Biometrika (1988), **75**, 2, *pp.* 189–99
Printed in Great Britain

The estimation of residual variance in nonparametric regression

BY M. J. BUCKLEY AND G. K. EAGLESON
CSIRO Division of Mathematics and Statistics, Lindfield, N.S.W. 2070, Australia

AND B. W. SILVERMAN
School of Mathematical Sciences, University of Bath, Bath BA2 7AY, U.K.

SUMMARY

A wide class of estimators of the residual variance in nonparametric regression is considered, namely those that are quadratic in the data, unbiased for linear regression, and always nonnegative. The minimax mean squared error estimator over a natural class of regression functions is derived. This optimal estimator has an interesting structure and is closely related to a minimax estimator of the regression curve itself.

Some key words: Curve estimation; Minimax; Roughness penalty; Smoothing; Spline.

1. INTRODUCTION

1·1. *The regression problem*

Consider the regression problem where we have observations Y_i at design points t_i for $i=1,\dots,n$, and the observations are assumed to satisfy

$$Y_i = g(t_i)+\varepsilon_i \quad (i=1,\dots,n).$$

It will be assumed that $0<t_1<\dots<t_n<1$. The errors ε_i will be assumed to be independently normally distributed with zero mean and variance σ^2. In this paper we consider the important problem of estimating σ^2. Apart from the intrinsic interest in σ^2 as a parameter of the model, an estimate of the variance is essential in order to make inferences about the curve g itself.

Reinsch (1967) proposed choosing, as the estimate of g, the curve that minimized $\Sigma\{Y_i-g(t_i)\}^2$ subject to

$$\int_0^1 \{g''(t)\}^2\,dt \leqslant C. \qquad (1·1)$$

The justification for (1·1) is that one wishes to obtain the best possible fit to the data subject to the curve not exhibiting too much local variation as measured by its integrated squared second derivative. The solution is called the spline smoothing estimate of g; the bound C is a smoothing parameter that determines how much the data are smoothed to produce the estimate. Just as Reinsch confined attention to curves g satisfying the above bound, we also will impose this restriction in the development of our estimates of σ^2. The formulation (1·1) is equivalent to the penalized least-squares approach reviewed by Silverman (1985). Many other methods for nonparametric regression are referred to in that paper and its published discussion.

1·2. *Estimating the variance*

One possible approach to the problem of estimating σ^2 is to use the idea, common in time series analysis, of differencing the data to remove trend. For example Rice (1984) suggested estimating σ^2 by

$$\tfrac{1}{2}(n-1)^{-1}\sum_{i=2}^{n}(Y_i - Y_{i-1})^2.$$

A recent paper of Gasser, Sroka & Jenner (1986) has suggested a similar idea for removing local trend effects by using

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

where $\hat{\varepsilon}_i$ is the difference between Y_i and the value at t_i of the line joining (t_{i-1}, Y_{i-1}) and (t_{i+1}, Y_{i+1}). The c_i^2 are chosen to ensure that $E(c_i^2\hat{\varepsilon}_i^2) = \sigma^2$ for all i when g is linear. This estimate is a weighted sum of squared second differences of the data and is in fact identical to a second estimate proposed by Rice (1984), the average of the one-degree-of-freedom estimates obtained by fitting a least-squares line to successive triples of points.

All the other suggestions for estimating σ^2 are based on the residual sum of squares about some estimator $\hat{g}$ of the curve g. For example, Wahba (1978, 1983) used a normalized residual sum of squares about the Reinsch smoothing spline estimate of g. Breiman & Meisel (1976) suggested a procedure where $\hat{g}$ is a piecewise linear, but not necessarily continuous, fit to the data. Cleveland (1979) used a robust local estimate of g.

Our approach is not to restrict attention to either of the two classes of methods outlined above, but rather to consider all estimators that are quadratic functions of the data and satisfy some other mild conditions. Since the regression function is unknown, it is desirable to use an estimator which will be effective for a wide class of regressions. We shall derive the estimator of σ^2 which has minimax mean squared error over a natural class of functions g. To elucidate the asymptotic behaviour of this estimate, and also to illustrate some connections with the estimation of g itself, we also investigate the minimax estimator for a criterion distinct from, but closely related to, mean squared error.

Although we shall concentrate on the case of one-dimensional regression, our theorems on the form of the optimal estimator apply equally to any problem with a quadratic constraint on g, and in particular to any univariate or multivariate nonparametric regression where a quadratic roughness penalty is used. Such problems include the thin-plate spline approach to spatial smoothing discussed briefly by Silverman (1985, § 8.2) and the rather nonstandard branching regression discussed by Silverman & Wood (1987).

2. MAIN RESULTS

2·1. *A general class of estimators*

In this section we provide a discussion and description of our main results. Technical details and proofs are postponed to § 3 below.

Before discussing what might be a reasonable measure of the efficacy of a particular estimator of σ^2, we first define the class of estimators which we are willing to consider. We shall search for estimators of σ^2 amongst those that are:

(i) quadratic in the data;

(ii) always nonnegative;

(iii) unbiased for σ^2 if $g(.)$ is a straight line.

Conditions (i) and (ii) restrict us to considering estimates that are multiples of the residual sum of squares of the data about some linear transformation AY of the data. Typically AY will be an estimate of the mean vector g, defined by $g_i = g(t_i)$ for each $i = 1, \ldots, n$. For example, the estimate of Wahba (1978) is of this kind with AY equal to the vector of values taken by a smoothing spline at the design points. Also included in the class of nonnegative-definite quadratic estimates are those of the kind considered by Rice (1984) and Gasser et al. (1986). However, estimates based on nonlinear smoothers, such as those of Cleveland (1979), are not, in general, quadratic in the data. Nevertheless a very wide range of estimators do fall within the class we consider.

The condition (iii) ensures that the unbiasedness properties of the standard variance estimators based on classical linear regression are not destroyed by the weakening of parametric assumptions. It is easily seen that the estimators considered must be of the form

$$\hat{\sigma}^2 = Y^{\mathrm{T}} D Y / \{\operatorname{tr}(D)\}, \qquad (2\cdot1)$$

where D is a symmetric $n \times n$ nonnegative-definite matrix. The divisor, $\operatorname{tr}(D)$, in (2·1) is a necessary consequence of condition (iii): if $g(t) \equiv 0$ we have $E(Y^{\mathrm{T}}DY) = E(\varepsilon^{\mathrm{T}} D \varepsilon) = \sigma^2 \operatorname{tr}(D)$.

The moments of $\hat{\sigma}^2$ can be calculated for a general g. We have

$$\hat{\sigma}^2 = (g+\varepsilon)^{\mathrm{T}} D (g+\varepsilon)/\operatorname{tr}(D) = (g^{\mathrm{T}} D g + 2 g^{\mathrm{T}} D \varepsilon + \varepsilon^{\mathrm{T}} D \varepsilon)/\operatorname{tr}(D)$$

and hence, by equation (15.47) of Kendall & Stuart (1977, p. 382),

$$E(\hat{\sigma}^2) = \sigma^2 + g^{\mathrm{T}} D g / \operatorname{tr}(D) \qquad (2\cdot2)$$

and, since $g^{\mathrm{T}} D \varepsilon$ and $\varepsilon^{\mathrm{T}} D \varepsilon$ are uncorrelated,

$$\operatorname{var}(\hat{\sigma}^2) = \{4\sigma^2 g^{\mathrm{T}} D^2 g + 2\sigma^4 \operatorname{tr}(D^2)\}/\{\operatorname{tr}(D)\}^2. \qquad (2\cdot3)$$

It is a consequence of conditions (ii) and (iii) that $\hat{\sigma}^2$ must be zero if the data lie exactly on a straight line. It follows that, in the special case where $g(.)$ is a straight line, $g^{\mathrm{T}} D g = 0$ and hence, by standard linear algebra, $Dg = 0$. It then follows that, for such g,

$$\operatorname{var}(\hat{\sigma}^2) = 2\sigma^4 \operatorname{tr}(D^2)/\{\operatorname{tr}(D)\}^2.$$

In our subsequent discussion we denote by Δ the set of symmetric $n \times n$ matrices D for which the estimate (2·1) satisfies conditions (i) to (iii). The set Δ depends, of course, on the design points, $\{t_1, \ldots, t_n\}$.

2·2. *The minimax mean squared error estimator*

Without some restriction on the regression curve $g(.)$ it is impossible to distinguish between signal and noise and therefore the regression problem cannot be properly posed. In order to make some progress, we restrict attention to a subclass of regressions, namely to those $g(.)$ for which (1·1) holds.

This restricts attention to regression functions $g(.)$ which are reasonably smooth, without forcing $g(.)$ into a finite-dimensional parametric family. We show in § 3·1 below that the condition can be replaced by one on the values taken by $g(.)$ at the design points, namely $g^{\mathrm{T}} \Omega g \leq C$, where Ω is a particular member of Δ which we shall describe later.

Define the mean squared error criterion $M(D)$ by

$$M(D) = \max_{g^{\mathrm{T}}\Omega g \leqslant C} \{E(\hat{\sigma}^2 - \sigma^2)^2\}$$

$$= \left[\max_{g^{\mathrm{T}}\Omega g \leqslant C} \{(g^{\mathrm{T}}Dg)^2 + 4\sigma^2 g^{\mathrm{T}}D^2 g\} + 2\sigma^4 \operatorname{tr}(D^2)\right] \Big/ \{\operatorname{tr}(D)\}^2, \qquad (2\cdot4)$$

substituting the expressions (2·2) and (2·3) for the bias and variance of $\hat{\sigma}^2$.

To describe the matrix that minimizes $M(D)$ over D in Δ, let $\omega_1 \leqslant \ldots \leqslant \omega_n$ be the eigenvalues of Ω and suppose $\phi_1, \ldots, \phi_n$ are the corresponding eigenvectors, orthonormalized so that

$$\sum_{i=1}^{n} \phi_{ji}\phi_{ki} = \begin{cases} 1 & (j = k), \\ 0 & \text{otherwise.} \end{cases}$$

It is known (Speckman, 1985) that $\omega_1 = \omega_2 = 0$ and ω_j is approximately $n^{-1}\rho j^4$, where ρ is a constant depending only on the design. If $t_{i+1} - t_i = n^{-1}$ for each i, then $\rho = \pi^4$.

The optimal matrix $D^\dagger$ is then specified by the following theorem, the proof of which is given in § 3·2 below.

Theorem 2·1. *For each $\alpha \geqslant 0$ let $D_\alpha^\dagger$ be the matrix with the same eigenvectors as Ω, with corresponding eigenvalues*

$$d_i^\dagger(\alpha) = \min\{1, \alpha\omega_i(1 + 4\sigma^2\omega_i/C)^{-\frac{1}{2}}\} \quad (i = 1, \ldots, n). \qquad (2\cdot5)$$

Then $D_\alpha^\dagger$ is in Δ for all $\alpha > 0$, and there is a positive quantity a depending on C/σ^2, n and the design such that $D_a^\dagger$ minimizes $M(D)$ over all D in Δ.

For any particular 'signal-to-noise ratio' nC/σ^2, the value of a, and hence the optimal estimator itself, can be found by minimizing $M(D_\alpha^\dagger)$ numerically. The value of $M(D_\alpha^\dagger)$ is given by

$$\{C^2\alpha^2 + 2\sigma^4 \sum d_i^\dagger(\alpha)^2\} / \{\sum d_i^\dagger(\alpha)\}^2$$

as will be shown in the course of the proof of Theorem 2·1.

While this solves the problem of finding the minimax estimator, for two reasons we shall investigate the estimator that is minimax under a slightly different accuracy criterion. One is in order to understand some asymptotic properties of $D_a^\dagger$ itself, and the other is to explore the relationship with a minimax estimator of the curve g itself.

2·3. *An alternative criterion and its minimax estimator*

For any D in Δ, define a criterion $K(D)$ by

$$K(D) = \left\{\left(\max_{g^{\mathrm{T}}\Omega g \leqslant C} g^{\mathrm{T}}Dg\right)^2 + 2\sigma^4 \operatorname{tr}(D^2)\right\} \Big/ \{\operatorname{tr}(D)\}^2.$$

This criterion is obtained from $M(D)$ by omitting the term $4\sigma^2 g^{\mathrm{T}}D^2 g$ from the maximum, and is the sum of the maximum squared bias of the estimator over g subject to (1·1) and the variance of $\hat{\sigma}^2$ for linear g.

The form of the optimal matrix D^* for this criterion is then specified by the following theorem.

Theorem 2·2. *For each $\beta \geqslant 0$ let D_β^* be the matrix with the same eigenvectors as Ω, with corresponding eigenvalues*

$$d_i^*(\beta) = \min(1, \beta\omega_i) \quad (i = 1, \ldots, n).$$

Then D^*_β is in Δ for all $\beta > 0$, and there is a positive b depending on C/σ^2, n and the design such that D^*_b minimizes $K(D)$ over all D in Δ.

We show in § 3·4 that the minimizer of $K(D)$ is very nearly a minimizer of $M(D)$ as well. It is important to bear this in mind in § 2·4 where the minimizer of $K(D)$ is discussed further.

2·4. *Minimax estimation of the curve itself*

We have developed minimax approaches to the estimation of the variance. It is interesting to make comparisons with some previous work on the estimation of the curve itself. The minimax estimate of the curve described in this section is due to Speckman (1985); for related work see also Nussbaum (1985) and the references in those two papers.

Given any estimator $\hat{g}$ of the curve g, define the expected summed squared error

$$T(\hat{g}, g) = E\left[\sum_{i=1}^{n} \{\hat{g}(t_i) - g(t_i)\}^2\right].$$

Consider, now, the class of estimators $\hat{g}$ that are linear in the data Y_i.

For each $j = 1, \ldots, n$, define a function $\phi_j(t)$ for t in $[0, 1]$ to be the natural cubic interpolating spline to the values $\phi_j(t_k) = \phi_{jk}$. Demmler & Reinsch (1975) show that the ϕ_j become progressively more oscillatory as j increases. That is

$$\int_0^1 \{\phi_j''(t)\}^2\, dt$$

increases as j increases.

Speckman (1985) shows that the linear estimator that minimizes the maximum of $T(\hat{g}, g)$ over g subject to (1·1) is given by

$$\hat{g}_\gamma(x) = \sum_{\nu=1}^{n} \{1 - \surd(\gamma\omega_\nu)\}_+ \tilde{Y}_\nu \phi_\nu(x).$$

Here γ is a smoothing parameter which depends on C/σ^2, n and the design, and the $\tilde{Y}_\nu$ are the coefficients of the expansion of the data points Y_i in terms of the eigenvectors ϕ_ν; that is, $Y_i = \Sigma\, \tilde{Y}_\nu \phi_{\nu i}$, with the sum over $\nu = 1, \ldots, n$, for each i, so that by the orthonormality of the eigenvectors $\tilde{Y}_\nu - \Sigma\, \phi_{\nu j} Y_j$, with the sum over $j = 1, \ldots, n$. Speckman essentially shows that γ is specified approximately by letting ν_1 be an integer such that

$$\nu_1 \doteq \{(15nC)/(2\rho\sigma^2)\}^{1/5} \qquad (2{\cdot}6)$$

and setting $\gamma = 1/\omega_{\nu_1}$.

The residual sum of squares about $\hat{g}_\gamma(.)$ satisfies

$$\text{RSS}\,(\hat{g}_\gamma) = \sum_{i=1}^{n} \{Y_i - \hat{g}_\gamma(t_i)\}^2 = Y^{\mathrm{T}}\Phi^{\mathrm{T}}\,\mathrm{diag}\,\{\min\,(1, \gamma\omega_\nu)\}\Phi\, Y, \qquad (2{\cdot}7)$$

where Φ is the matrix with elements ϕ_{ij}. Because Φ is formed of the eigenvectors of Ω, it follows from (2·7) and the definition of D^* in Theorem 2·2 that $\text{RSS}\,(\hat{g}_\gamma) = Y^{\mathrm{T}} D^*_\gamma Y$.

This provides an interesting connection between Speckman's work and our own. Let σ_b^{*2} be the optimal estimator for the criterion $K(D)$,

$$\sigma_b^{*2} = Y^{\mathrm{T}} D_b^* Y/\mathrm{tr}\,(D_b^*).$$

We show in § 3·3 that the value of b is specified approximately as $b = 1/\omega_{\nu_0}$, where ν_0 is an integer satisfying

$$\nu_0 \simeq \{(45C^2n^2)/(8\rho^2\sigma^4)\}^{1/9}. \tag{2·8}$$

The integer ν_0 is the 'cut-off point' where the eigenvalues $d_i^*(b)$ become equal to 1. Our estimator σ_b^{*2} is the normalized residual sum of squares about a Speckman estimator of the curve itself. However, for the same value of C/σ^2 and the same design, the smoothing parameters b and γ, or equivalently the cut-off points ν_0 and ν_1, have different values. Some elementary manipulation from (2·6) and (2·8) shows that

$$\nu_0 \simeq \nu_1(\nu_1/10)^{1/9}. \tag{2·9}$$

Thus, in large samples, ν_0 will be larger than the optimal ν_1 for the estimation of the curve itself, and the variance estimator σ_b^{*2} will be constructed from an undersmoothed curve. However, in many practical situations, particularly bearing in mind all the approximations leading to (2·9), the message of (2·9) is, perhaps, that little will be lost by setting $\nu_0 = \nu_1$ and hence basing the estimate of σ^2 directly on the Speckman estimate of the curve itself.

3. Technical details

3·1. *Preliminaries*

In this section further details of various technical matters are given. First we explain the genesis of the matrix Ω and demonstrate why the conditions (1·1) and $g^T\Omega g \leq C$ are equivalent. In § 3·2 the proofs of the main theorems are given, and in § 3·3 the asymptotic expression (2·8) for ν_0 is justified. Finally in § 3·4 the mean squared error properties of the estimator based on the criterion $K(D)$ are discussed.

Let S be the space of natural cubic splines $s(.)$ on $[0, 1]$ with knots at the design points. This is to say that each $s(.)$ is a piecewise cubic, twice continuously differentiable everywhere and with joins between the cubics at the design points; furthermore the second and third derivatives of s are zero outside $[t_1, t_n]$. Demmler & Reinsch (1975) show that there is a basis for S, $\{\phi_1(.), \ldots, \phi_n(.)\}$, determined essentially uniquely by

$$\sum_{i=1}^{n} \phi_j(t_i)\phi_k(t_i) = \delta_{jk}, \quad \int_0^1 \phi_j''(t)\phi_k''(t)\,dt = \delta_{jk}\omega_k,$$

with $0 = \omega_1 = \omega_2 < \ldots < \omega_n$. Here $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. The ω_k are the eigenvalues of Ω and the first two basis functions $\phi_1(t)$ and $\phi_2(t)$ span the space of linear functions. The eigenvectors ϕ_i of Ω are defined by $\phi_{ij} = \phi_i(t_j)$ for all i and j, and hence, since every member of S is determined by its values at the design points, the definition given in § 2·3 of $\phi_i(t)$ as the natural spline interpolator of the values ϕ_{ij} is justified.

Given any spline function $s(.)$ in S, let $s(.) = \Sigma\, \tilde{s}_\nu \phi_\nu(.)$ be the expansion of s in terms of the basis $\phi_\nu(.)$. It is immediate that

$$\int_0^1 \{s''(t)\}^2\,dt = \sum \omega_\nu \tilde{s}_\nu^2.$$

Now, given any function $g(.)$ with absolutely continuous derivative, define a vector $g_i = g(t_i)$ $(i = 1, \ldots, n)$ and let $\tilde{g}_\nu$ be the coefficients of the expansion of g in terms of the eigenvectors ϕ_ν, so that $g_i = \Sigma\, \tilde g_\nu \phi_{i\nu}$ for all i. Then the function $s_g(t) = \Sigma\, \tilde g_\nu \phi_\nu(t)$ is the natural interpolating spline to the values $g(t_i)$, and hence, by standard spline theory,

$$\int_0^1 \{g''(t)\}^2\, dt \geq \int_0^1 \{s_g''(t)\}^2\, dt = \sum \omega_\nu \tilde g_\nu^2 = g^{\mathrm{T}}\Omega g. \qquad .$$

It now follows that $g^{\mathrm{T}}\Omega g \leq C$ if and only if there exists a curve $g(.)$ with

$$\int_0^1 \{g''(t)\}^2\, dt \leq C, \quad g(t_i) = g_i \quad (i = 1, \ldots, n).$$

Hence for any functional $\tau(g)$ which depends on g only through the values $g(t_1), \ldots, g(t_n)$, the maxima of $\tau(g)$ subject to (1·1) and subject to $g^{\mathrm{T}}\Omega g \leq C$ are the same.

3·2. *Proof of the theorems*

The proof of Theorem 2·1 proceeds in two stages. First we prove that the minimizer of $M(D)$ over D in Δ may be taken to have the same eigenvectors as Ω. This stage is related to an argument given in the Appendix of Speckman (1985).

Given any vector g, let $\tilde g = \Phi g$ so that g has the eigenvector expansion $g = \Sigma\, \tilde g_\nu \phi_\nu$. Given any D in Δ, let $\tilde D = \Phi D \Phi^{\mathrm{T}}$, the representation of D in the basis $\{\phi_i\}$, and let $\tilde\Delta$ be the space of possible $\tilde D$ with D in Δ; that is, $\tilde\Delta = \Phi\Delta\Phi^{\mathrm{T}}$. Then $g^{\mathrm{T}}\Omega g = \Sigma\, \omega_\nu \tilde g_\nu^2$ so that the problem of minimizing $M(D)$ is equivalent to finding $\tilde D$ to solve

$$\min_{\tilde D \in \tilde\Delta} \left[\max_{\Sigma \omega_\nu \tilde g_\nu^2 \leq C} \{(\tilde g^{\mathrm{T}} \tilde D \tilde g)^2 + 4\sigma^2 \tilde g^{\mathrm{T}} \tilde D^2 \tilde g\} + 2\sigma^4 \operatorname{tr}(\tilde D^2) \right]$$

subject to $\operatorname{tr}(\tilde D) = 1$. Replace the constraint $\operatorname{tr}(\tilde D) = 1$ by a Lagrange multiplier term and define, for $\tilde D$ in $\tilde\Delta$,

$$L(\tilde D) = \max_{\Sigma \omega_\nu \tilde g_\nu^2 \leq C} \{(\tilde g^{\mathrm{T}} \tilde D \tilde g)^2 + 4\sigma^2 \tilde g^{\mathrm{T}} \tilde D^2 \tilde g\} + 2\sigma^4 \operatorname{tr}(\tilde D^2) - \lambda \operatorname{tr}(\tilde D). \qquad (3·1)$$

The optimizing $\tilde D$ will then minimize $L(\tilde D)$ for a suitably chosen λ.

Let $\omega_i^\dagger = \omega_i(1 + 4\sigma^2 \omega_i / C)^{-\frac{1}{2}}$ and define

$$L_0(\tilde D) = C^2 \max_{3 \leq i \leq n} (\tilde d_{ii}/\omega_i^\dagger)^2 + 2\sigma^4 \sum_{i=1}^{n} \tilde d_{ii}(\tilde d_{ii} - \lambda/2\sigma^4).$$

By restricting the maximum in (3·1) to the maximum over coordinate vectors, and using the fact that $\operatorname{tr}(\tilde D^2) \geq \Sigma\, \tilde d_{ii}^2$, it follows that $L(\tilde D) \geq L_0(\tilde D)$ for all $\tilde D \in \tilde\Delta$.

Since $\phi_1(t)$ and $\phi_2(t)$ are linear functions, requirements (ii) and (iii) of § 2·1 force $\tilde d_{11} = \tilde d_{22} = 0$. Now suppose that $\tilde D$ is diagonal, and write $d_i = \tilde d_{ii}$ for each i. Define $A_0(\tilde g)$ by

$$A_0(\tilde g) = (\tilde g^{\mathrm{T}} \tilde D \tilde g)^2 + 4\sigma^2 \tilde g^{\mathrm{T}} \tilde D^2 \tilde g = \left(\sum_{i=1}^{n} d_i \tilde g_i^2 \right)^2 + 4\sigma^2 \sum_{i=1}^{n} d_i^2 \tilde g_i^2$$

and, for $n \times 1$ vectors h, define $A(h)$ by

$$A(h) = \left(\sum_{i=1}^{n} d_i h_i \right)^2 + 4\sigma^2 \sum_{i=1}^{n} d_i^2 h_i.$$

Then on substituting $h_i = \tilde g_i^2$, it is clear that the maximum of $A_0(\tilde g)$ subject to $\Sigma\, \omega_i \tilde g_i^2 \leq C$ is identical with the maximum of $A(h)$ subject to the linear constraints $h_j \geq 0$ for all j and $\Sigma\, \omega_j h_j \leq C$.

The feasible region for h is a convex polyhedron $\mathscr{H}$ in $\mathscr{R}^n$, with extreme points the $(n-1)$ vectors $(0, \ldots, 0)$ and $(0, \ldots, C\omega_i^{-1}, 0, \ldots, 0)$ for $i = 3, \ldots, n$. Since the function A is convex, its maximum over $\mathscr{H}$ is attained at one of the extreme points, and so is equal to

$$\max_{3\leq i\leq n} (d_i^2 C^2 \omega_i^{-2} + 4\sigma^2 d_i^2 C\omega_i^{-1}) = C^2 \max_{3\leq i \leq n} (d_i/\omega_i^{\dagger})^2.$$

It follows that, for diagonal $\tilde{D}$, $L(\tilde{D}) = L_0(\tilde{D})$. Now let $\tilde{\Delta}_0$ be the space of diagonal matrices in $\tilde{\Delta}$. Given any $\tilde{D}$ in $\tilde{\Delta}$, define $\tilde{D}_0$ to be equal to $\tilde{D}$ on the main diagonal and zero elsewhere. Then $L(\tilde{D}) \geq L_0(\tilde{D}) = L_0(\tilde{D}_0) = L(\tilde{D}_0)$ since the value of $L_0(\tilde{D})$ depends only on the diagonal elements of $\tilde{D}$. Hence the minimizer of L over $\tilde{\Delta}_0$ will minimize L over $\tilde{\Delta}$. Note that $\tilde{D} \in \tilde{\Delta}_0$ if and only if the corresponding $D = \Phi^{\mathrm{T}}\tilde{D}\Phi$ is in Δ and has eigenvectors equal to those of Ω.

In the second stage of the proof we find the form of the optimal D. Suppose $\tilde{D} = \operatorname{diag}(0, 0, d_3, \ldots, d_n) \in \tilde{\Delta}_0$. Let $m = \max(d_i/\omega_i^{\dagger}: 3 \leq i \leq n)$ be fixed for the moment; then

$$L(\tilde{D}) = C^2 m^2 + 2\sigma^4 \sum_{i=3}^{n} d_i(d_i - \tfrac{1}{2}\lambda/\sigma^4). \tag{3·2}$$

Subject to $d_i/\omega_i^{\dagger} \leq m$, (3·2) is minimized by $d_i = \min(m\omega_i^{\dagger}, \tfrac{1}{4}\lambda/\sigma^4)$; requirement (ii) of § 2·1 implies that $d_i \geq 0$ for all i and hence $\lambda > 0$. The criterion $M(D)$ is unaltered by multiplying D by a constant, and hence (3·2) is minimized by specifying $d_i = \min(1, \alpha\omega_i^{\dagger}) = d_i^{\dagger}(\alpha)$ as defined in Theorem 2·1, in other words setting $D = D_\alpha^{\dagger}$ for some α as yet unspecified. Setting a equal to the minimizer over α of $M(D_\alpha^{\dagger})$ completes the proof of Theorem 2·1, since it is immediate that $\sigma^{-4}M(D_\alpha^{\dagger})$ depends only on C/σ^2, n and the design. □

Theorem 2·2 is proved in exactly the same way, leading to a minimizer of $K(D)$ of the form D_β^* for some suitable β. Since the eigenvalues ω_i are increasing, and since $d_i^*(\beta) = \min(1, \beta\omega_i)$, we have the equivalent specification

$$d_i^*(\beta) = \begin{cases} \beta\omega_i & (i \leq \nu_0), \\ 1 & (i > \nu_0), \end{cases} \tag{3·3}$$

where β is as yet unspecified and $\nu_0 = \max(i: \beta\omega_i \leq 1)$. To find β, note that

$$K(D_\beta^*)/\sigma^4 = \frac{\beta^2 C^2/\sigma^4 + 2\{\beta^2(\omega_1^2 + \ldots + \omega_{\nu_0}^2) + n - \nu_0\}}{\{\beta(\omega_1 + \ldots + \omega_{\nu_0}) + n - \nu_0\}^2}. \tag{3·4}$$

Minimizing (3·4) over β will then give the value of b. In § 3·3 we use approximate values of ω_i to simplify (3·4) and hence derive the approximate formula (2·8). □

3·3. *Approximating the cut-off point* ν_0

It has already been noted that the ω_j are approximately $n^{-1}\rho j^4$, where ρ is a constant depending only on the design. If the design points have limiting density $p(x)$ in some suitable sense, then (Speckman, 1985)

$$\rho = \pi^4 \left[\int \{p(x)\}^{1/4}\, dx\right]^{-4}$$

and (Silverman, 1984) a density estimation method could be used to estimate $p(x)$ and hence ρ. If the design points are equally spaced on $[a, b]$, then $p(x) = (b-a)^{-1}$ and $\rho = \pi^4(b-a)^{-3}$.

The sums in (3·4) can now be approximated by integrals. We have, since $\beta\omega_{\nu_0} \simeq 1$,

$$\beta^2 \sum_{i=1}^{\nu_0} \omega_i^2 \simeq \omega_{\nu_0}^{-2} \sum_{i=1}^{\nu_0} \omega_i^2 \simeq \sum_{i=1}^{\nu_0} (i/\nu_0)^8 \simeq \nu_0 \int_0^1 t^8\,dt = \tfrac{1}{9}\nu_0$$

and, similarly

$$\beta \sum_{i=1}^{\nu_0} \omega_i \simeq \sum_{i=1}^{\nu_0} (i/\nu_0)^4 \simeq \tfrac{1}{5}\nu_0.$$

Finally

$$\beta^2 \simeq 1/\omega_{\nu_0}^2 \simeq n^2\rho^{-2}\nu_0^{-8}.$$

Substitution of these approximations into (3·4) yields

$$K(D_\beta^*)/\sigma^4 \simeq \frac{n^2C^2\sigma^{-4}\rho^{-2}\nu_0^{-8} + 2(n - \frac{8}{9}\nu_0)}{(n - \frac{4}{5}\nu_0)^2}. \qquad (3\cdot5)$$

Let $\lambda_0(z)$ be the minimizer in $[0, 1]$ over λ of

$$\{z\lambda^{-8} + 2(1 - \tfrac{8}{9}\lambda)\}/(1 - \tfrac{4}{5}\lambda)^2.$$

Now $\lambda_0(z) = \min\{1, (45z/8)^{1/9}\}$ and hence the ν_0 that minimizes (3·5) is given approximately by $\nu_0/n \simeq \lambda_0(n^{-7}C^2\sigma^{-4}\rho^{-2})$, which yields, for sufficiently large n,

$$\nu_0 \simeq \{(45C^2n^2)/(8\rho^2\sigma^4)\}^{1/9} \qquad (3\cdot6)$$

as in (2·8). Substitution of this approximation into (3·5) yields the approximation

$$K(D_b^*) = 2n^{-1}\sigma^4\{1 + O(n^{-7/9})\} \qquad (3\cdot7)$$

which we use in § 3·4.

3·4. *Mean squared error optimality*

In this section we show that the matrix D_b^* is not only optimal for the criterion $K(D)$ but also very nearly optimal for the mean squared error criterion $M(D)$ defined in (2·4). Write, for D in Δ,

$$M(D, g) = E\{Y^{\mathrm{T}}DY/\mathrm{tr}\,(D) - \sigma^2\}^2,$$

given that the true regression is g; then

$$M(D) = \max\{M(D, g) \colon g^{\mathrm{T}}\Omega g \le C\}.$$

The quantity $M(D, g)$ is precisely the mean squared error $E(\hat\sigma^2 - \sigma^2)^2$ and, by (2·2) and (2·3),

$$M(D, g) = \{(g^{\mathrm{T}}Dg)^2 + 4\sigma^2 g^{\mathrm{T}}D^2g + 2\sigma^4\,\mathrm{tr}\,(D^2)\}/\{\mathrm{tr}\,(D)\}^2.$$

Let

$$K(D, g) = \{(g^{\mathrm{T}}Dg)^2 + 2\sigma^4\,\mathrm{tr}\,(D^2)\}/\{\mathrm{tr}\,(D)\}^2,$$

$$A(D, g) = M(D, g) - K(D, g) = 4\sigma^2 g^{\mathrm{T}}D^2g/\{\mathrm{tr}\,(D)\}^2.$$

By elementary manipulation

$$\min_{D\in\Delta} M(D) \le M(D_b^*) \le K(D_b^*) + \max_{g^{\mathrm{T}}\Omega g \le C} A(D_b^*, g). \qquad (3\cdot8)$$

Since $K(D, g) \leq M(D, g)$ for all D and g,

$$\min_{D\in\Delta} M(D) \geq \min_{D\in\Delta} K(D) = K(D_b^*). \tag{3·9}$$

The eigenvalues of D_b^{*2} are the squares of those given in (3·3) and hence

$$\max_{g^{\mathrm{T}}\Omega g \leq C} g^{\mathrm{T}} D_b^{*2} g = C \max_i [\{d_i^*(b)\}^2/\omega_i] \leq C/\omega_{\nu_0} \tag{3·10}$$

since the ω_i are increasing. Combine (3·10) with (3·8) and the approximation (3·7) to obtain

$$K(D_b^*)^{-1} \max_{g^{\mathrm{T}}\Omega g \leq C} A(D_b^*, g) \simeq 2(8/45)^{4/9}\{C/(\rho\sigma^2)\}^{1/9} n^{-8/9}. \tag{3·11}$$

It follows from (3·8), (3·9) and (3·11) that

$$\{M(D_b^*) - \min_{D\in\Delta} M(D)\}/\min_{D\in\Delta} M(D) = O(n^{-8/9}).$$

Thus the mean squared error attained by using D_b^* is within a relative error of $O(n^{-8/9})$ of the minimum possible $M(D)$.

It is interesting to compare the mean squared error of other estimators of variance with the minimax value we have obtained. For example, the estimator suggested by Gasser et al. (1986) satisfies the conditions of § 2·1, and elementary calculations give its asymptotic mean squared error on a uniform design as $35n^{-1}\sigma^4/9$, nearly twice that of our minimax estimator.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the helpful comments of C. Jennison and the financial support of the Science and Engineering Research Council, the National Science Foundation and the U.S. Army European Research Office.

REFERENCES

BREIMAN, L. & MEISEL, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Am. Statist. Assoc.* **71**, 301-7.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* **74**, 829-36.

DEMMLER, A. & REINSCH, C. (1975). Oscillation matrices with spline smoothing. *Numer. Math.* **24**, 375-82.

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

KENDALL, M. G. & STUART, A. (1977). *The Advanced Theory of Statistics*, **1**, 4th ed. London: Griffin.

NUSSBAUM, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13**, 984-97.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10**, 177-83.

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

SILVERMAN, B. W. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Am. Statist. Assoc.* **79**, 584-9.

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.

SILVERMAN, B. W. & WOOD, J. T. (1987). The nonparametric estimation of branching curves. *J. Am. Statist. Assoc.* **82**, 551-8.

SPECKMAN, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13**, 970-83.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc.* B **40**, 364-72.

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

[*Received June* 1987. *Revised October* 1987]