ for $1 \leq k \leq m$, and the minimum equals $(2m)^{-1}$. We showed in the previous paragraph that there exists a difference sequence $\{d_j\}$ with the property $\Sigma_j d_j d_{j+k} = -(2m)^{-1}$, and (A·6), and so this sequence must produce the minimum.

Appendix 3

Computation of optimal difference sequences

For general m, observe that

$$D(d_0, \ldots, d_m) = \tfrac{1}{2}\delta = \tfrac{1}{2} \sum_{k \neq 0} \left(\sum_j d_j d_{j+k} \right)^2$$

$$= (d_0 d_m)^2 + (d_0 d_{m-1} + d_1 d_m)^2 + \ldots + (d_0 d_1 + \ldots + d_{m-1} d_m)^2. \qquad \text{(A·7)}$$

Define

$$s_1 = -(d_0 + d_m), \quad s_2^2 = 1 - (d_0^2 + d_m^2), \quad t_1 = (\tfrac{1}{2} - \tfrac{1}{4}s_1^2 - \tfrac{1}{2}s_2^2)^{\frac{1}{2}}.$$

Then $d_0 = -\frac{1}{2}s_1 + t_1$, $d_m = -\frac{1}{2}s_1 - t_1$. Using these formulae for d_0 and d_m, but taking $s_1 = d_1 + \ldots + d_{m-1}$ and $s_2^2 = d_1^2 + \ldots + d_{m-1}^2$, and substituting for d_0 and d_m in (A·7), we obtain an expression for $D(d_0, \ldots, d_m)$ as a function of $d_1, \ldots, d_{m-1}$ alone. This formula incorporates the constraints $\Sigma d_j = 0$ and $\Sigma d_j^2 = 1$, and may be minimized over $d_1, \ldots, d_{m-1}$ by using a standard optimization routine. We used NAG routine EØ4 JAF.

When $m = 2$, the conditions $\Sigma d_j = 0$ and $\Sigma d_j^2 = 1$ entail

$$D(d_0, d_1, d_2) = \{(d_0 + d_2)^2 - \tfrac{1}{2}\}^2 + (d_0 + d_2)^4.$$

The function $(x^2 - \frac{1}{2})^2 + x^4$ is minimized with $x = \pm\frac{1}{2}$, and so we should take $d_0 + d_2 = \pm\frac{1}{2}$. Thus, $d_1 = -(d_0 + d_2) = \mp\frac{1}{2}$, and $1 = d_0^2 + d_1^2 + d_2^2 = d_0^2 + \frac{1}{4} + (\frac{1}{2} \mp d_0)^2$. Solving this quadratic equation for d_0 we deduce that $(d_0, d_1, d_2) = (\frac{1}{4}(5^{\frac{1}{2}} + 1), -\frac{1}{2}, -\frac{1}{4}(5^{\frac{1}{2}} - 1))$, or one of the sign-changed and/or order-reversed variants.

When $m = 3$ it is convenient to make use of the identities $\Sigma d_j d_{j+k} = -\frac{1}{6}$ $(1 \leq k \leq 3)$ and $d_0^2 + d_1^2 + d_2^2 + d_3^2 = 1$. The equation for $k = 3$ implies that $d_3 = -(6d_0)^{-1}$. Substituting into the equations for $k = 1$ and $k = 2$ we may now express d_2 as a function of d_1 and d_0 alone. Using these formulae for d_2 and d_3 in either of the $k = 1$ and $k = 2$ equations, we produce a quadratic in d_1 with coefficients depending only on d_0. Solving for d_1, we may now express d_1, d_2, d_3 as functions of d_0 alone. The value of d_0 may be obtained from $d_0^2 + d_1^2 + d_2^2 + d_3^2 = 1$.

References

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

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

EAGLESON, G. K. (1990). Curve estimation—whatever happened to the variance? In *Proc 47th Session of the International Statistical Institute.* To appear.

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

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

MÜLLER, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*, Springer Lecture Notes in Statistics 46. New York: Springer.

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

MÜLLER, H.-G. & STADTMÜLLER, U. (1989). Detecting dependencies in smooth regression models. *Biometrika* **75**, 639–50.

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

[*Received September* 1989. *Revised January* 1990]