The Annals of Statistics
1985, Vol. 13, No. 3, 970–983

SPLINE SMOOTHING AND OPTIMAL RATES OF CONVERGENCE IN NONPARAMETRIC REGRESSION MODELS[1]

BY PAUL SPECKMAN

University of Missouri, Columbia

Linear estimation is considered in nonparametric regression models of the form $Y_i = f(x_i) + \varepsilon_i$, $x_i \in (a, b)$, where the zero mean errors are uncorrelated with common variance σ^2 and the response function f is assumed only to have a bounded square integrable qth derivative. The linear estimator which minimizes the maximum mean squared error summed over the observation points is derived, and the exact minimax rate of convergence is obtained. For practical problems where bounds on $\|f^{(q)}\|^2$ and σ^2 may be unknown, generalized cross-validation is shown to give an adaptive estimator which achieves the minimax optimal rate under the additional assumption of normality.

1. The model.

Consider the nonparametric regression model

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \cdots, n,$$

where observations are taken at distinct points on a finite interval $[a, b]$. The usual assumptions on the random errors are in force, i.e., $E\varepsilon_i = 0$, $E\varepsilon_i\varepsilon_j = \delta_{ij}\sigma^2$, where δ_{ij} is the Kronecker delta, but the response function f is assumed only to belong to a qth-order Sobolev space $W^q = \{f: f \text{ has } q-1 \text{ absolute continuous derivatives}, \|f^{(q)}\|^2 = \int_a^b (f^{(q)}(x))^2\,dx < \infty\}$.

The model is motivated by certain robustness considerations. For small $\alpha > 0$, the class $\mathscr{F}_{q,\alpha} = \{f \in W^q: \|f^{(q)}\| \leq \alpha\}$ can be viewed as a collection of response functions at least locally well-approximated by polynomials of degree $q-1$ (or order q). If a regression method is uniformly good within this class, it is robust to arbitrary small departures from the standard qth-order polynomial model. This concept of robustness is related to the models of Sacks and Ylvisaker (1978).

This paper deals with global convergence for linear estimators. Let $\mathscr{L}$ be the class of all estimators $\hat{f}(x)$ which are linear in the observations, and define

$$T(\hat{f}, f) = (1/n)\sum_{i=1}^n (\hat{f}(x_i) - f(x_i))^2.$$

In Section 2 a family of linear estimators $\hat{f}_\gamma$, $\gamma > 0$, is introduced which is optimal in the following sense. For any $\alpha > 0$ and $\sigma > 0$, there exists a $\gamma_0 = \gamma_0(\alpha, \sigma)$ such that

$$\min_{\hat{f}\in\mathscr{L}} \max_{f\in\mathscr{F}_{q,\alpha}} ET(\hat{f}, f) = \max_{f\in\mathscr{F}_{q,\alpha}} ET(\hat{f}_{\gamma_0}, f).$$

Received September 1981; October 1984.

[1] Research was supported in part by NSF Grant MCS 80-02754.

AMS 1980 *subject classifications.* Primary 62J05; secondary 62G35, 41A15.

Key words and phrases. Cross-validation, mean square linear estimation, nonparametric regression, splines.

Being minimax, $\hat{f}_\gamma$ is model robust in the sense of Sacks and Ylvisaker.

The behavior of the minimax estimator is studied in Section 3 for suitably regular sequences of observation sets, and the exact asymptotic rate of convergence is obtained. In particular, Theorem 3.4 shows that $O(n^{-2q/(2q+1)})$ is the best possible rate for any linear estimator. This is exactly the optimal global rate established by Stone (1982) for arbitrary estimators assuming q-times differentiable f. Stone's models are more general since there is no restriction to the linear case. However, under the assumptions here, the best in a large class of linear estimators with the "correct" rate of convergence is studied.

The minimax estimator is a variant of spline smoothing. To describe it, natural splines are defined in Section 2 and the spline basis of Demmler and Reinsch (1975) is introduced. The work here is related to the least squares and bias minimizing splines studied by Agarwal and Studden (1980), but the B-spline basis they employed is different. There is a close relationship between the minimax estimator and the usual smoothing spline of Reinsch (1967). The statistical properties of the latter method have received considerable attention recently (see, e.g., Wahba, 1975; Wahba and Wold, 1975; and Craven and Wahba, 1979), and some of the results here extend the work of these authors for ordinary smoothing splines.

Section 4 is devoted to a practical method for applying the estimator with the best uniform rate of convergence. The idea is to find an estimator which has the same asymptotic rate of convergence as the minimax one calculated when $\|f^{(q)}\|$ and σ^2 are known. The problem is to find a good estimate for γ, the parameter indexing the family of estimators, based on the data alone. This kind of problem seems to be intrinsic to any nonparametric method in one form or another. Here γ is estimated by the method of generalized cross-validation introduced in Craven and Wahba (1979) and Golub, Heath and Wahba (1979). The main result, considerably stronger than the comparable one Craven and Wahba obtained, is that there is a sequence $\hat{\gamma}_n$ depending only on the observations such that

$$ET_n(\hat{f}_{\hat{\gamma}_n}, f)/\inf_\gamma ET_n(\hat{f}_\gamma, f) \to 1$$

as $n \to \infty$ for suitably regular sequences of observations.

2. A basis for natural splines and the minimax estimator. Given a distinct knot set $\{x_1, \cdots, x_n\} \subset (a, b)$, let $\mathscr{S}_n^q$ denote the n-dimensional space of natural polynomial splines of degree $2q-1$ with simple knots at the prescribed points. Specifically, $s \in \mathscr{S}_n^q$ if and only if $s \in C^{2q-2}(a, b)$, the space of functions on (a, b) with $2q-2$ continuous derivatives, $s^{(2q-1)}$ is constant on (x_i, x_{i+1}), $i = 1, \cdots, n-1$, and $s^{(j)} \equiv 0$ on (a, x_1), (x_n, b), $j = q, \cdots, 2q-1$. Thus s is a polynomial of degree $q-1$ on $[a, x_1]$ and $[x_n, b]$ and of degree $2q-1$ on (x_i, x_{i+1}), $i = 1, \cdots, n-1$, with jumps in the $(2q-1)$st derivative only at the knots. Natural splines have a well-known minimum norm property (cf. Greville, 1968). (Assume $n \geq q$.) For any $f \in W^q$, there is a unique $s \in \mathscr{S}_n^q$ such that $s(x_i) = f(x_i)$, $i = 1, \cdots, n$, and $\|s^{(q)}\| \leq \|g^{(q)}\|$ for any $g \in W^q$ interpolating f at the knots. In particular, $\|s^{(q)}\| \leq \|f^{(q)}\|$.

The basis $\{\varphi_1, \cdots, \varphi_n\}$ for $\mathscr{S}_n^q$ of Demmler and Reinsch (1975) is determined (essentially uniquely) by the conditions

$$\sum_{\ell=1}^n \varphi_i(x_\ell)\varphi_j(x_\ell) = \delta_{ij}$$

$$\int_a^b \varphi_i^{(q)}(x)\varphi_j^{(q)}(x)\,dx = \delta_{ij}\lambda_j \tag{2.1}$$

$$0 = \lambda_1 = \cdots = \lambda_q < \lambda_{q+1} \leq \cdots \leq \lambda_n$$

for $i, j = 1, \cdots, n$. The eigenfunctions $\{\varphi_1, \cdots, \varphi_q\}$ corresponding to the zero eigenvalues span the space of polynomials of order q. Demmler and Reinsch showed that φ_k has exactly $k-1$ oscillations for $k > q$. Note that if $f = \sum \beta_k\varphi_k \in \mathscr{S}_n^q$, then $\|f^{(q)}\|^2 = \sum \beta_k^2\lambda_k$. (All summations range from 1 to n unless otherwise noted.)

To construct an estimator, let $\hat{\beta}_k = \sum Y_i\varphi_k(x_i)$. Since the basis defined by (2.1) is orthogonal, $\hat{\beta}_k$ is the least squares estimate of β_k for $f \in \mathscr{S}_n^q$ and

$$\sum_{k=1}^q \hat{\beta}_k\varphi_k(x)$$

is exactly the least squares polynomial of order q if $n \geq q$. The global minimax estimator $\hat{f}_\gamma$, derived in the next section, can be viewed as the natural spline

$$\hat{f}_\gamma(x) = \sum_{k=1}^n (1 - \sqrt{\gamma\lambda_k})_+ \hat{\beta}_k\varphi_k(x), \tag{2.2}$$

where $\gamma \geq 0$ is a smoothing parameter and the standard notation $(t)_+ = \max\{t, 0\}$ for t real is used. Thus $\hat{f}_\gamma(x)$ is the least squares polynomial of order q plus added terms depending on the value of γ.

REMARK 2.1. Reinsch's smoothing spline has the form

$$\hat{f}_{sm}(x; \lambda) = \sum_{k=1}^n (1/1 + \lambda\lambda_k)\hat{\beta}_k\varphi_k(x)$$

(see Demmler and Reinsch for details with $p = \lambda^{-1}$). Speckman (1980) and Li (1982) have shown that $\hat{f}_{sm}$ has a local minimax property.

The main results of this paper rely on a version of the asymptotic characterization of the eigenvectors and eigenvalues (2.1) first given by Utreras (1979, 1980).

Let $G \in W^2$ be a continuously differentiable function mapping $[0, 1]$ *onto* $[a, b]$ with $G'(x) \geq c > 0$ on $[0, 1]$ for some constant c. We consider the sequence of sets $\{a < x_{1n} < \cdots < x_{nn} < b\}$, $n = 1, 2, \cdots$, generated by $x_{in} = G((2i-1)/2n)$. Equivalently, let $p(x) = 1/G'(G^{-1}(x))$. Then p is a continuous density on $[a, b]$ and

$$(2i-1)/2n = \int_a^{x_{in}} p(x)\,dx \tag{2.3}$$

uniquely determines the set. (This is almost the regular sequence of Sacks and Ylvisaker, 1970.)

With this sequence,

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_{kn}(x_{in})\varphi_{jn}(x_{in}) \sim \int_a^b \varphi_{kn}(x)\varphi_{jn}(x)p(x)\,dx,$$

suggesting a comparison with the continuous version: find a set $\{\psi_k\}_{k=1}^{\infty}$ spanning W^q and eigenvalues $\{\nu_k\}_{k=1}^{\infty}$ such that

$$\int_a^b \psi_k(x)\psi_j(x)p(x)\,dx = \delta_{kj}$$

(2.4) $$k, j = 1, 2, \cdots$$

$$\int_a^b \psi_k^{(q)}(x)\psi_j^{(q)}(x)\,dx = \delta_{kj}\nu_k.$$

But this is equivalent to the eigensystem of the differential equation (cf. Utreras, 1979)

$$\psi^{(2q)}(x) = (-1)^q\nu\psi(x)p(x),$$
$$\psi^{(j)}(a) = \psi^{(j)}(b) = 0, \quad j = q, \cdots, 2q-1,$$

which can be solved or approximated in certain cases. Moreover, $\{\psi_k\}_{k=1}^{\infty}$ is a complete orthonormal system in $L^2(a, b)$ under the inner product

$$(f, q) = \int_a^b f(t)q(t)p(t)\,dt,$$

and

$$f = \sum_{k=1}^{\infty}\beta_k\psi_k \in L^2$$

belongs to W^q if and only if

$$\|f^{(q)}\|^2 = \sum_{k=1}^{\infty}\beta_k^2\nu_k < \infty.$$

In view of (2.1) and (2.4), it is reasonable to expect that $\varphi_{kn} \approx n^{-1/2}\psi_k$ and $\lambda_{kn} \approx \nu_k/n$. The next result makes the approximation precise.

THEOREM 2.2. *Assume* $\{x_{in}\}_{i=1}^{n}$ *satisfies* (2.3), *and let*

$$\beta_{kn} = \sum_{i=1}^{n} f(x_{in})\phi_{kn}(x_{in})$$

and

$$\beta_k = \int_a^b f(x)\psi_k(x)p(x)\,dx.$$

Then for fixed $k = 1, 2, \cdots$,

(2.5a) $$\lim_{n\to\infty} n\lambda_{kn} = \nu_k,$$

(2.5b) $$\lim_{n\to\infty} \beta_{kn}^2/n = \beta_k^2,$$

and

(2.5c) $$\lim_{n\to\infty}\beta_{kn}^2\lambda_{kn} = \beta_k^2\nu_k.$$

Moreover

(2.5d) $$\lambda_{q+k,n} = n^{-1}(kc_q(p))^{2q}(1 + o(1)).$$

Where

(2.5e) $$c_q(p) = \pi\left(\int_a^b p(x)^{1/2q}\,dx\right)^{-1}$$

In (2.5d), $o(1)$ *denotes a term tending to zero as* $n \to \infty$ *uniformly for* $k_{1n} \le k \le k_{2n}$ *for any sequences* $k_{1n} \to \infty$ *and* $k_{2n} = o(n^{2/(2q+1)})$.

PROOF. Utreras (1979, 1980) proved (2.5a) for k fixed. His method can be extended to show that $n^{1/2}\phi_{kn} \to \psi_k$ in q.m., yielding (2.5b) and (2.5c). (2.5d) is Corollary 5.4 of Speckman (1984).

3. The minimax estimator. The derivation of the minimax estimator depends on the following version of a result of Kuks and Olman (1971).

LEMMA 3.1. *Let* $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\}$ *be a nonnull diagonal matrix with* $0 = \lambda_1 = \cdots = \lambda_q < \lambda_{q+1} \le \cdots \le \lambda_n < \infty$. *The solution to the minimax problem*

$$\min_{\mathbf{B}}\max_{\beta'\Lambda\beta\le\alpha^2}\|(\mathbf{I} - \mathbf{B})\beta\|^2 + \sigma^2\operatorname{tr}(\mathbf{B}'\mathbf{B})$$

where the minimum is taken over $n \times n$ *matrices* $\mathbf{B}$ *is achieved when* $\mathbf{B} = \mathbf{B}(\gamma) = \operatorname{diag}\{(1 - \sqrt{\gamma\lambda_i})_+\}$ *for some constant* $\gamma > 0$.

PROOF. See the Appendix.

THEOREM 3.2. *With* $\hat{f}_\gamma$ *given by* (2.2),

$$\min_{\hat{f}\in\mathscr{L}}\max_{f\in\mathscr{F}_{q,\alpha}} ET(f, \hat{f}) = \min_\gamma \max_{f\in\mathscr{F}_{q,\alpha}} ET(f, \hat{f}_\gamma)$$
$$= \min_{0\le\gamma\le\lambda_{q+1}^{-1}}(1/n)\{\alpha^2\gamma + \sigma^2\textstyle\sum_{k=1}^n (1 - \sqrt{\gamma\lambda_k})_+^2\}.$$

PROOF. If $\hat{f}$ is any linear estimator, there is an $n \times n$ matrix $\mathbf{A}$ such that $\hat{\mathbf{f}}' = (\hat{f}(x_1), \cdots, \hat{f}(x_n))' = \mathbf{A}(y_1, \cdots, y_n)'$. Let $\mathbf{f} = (f(x_1), \cdots, f(x_n))'$ and write

$$ET(f, \hat{f}) = (1/n)\{\|(\mathbf{I} - \mathbf{A})\mathbf{f}\|^2 + \sigma^2\operatorname{tr}(\mathbf{A}'\mathbf{A})\}.$$

Since this clearly depends on f only through its values at the observation points, $ET(\hat{f}, f) = ET(\hat{f}, s)$ where $s \in \mathscr{S}_n^q$ is the (unique) interpolating natural spline satisfying $s(x_i) = f(x_i)$, $i = 1, \cdots$. Thus by the minimum norm property of s, $\|s^{(q)}\| \le \|f^{(q)}\|$ and

(3.1) $$\sup_{f\in\mathscr{F}_{q,\alpha}} ET(\hat{f}, f) = \sup_{s\in\mathscr{S}_n^q\cap\mathscr{F}_{q,\alpha}} ET(\hat{f}, s).$$

Now let $\Phi = [\phi_j(x_i)]_{i,j=1,\cdots,n}$ be the $n \times n$ orthogonal matrix determined by the basis (2.1), and let $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\}$ be the diagonal matrix of eigenvalues.

For $s \in \mathscr{S}_n^q$, there exists $\beta \in \mathbb{R}^n$ with $\mathbf{s} = (s(x_1), \cdots, s(x_n))' = \Phi\beta$ and $\| s^{(q)} \|^2 = \beta'\Lambda\beta$. Then for any linear estimator $\hat{\mathbf{f}} = \mathbf{A}\mathbf{y}$, (3.1) implies

$$\sup_{f \in \mathscr{S}_{q,\alpha}} nET(\hat{f}, f) = \max_{\beta'\Lambda\beta \le \alpha^2} (1/n)\{\| (\mathbf{I} - \mathbf{A})\Phi\beta \|^2 + \sigma^2 \operatorname{tr} \mathbf{A}'\mathbf{A}\}.$$

Now by lemma 3.1 with $\mathbf{B} = \Phi'\mathbf{A}\Phi$, the minimax value is achieved with $\mathbf{B} = \mathbf{D}(\gamma) = \operatorname{diag}\{(1 - \sqrt{\gamma\lambda_i})_+\}$ for some $\gamma > 0$. Thus

$$\mathbf{A}(\gamma) = \Phi\mathbf{D}(\gamma)\Phi' \tag{3.2}$$

is minimax for some γ. The expression for the minimax value follows since

$$\max_{\beta'\Lambda\beta \le \alpha^2} \| (\mathbf{I} - \mathbf{A}(\gamma))\Phi\beta \|^2 = \begin{cases} \alpha^2\gamma, & 0 \le \gamma \le \lambda_{q+1}^{-1} \\ \alpha^2\lambda_{q+1}^{-1}, & \gamma > \lambda_{q+1}^{-1} \end{cases}$$

and the proof is complete.

To obtain asymptotic results, we consider the sequence of observation sets specified by (2.3). For fixed n, let $d_{kn}(\gamma) = (1 - \sqrt{\gamma\lambda_{kn}})_+$, $k = 1, \cdots, n$ and let

$$\mu_{2n}(\gamma) = (q/n) + (1/n) \sum_{k=q+1}^{n} (1 - \sqrt{\gamma\lambda_{kn}})_+^2.$$

Also, define $e_n(\gamma) = \sup_{f \in \mathscr{S}_{q,\alpha}} ET(f, \hat{f}_\gamma)$. Then by Theorem 3.2, for $0 \le \gamma \le \lambda_{q+1,n}^{-1}$,

$$e_n(\gamma) = n^{-1}\gamma\alpha^2 + \sigma^2\mu_{2n}(\gamma).$$

The asymptotics are simplified somewhat by a suitable change of scale. Let $r = (2q + 1)^{-1}$ and $\delta = n^{-r}\gamma$. Throughout the ensuing discussion, if g is any function of γ, a function $\tilde{g}$ of δ will be defined by $\tilde{g}(\delta) = n^{1-r}g(n^r\delta) = n^{1-r}g(\gamma)$. In particular,

$$\tilde{e}_n(\delta) = \delta\alpha^2 + \sigma^2\tilde{\mu}_{2n}(\delta), \quad 0 \le \delta \le n^{-r}\lambda_{q+1,n}^{-1}. \tag{3.3}$$

The exact optimal asymptotic rate of convergence of e_n is established with the aid of a lemma. Recalling the definition of $c_q(p)$ in (2.5e), define

$$m_q = 2q^2c_q(p)/((2q + 1)(q + 1))$$

and

$$h(\delta) = \delta\alpha^2 + \sigma^2\delta^{-1/(2q)}m_q, \quad \delta > 0.$$

LEMMA 3.3. *As* $n \to \infty$,

$$\tilde{\mu}_{2n}(\delta) = \delta^{-1/(2q)}m_q + o(1) \tag{3.4a}$$

and

$$\tilde{e}_n(\delta) = h(\delta) + o(1), \tag{3.4b}$$

where $o(1)$ *denotes a term tending to* 0 *independent of* $\delta \in I$ *for any fixed interval* $I \subset (0, \infty)$.

PROOF. If the estimate of (2.5d) were exact, then $\mu_{2n}(\gamma)$ would equal

$$(q/n) + (1/n)\sum_{k=1}^{n-q}(1 - \sqrt{uk^{2q}})_+^2$$

with $u = c_q(p)^{2q}/n$. Assuming $0 < \delta^* \le n^{-r}\gamma \le \delta^{**} < \infty$, then there exist positive constants Δ^* and Δ^{**} such that

$$\Delta^* n^{r-1} < u < \Delta^{**} n^{r-1}. \tag{3.5}$$

But in this range, one may check that

$$\sum_{k=1}^{n-q}(1 - \sqrt{uk^{2q}})_+^2 = u^{1/(2q)}\int_0^2 (1 - y^q)^2\,dy(1 + o(1)) \tag{3.6}$$

with error term $o(1)$ tending to zero uniformly for u satisfying (3.5) as $n \to \infty$. Thus, under (3.5) the summation in (3.6) contains $O(n^r)$ terms and is bounded below by cn^r for some constant c. This implies that the first $n^{r/2}$ terms (say) can be disregarded, and by the uniformity of (2.5d), for $n^{r/2} < k \le cn^r$,

$$\mu_{2n}(\gamma) = (1/n)\{\textstyle\sum_{n^{r/2}<k\le n-q}(1 - \sqrt{uk^{2q}})_+^2\}(1 + o(1)).$$

Evaluating the right side of (3.6) yields

$$\mu_{2n}(\gamma) = n^{r-1}\gamma^{-r}m_q(1 + o(1)),$$

and a change of variables gives (3.4a). Finally, (3.4b) follows directly by definition.

THEOREM 3.4. *If* $\{x_{in}\}_{i=1}^n$ *satisfies* (2.3),

$$\min_{\hat f \in \mathscr{F}} \max_{f \in \mathscr{F}_{q,\alpha}} ET(f, \hat f) = n^{r-1}C_q(\alpha, \sigma)(1 + o(1)),$$

where $C_q(\alpha, \sigma) = \alpha^{2r}(\sigma^2 m_q/(2q))^{1-r}(2q + 1)$. *Moreover* $\gamma_n = n^r\delta_0$, *with* $\delta_0 = [\sigma^2 m_q/(2q\alpha^2)]^{1-r}$, *is asymptotically optimal in the sense that*

$$\max_{f \in \mathscr{F}_{q,\alpha}} ET(\hat f_{\gamma_n}, f) = n^{r-1}C_q(\alpha, \sigma)(1 + o(1)).$$

PROOF. Let γ_n be any minimizer of $e_n(\gamma)$ and let $\delta_n = n^{-r}\gamma_n$. Since $\gamma_n \in [0, \lambda_{q+1,n}^{-1}]$, $\delta_n = O(n^{1-r})$ by (2.5d). Let $\delta_n^* \equiv 1$ and consider $\tilde e_n(\delta_n^*) = h(1) + o(1)$. By optimality, $\tilde e_n(\delta_n) \le \tilde e_n(1) = \alpha^2 + m_q + o(1)$ from (3.3) and (3.4a). Since $\tilde e_n(\delta_n) \ge \delta_n\alpha^2$, this implies that lim sup $\delta_n < \infty$. Moreover, $\tilde e_n(\delta) \ge \sigma^2\tilde\mu_{2n}(\delta)$, and by (3.4a) and the fact that $\tilde\mu_{2n}(\delta)$ is a nonincreasing function of δ, lim inf $\delta_n > 0$. Thus there is a compact interval $I = [\delta^*, \delta^{**}] \subset (0, \infty)$ such that $\delta_n \in I$ for all n sufficiently large. One can check that $h(\delta)$ is strictly convex and has a unique minimum at δ_0 with $h(\delta_0) = C_q(\alpha, \sigma^2)$. It is then routine to show that $\delta_n \to \delta_0$ and $\tilde e_n(\delta_n) \to h(\delta_0)$ using the fact that $\tilde e_n(\delta) \to h(\delta)$ uniformly on I.

4. The adaptive estimator. This section develops the asymptotic properties of a practical method which achieves the rate of convergence of Theorem 3.4 for arbitrary $f \in W^q$, $\|f^{(q)}\| = \alpha > 0$, even if α and σ^2 are unknown. We begin with a lemma giving the asymptotic rate of convergence for fixed f and showing incidentally that $\hat f_\gamma$ is an asymptotic equalizer rule. From now on there is no ambiguity in the estimator used and f is assumed fixed, so for simplicity let

$T_n(\gamma) = T(\hat{f}_\gamma, f)$. In the notation of Section 2,

$$T_n(\gamma) = n^{-1} \sum_{i=1}^n (\hat{f}_\gamma(x_{in}) - f(x_{in}))^2$$
$$= n^{-1} \sum_{k=1}^n (\beta_{kn} - d_{kn}(\gamma)\hat{\beta}_{kn})^2.$$

LEMMA 4.1. *Assume* $\|f^{(q)}\| = \alpha > 0$ *and let* γ_n^* *be a minimizer of* $ET_n(\gamma)$, $n = 1, 2, \cdots$. *Then*

$$\gamma_n^* = n^r \delta_0 (1 + o(1)) \tag{4.1}$$

and

$$ET_n(\gamma_n^*) = n^{r-1} C_q(\|f^{(q)}\|, \sigma^2)(1 + o(1)),$$

where δ_0 *and* C_q *are given in Theorem* 3.4.

PROOF. The first step of the proof is to show that $\delta_n^* = n^{-r}\gamma_n^*$ is bounded. By Theorem 3.4 $\limsup n^{1-r} ET_n(\gamma_n^*) < \infty$. Since $ET_n(\gamma) = n^{-1} \sum (1 - d_{kn}(\gamma))^2 \beta_{kn}^2 + \sigma^2 \mu_{2n}(\gamma)$, (3.4a) implies that $\liminf n^{-r}\gamma_n^* > 0$. By assumption, $0 < \|f^{(q)}\|^2 = \sum \beta_k^2 \nu_k < \infty$, so there is at least one index j for which $\beta_j^2 \nu_j > 0$. But $n^{-1}\beta_{jn}^2 \to \beta_j^2$, so we must have $d_{jn}(\gamma) \to 1$ in order for $ET_n(\gamma_n^*) \to 0$. In particular, $ET_n(\gamma_n^*) > n^{-1}\gamma_n^* \lambda_{jn} \beta_{jn}^2$ for n sufficiently large. Now let $\tilde{T}_n(\delta) = n^{1-r} T_n(n^r \delta)$ and $\delta_n^* = n^{-r}\gamma_n^*$. Since $\lambda_{jn}\beta_{jn}^2 \to \nu_j \beta_j^2 > 0$, $\limsup E\tilde{T}_n(\delta_n^*) \geq \mu_j \beta_j^2 \limsup \delta_n^*$, implying $\limsup_{n\to\infty} \delta_n^* < \infty$. Thus there is compact interval $I \subset (0, \infty)$ such that $\delta_n^* \in I$ for all n sufficiently large.

To finish the proof, it suffices to show that $E\tilde{T}_n(\delta) \to h(\delta)$ uniformly for $\delta \in I$ as in the proof of Theorem 3.4. Only a lower bound is necessary since $E\tilde{T}_n(\delta) \leq \tilde{e}_n(\delta)$, so let $\varepsilon > 0$ be arbitrary. There is an integer M such that

$$\sum_{k=1}^M \nu_k \beta_k^2 > \|f^{(q)}\|^2 (1 - \varepsilon);$$

hence applying (2.5c) and (3.4a), we have

$$\liminf_{n\to\infty} E\tilde{T}_n(\delta) \geq \liminf_{n\to\infty} \{\delta \sum_{k=1}^M \lambda_{kn}\beta_{kn}^2 + \delta^{-1/2q} m_q \sigma^2\}$$
$$\geq \delta \|f^{(q)}\|^2 (1 - \varepsilon) + \delta^{-1/2q} m_q \sigma^2$$

with uniform convergence on I, and the proof is complete.

To estimate the optimal parameter γ_n^*, we use generalized cross-validation, or GCV, as introduced by Craven and Wahba (1979) for ordinary smoothing splines. The GCV function is

$$V_n(\gamma) = n^{-1} \|\mathbf{y} - \hat{\mathbf{f}}_\gamma\|^2 / [n^{-1} \mathrm{tr}(\mathbf{I} - \mathbf{A}(\gamma))]^2,$$

where $\mathbf{A}(\gamma)$ is given by (3.2) and $\hat{\mathbf{f}}_\gamma = \mathbf{A}(\gamma)\mathbf{y}$. Because the spline basis is orthogonal,

$$V_n(\gamma) = n^{-1} \sum_{k=1}^n (1 - d_{kn}(\gamma))^2 \hat{\beta}_{kn}^2 / (1 - \mu_{1n}(\gamma))^2 \tag{4.2}$$

with

$$\mu_{1n}(\gamma) = n^{-1} \sum_{k=1}^n d_{kn}(\gamma).$$

The minimizer of V_n, a function of the data alone, is taken as an estimate of γ_n^*.

The derivation in Craven and Wahba (1979) shows that GCV is ordinary cross-validation in transformed coordinates. In addition, $V_n(\gamma)$ is an estimator of $ET_n(\gamma)$ with near constant bias, hence intuitively the two functions should be minimized at approximately the same point. A weak version of the desired result is given in the GCV theorem of Golub, Heath and Wahba (1979), namely

$$(EV_n(\gamma) - \sigma^2 - ET_n(\gamma)/ET_n(\gamma) \leq [2\mu_{1n}(\gamma) + (\mu_{1n}^2(\gamma)/\mu_{2n}(\gamma))]/(1 - \mu_{2n}(\gamma))^2.$$

As in Lemma 3.3, it is not hard to show that

$$\tilde{\mu}_{1n}(\delta) = O(1); \tag{4.3}$$

hence

$$(EV_n(\gamma) - \sigma^2 - ET_n(\gamma))/ET_n(\gamma) = o(1) \tag{4.4}$$

uniformly for $\gamma \in n^r I$, where $I \subset (0, \infty)$ is an arbitrary closed interval. Moreover (cf. Golub, Heath and Wahba, 1979, Corollary 1), the minimizer of $EV_n(\gamma)$, say γ_n^0, satisfies $ET_n(\gamma_n^0)/ET_n(\gamma_n^*) \to 1$. This result is suggestive but incomplete since $EV_n(\gamma)$ is also unobservable. The purpose of this section is to extend the arguments in Craven and Wahba (1979) to obtain a comparable result for the sample version. This is accomplished under

ASSUMPTION 4.2. (a) $Y_1, \cdots, Y_n$ are independent, normally distributed with mean $EY_i = f(x_{in})$ and variance σ^2.

$$\text{(b)} \quad \|f^{(q)}\| = \alpha > 0.$$

The results below establish weak consistency for a sequence $\hat{\gamma}_n$ of possibly local minimizers of $V_n(\gamma)$. Specifically, let $I = [\delta^*, \delta^{**}] \subset (0, \infty)$ be arbitrary with $\delta_0 \in I$, and define $\hat{\gamma}_n$ to be the minimizer of $V_n(\gamma)$ over I.

REMARK. Note that because $V_n(\gamma)$ is continuous in γ, $\hat{\gamma}_n$ is measurable (in "ω" of the underlying probability space) as in the standard proof of measurability of maximum likelihood estimators.

THEOREM 4.3. *Under Assumption* 4.2, *there exists a sequence of* (*possibly local*) *minimizers* $\{\hat{\gamma}_n\}$ *of* $V_n(\gamma)$ *such that* $\hat{\gamma}_n/\gamma_n^* \to 1$ *in probability.*

REMARK. As a referee has pointed out, it remains unclear whether or not the unconstrained application of cross-validation also leads to "efficient" (or even consistent) results. This is important because the point of cross-validation is to provide a completely data-driven choice of the smoothing parameter. In its present form, the theorem requires an a priori choice of δ^* and δ^{**}, which are themselves smoothing parameters. The efficiency of the unconstrained version remains an open problem.

The proof is broken down into several steps, beginning with an extension of Kolmogorov's inequality.

LEMMA 4.4. *Let $Z_1, \cdots, Z_n$ be independent mean zero random variables with* $\operatorname{Var}(Z_1 + \cdots + Z_n) = \sigma_n^2 < \infty$, *and let*

$$\mathscr{C} = \{(c_1, \cdots, c_n) \in \mathbb{R}^n: 1 \geq c_1 \geq c_2 \geq \cdots \geq c_n \geq 0\}.$$

Then

$$P[\sup_{\mathscr{C}} |\textstyle\sum_{i=1}^n c_i Z_i| > \varepsilon] \leq \sigma_n^2/\varepsilon^2$$

for all $\varepsilon > o$.

PROOF. The set of extreme points of $\mathscr{C}$ is exactly the set $\{e_0, \cdots, e_n\}$ where $e_k = (e_{k1}, \cdots, e_{kn})$ is defined by $e_{ki} = 1$ if $i \leq k$, 0 if $i > k$. Since $|\sum c_i Z_i|$ is a convex function on $\mathscr{C}$ for arbitrary $\{Z_1, \cdots, Z_n\}$,

$$\sup_{\mathscr{C}} |\textstyle\sum_{i=1}^n c_i Z_i| = \max_{0 \leq k \leq n} |\sum_{i=1}^k e_{ki} Z_i| = \max_{1 \leq k \leq n} |\sum_{i=1}^k Z_i|.$$

Therefore

$$P[\sup_{\mathscr{C}} |\textstyle\sum_{i=1}^n c_i Z_i| > \varepsilon] = P[\max_k |\sum_{i=1}^k Z_i| > \varepsilon] \leq \sigma_n^2/\varepsilon^2.$$

by Kolmogorov's inequality (cf. Chung, 1974; Theorem 5.3.1), and the lemma is proved.

We return now to the discussion of GCV. Let

$$K = K(n, \delta^*) = \min\{k: n^r \delta^* \lambda_{kn} \geq 1\},$$

so $d_{kn}(\gamma) = 0$ for all $\gamma = n^r \delta \in n^r I$ and $k > K$. Note that

$$\lambda_{Kn} = O(n^{-r}) \tag{4.5a}$$

and

$$K = O(n^r) \tag{4.5b}$$

from (2.5 d). Also define $X_k = \hat{\beta}_{kn}^2 - \beta_{kn}^2 - \sigma^2$ and

$$S_n = n^{-1} \textstyle\sum_{k=K+1}^n X_k,$$

and let

$$W_n(\gamma) = V_n(\gamma) - \sigma^2 - S_n.$$

W_n and V_n differ only by a random quantity independent of γ; hence they have the same minimizer. In addition,

$$EW_n(\gamma) = EV_n(\gamma) - \sigma^2 \tag{4.6}$$

and the mean functions have the same minimizer, γ_n^0, as well. The point of introducing W_n is that it seems more tractable because it depends essentially on only $K = O(n^r)$ terms. The key step provided by the next lemma is to show that

the error $W_n - EW_n$ is sufficiently small. To that end, normalize again by letting $\tilde{W}_n(\delta) = n^{1-r}W_n(n^r\delta)$.

LEMMA 4.5. $\max_{\delta\in I}|\tilde{W}_n(\delta) - E\tilde{W}_n(\delta)| \to 0$ *in probability.*

PROOF. For ease of notation, we suppress the double subscripts and expand (4.2) letting $d_k = d_{kn}(n^r\delta)$ and

$$U_n(\delta) = n^{-r}\sum_{k=1}^{K}(1 - d_k)^2X_n$$

to obtain

$$\begin{aligned}\tilde{W}_n(\delta) - E\tilde{W}_n(\delta) &= n^{-r}(1-\mu_1)^{-2}\sum_{k=1}^{n}(1-d_k)^2X_k - n^{1-r}S_n \\ &= (1-\mu_1)^{-2}[U_n(\delta) - n^{1-r}\mu_1(2-\mu_1)S_n].\end{aligned} \tag{4.7}$$

After some manipulation, one can write

$$U_n(\delta) = \delta\sum_{k=1}^{K}c_k(\delta)\lambda_{kn}X_k,$$

where $c_k(\delta) = \min\{1, (n^r\delta\lambda_{kn})^{-1}\}$. Note that $1 \geq c_1(\delta) \geq \cdots \geq c_n(\delta) \geq 0$ for all $\delta \in I$ and by normality $Z_k = \lambda_{kn}X_k$ satisfies the assumptions of Lemma 4.4. Using normality again,

$$\begin{aligned}\operatorname{Var}\Big(\sum_{k=1}^{K}\lambda_{kn}X_k\Big) &= \sum_{k=1}^{K}\lambda_{kn}^2(2\sigma^4 + 4\sigma^2\beta_{kn}^2) \\ &= 2\sigma^4\sum_{k=1}^{K}\lambda_{kn}^2 + 4\sigma^2\sum_{k=1}^{K}\lambda_{kn}^2\beta_{kn}^2.\end{aligned}$$

By (2.5d) and (4.5b), the first term is $O(n^{-r})$. The second term is bounded by

$$\begin{aligned}4\sigma^2\lambda_{Kn}\sum_{k=1}^{K}\lambda_{kn}\beta_{kn}^2 &\leq 4\sigma^2\lambda_{Kn}\sum_{k=1}^{n}\lambda_{kn}\beta_{kn}^2 \\ &\leq 4\sigma^2\lambda_{Kn}\|f^{(q)}\|^2,\end{aligned}$$

which is also $O(n^{-r})$ from (4.5a). Thus by Lemma 4.4, $\max_{\delta\in I}U_n(\delta) \to 0$ in probability.

The second term in (4.7) is easier. From (4.3), $n^{1-r}\mu_1$ is bounded and $\mu_1 \to 0$ for $\delta \in I$, so it suffices to show that $S_n \to 0$. But $ES_n = 0$ and, by normality (2.5d), and (4.5a),

$$\begin{aligned}\operatorname{Var}(S_n) &= n^{-2}\sum_{k=K+1}^{n}(2\sigma^4 + 4\sigma^2\beta_{kn}^2) \\ &\leq 2\sigma^4/n + (4\sigma^2/(n^2\lambda_{Kn}))\sum_{k=K+1}^{n}\lambda_{kn}\beta_{kn}^2 \\ &\leq 2\sigma^4/n + 4\sigma^2\|f^{(q)}\|^2/(n^2\lambda_{Kn}) = O(n^{-1}).\end{aligned}$$

Thus the term involving S_n is also negligible. Since $(1-\mu_1)^{-2} \to 1$ uniformly on I, the proof is complete.

PROOF OF THEOREM 4.3. From the proof of Lemma 4.1, $E\tilde{T}_n(\delta) \to h(\delta)$ uniformly on I. Thus using (4.4), (4.6), and Lemma 4.5, $\max_{\delta\in I}|\tilde{W}_n(\delta) - h(\delta)| \to_P 0$. Once again, since h is convex with unique minimum at δ_0, we have $\hat{\delta}_n \to_P \delta_0$, where $\hat{\delta}_n$, the minimizer of $\tilde{V}_n$, is also the minimizer on I of $\tilde{W}_n$. But $\hat{\gamma}_n = n^r\hat{\delta}_n$, so (4.1) gives the desired result.

THEOREM 4.6. $\lim_{n\to\infty} ET_n(\hat{\gamma}_n)/ET_n(\gamma_n^*) = 1$.

We need a lemma whose proof is similar to that of Lemma 4.5 and is omitted.

LEMMA 4.7. $\max_{\delta\in I} | \tilde{T}_n(\delta) - h(\delta) | \to_P 0$.

PROOF OF THEOREM 4.6. It is equivalent to show that $E\tilde{T}_n(\hat{\delta}_n) - E\tilde{T}_n(\delta_n^*) \to 0$. Since $\tilde{T}_n(\hat{\delta}_n) - E\tilde{T}_n(\delta_n^*) \to_P 0$ by Lemma 4.7 and Theorem 4.3, it suffices to show that $\tilde{T}_n(\hat{\delta}_n)$ is a uniformly integrable sequence (cf. Chung 1974, Theorem 4.5.4). Let $\hat{d}_k = d_{kn}(n^r\hat{\delta}_n)$. Since $\hat{\delta}_n \le \delta^{**}$, we have $(1 - \hat{d}_k)^2 \le n^r\delta^{**}\lambda_{kn}$ and, using a standard inequality,

$$\begin{aligned}\tilde{T}_n(\hat{\delta}_n) &= n^{-r}\textstyle\sum_{k=1}^n (\hat{d}_k\hat{\beta}_{kn} - \beta_{kn})^2 \\ &\le 2n^{-r}[\textstyle\sum_{k=1}^n \hat{d}_k^2(\hat{\beta}_{kn} - \beta_{kn})^2 + \sum_{k=1}^n \beta_{kn}^2(1 - \hat{d}_k)^2] \\ &\le 2n^{-r}\textstyle\sum_{k=1}^K (\hat{\beta}_{kn} - \beta_{kn})^2 + 2\delta^{**}\| f^{(q)} \|^2.\end{aligned}$$

The last term involving the sum is uniformly integrable because $K = O(n^r)$ and $\{\hat{\beta}_{kn} - \beta_{kn}\}_{k=1}^n$ by assumption is a sequence of independent standard normal random variables. This completes the proof.

REMARK. The normality assumption seems to be used most critically in applying the version of Kolmogorov's inequality, where the $\hat{\beta}_{kn}$'s must be independent. It seems likely that the results are true in greater generality.

5. Remarks. In order to compute $\hat{f}_\gamma$, at least the first few eigenvectors and eigenvalues in (2.1) must be generated. Demmler and Reinsch (1975) outlined a relatively economical computational scheme which is especially suitable if only the first few terms are needed.

For $q = 2$, Craven and Wahba (1979) gave a method for generating the entire sequence $\{\lambda_3, \cdots, \lambda_n\}$ and $\{\varphi_3, \cdots, \varphi_n\}$ based on a singular value decomposition. In their notation, $d_k^2 = \lambda_{k+2}$, $k = 1, \cdots, n-2$, and $U = (\varphi_3, \cdots, \varphi_n)$ is a suborthogonal matrix. The first two terms in (2.2), the projection onto the subspace of linear polynomials, are computed by ordinary least squares. As Craven and Wahba point out, the expensive task of producing the entire basis need only be performed once for each set of x_i's; so for many problems, the procedure is relatively inexpensive. However, for n larger than 150 or so, this approach is probably not feasible.

APPENDIX

We give a new proof of the theorem of Kuks and Olman (1971).

PROOF OF LEMMA 3.1. Let $\mathbf{B} = [b_{ij}]$ be an $n \times n$ matrix and let

$$J(\mathbf{B}) = \max_{\beta'\Lambda\beta\le 1} \| (\mathbf{I} - \mathbf{B})\beta \|^2 + (\sigma^2/\alpha^2)\mathrm{tr}(\mathbf{B}'\mathbf{B}).$$

By a simple transformation the quantity to be minimized is $\alpha^2 J(\mathbf{B})$. Also define

$$J_0(\mathbf{B}) = \max_{q<i\le n}(1 - b_{ii})^2/\lambda_i + (\sigma^2/\alpha^2)\sum_{i=1}^n b_{ii}^2. \tag{A.1}$$

Note that if $\mathbf{B}$ is a diagonal matrix such that

$$b_{ii} = 1, \quad i = 1, \cdots, q, \tag{A.2}$$

then $J(\mathbf{B}) = J_0(\mathbf{B})$. Moreover, if $J(\mathbf{B}) < \infty$ for any matrix $\mathbf{B}$, then (A.2) must hold. Let $\mathscr{B}$ denote the class of $n \times n$ matrices satisfying (A.2), and let $\mathbf{e}_i$ denote the ith coordinate unit vector in $\mathbb{R}^k$. If $\mathbf{B} \in \mathscr{B}$,

$$\begin{aligned} J(\mathbf{B}) &\ge \max_{q<i\le n} \|(\mathbf{I} - \mathbf{B})\mathbf{e}_i\|^2/\lambda_i + (\sigma^2/\alpha^2)\sum_{i=1}^n b_{ii}^2 \\ &= \max_{q<i\le n}[(1 - b_{ii})^2 + \sum_{j\ne i} b_{ij}^2]/\lambda_i + (\sigma^2/\alpha^2)\sum_{i=1}^n b_{ii}^2 \\ &\ge J_0(\mathbf{B}) \end{aligned}$$

with equality if and only if $\mathbf{B}$ is diagonal. Let $\mathbf{D} \in \mathscr{B}$ be the diagonal matrix which minimizes J_0. Then for any $\mathbf{B}$,

$$J(\mathbf{D}) = J_0(\mathbf{D}) \le J_0(\mathbf{B}) \le J(\mathbf{B}),$$

hence $\mathbf{D}$ is minimax. Write $\mathbf{D} = \operatorname{diag}(d_1, \cdots, d_n)$, and let

$$\gamma = \max_{q<i\le n}(1 - d_i)^2/\lambda_i. \tag{A.3}$$

From (A.2), it is clear that $0 \le d_i \le 1$ for $1 \le i \le n$, for if either $d_i < 0$ or $d_i > 1$, $J_0(D)$ could be made smaller by replacing d_i by 0 or 1, respectively. Thus if $q < i \le n$ and equality holds in (A.3), then $d_i = (1 - \sqrt{\gamma\lambda_i})$. Similarly, if $\gamma > (1 - d_i)^2/\lambda_i$, it follows that $d_i = 0$ since otherwise $J_0(D)$ could be decreased by making d_i smaller. From these conditions, $d_i = (1 - \gamma\sqrt{\lambda_i})_+$ and by (A.3), $\gamma \le \lambda_{q+1}^{-1}$.

REMARK. The general minimax problem defined by

$$\min_{\mathbf{B}}\max_{\beta'\Lambda\beta\le 1}\beta'(I - \mathbf{B})'W(I - \mathbf{B})\beta + \operatorname{tr}(\mathbf{B}'\mathbf{B}),$$

where W is a symmetric nonnegative definite matrix, is much more difficult. The only cases that have been solved explicitly are when W is diagonal or has rank 1. (cf. Kuks and Olman (1972), Speckman (1980) for the latter.) Laüter (1975) treated the general case, but there is no explicit solution.

Acknowledgement. The author is indebted to an anonymous referee for an extremely thorough reading of the original and for many helpful comments.

REFERENCES

AGARWAL, G. and STUDDEN, W. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8** 1307–1325.

CHUNG, K. L. (1974). *A Course in Probability Theory,* 2nd ed. Academic, New York.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

DEMMLER, A. and REINSCH, C. (1975). Oscillation matrices with spline smoothing. *Numer. Math.* **24** 375–382.

GOLUB, G., HEATH, M. and WAHBA, G. (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

GREVILLE, T. N. E. (1968). Introduction to spline functions. In *Theory and Application of Spline Functions* (T. N. E. Greville, ed.). Academic, New York.

KUKS, J. A. and OLMAN, V. (1971). A minimax linear estimator of regression coefficients II (in Russian). *Izv. Akad. Nauk Eston. SSR* **20** 480–482.

KUKS, J. A. and OLMAN, V. (1972). A minimax linear estimator of regression coefficients (in Russian). *Izv. Akad. Nauk Eston. SSR* **21** (66–72).

LAÜTER, H. (1975). A minimax linear estimator for linear parameters under restrictions in form of inequalities. *Math. Operationsforsch. Statist.* **6** 689–695.

LI, K.-C. (1982). Minimaxity of the method of regularization on stochastic processes. *Ann. Statist.* **10** 937–942.

REINSCH, C. (1967) Smoothing by spline functions. *Numer. Math.* **10** 177–183.

SACKS, J. and YLVISAKER, D. (1970). Design for regression problems with correlated errors III. *Ann. Math. Statist.* **41** 2057–2074.

SACKS, J. and YLVISAKER, D. (1978). Linear estimation for approximately linear models. *Ann. Statist.* **6** 1122–1137.

SPECKMAN, P. (1980). Minimax estimates of linear functionals in a Hilbert space. Unpublished manuscript.

SPECKMAN, P. (1984). The asymptotic integrated mean square error for smoothing noisy data by splines. To appear in *Numer. Math.*

STONE, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

UTRERAS, F. (1979). Cross validation techniques for smoothing in one or two dimensions. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.). *Lecture Notes in Math.* **757**. Springer, New York.

UTRERAS, F. (1980). Sur le choix des parametre d'adjustement dans le lissage par functions spline. *Numer. Math.* **34** 15–28.

WAHBA, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.* **24** 333–343.

WAHBA, G. and WOLD, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation. *Comm. Statist.* **4** 1–17.

DEPARTMENT OF STATISTICS
UNIVERSITY OF MISSOURI
COLUMBIA, MISSOURI 65211