The Annals of Statistics
1991, Vol. 19, No, 4, 2003–2014

MINIMAX BAYES ESTIMATION IN NONPARAMETRIC REGRESSION

BY NANCY E. HECKMAN[1] AND MICHAEL WOODROOFE[2]

University of British Columbia and University of Michigan

One observes n data points, $(\mathbf{t}_i, Y_i)$, with the mean of Y_i, conditional on the regression function f, equal to $f(\mathbf{t}_i)$. The prior distribution of the vector $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^t$ is unknown, but lies in a known class Ω. An estimator, $\hat{\mathbf{f}}$, of $\mathbf{f}$ is found which minimizes the maximum $E\|\hat{\mathbf{f}} - \mathbf{f}\|^2$. The maximum is taken over all priors in Ω and the minimum is taken over linear estimators of $\mathbf{f}$. Asymptotic properties of the estimator are studied in the case that $\mathbf{t}_i$ is one-dimensional and Ω is the set of priors for which f is smooth.

1. Introduction. Suppose that one observes n data points $(\mathbf{t}_i, Y_i)$ in order to estimate $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^t$, with f the regression function $E(Y|f)$. In standard estimation procedures, one assumes a specific form of the function f, typically depending upon a few (much less than n) parameters. A Bayesian analysis involves the choice of a loss function and a prior distribution on the parameters. However, the assumption of a particular form of f may be arbitrary, and thus the specification of a particular prior will be suspect.

The approach to the estimation problem developed here is to treat the components of the vector $\mathbf{f}$ as the unknown parameters and to choose an estimator that performs well over a large class of priors on $\mathbf{f}$. Specifically, one assumes that

$$Y_i = f(\mathbf{t}_i) + \varepsilon_i \tag{1}$$

with $\mathbf{t}_i$ nonrandom and in $\mathscr{R}^m$, the ε_i's uncorrelated, mean zero, variance σ^2 and the $f(\mathbf{t}_i)$'s and ε_j's uncorrelated. The estimator $\hat{\mathbf{f}} = C\mathbf{Y}$ derived below is a minimax linear estimate; that is, $\hat{\mathbf{f}}$ minimizes (over linear estimators) the maximum of $E\|\hat{\mathbf{f}} - \mathbf{f}\|^2$. The expectation is taken with respect to both the conditional distribution of $\mathbf{Y}$ given $\mathbf{f}$ and the prior distribution of $\mathbf{f}$. The maximum is taken over Ω, a class of priors that suitably reflects one's beliefs about $\mathbf{f}$. Since the criterion is mean squared error, one may define Ω by placing restrictions on the mean and covariance of $\mathbf{f}$.

For Ω satisfying certain conditions, we find an explicit form for $\hat{\mathbf{f}}$ and study its frequentist and Bayesian asymptotic properties. Computation of $\hat{\mathbf{f}}$ involves

Received July 1989; revised July 1990.
[1]Supported under the Natural Sciences and Engineering Research Council of Canada Grant 5-87969.
[2]Supported under NSF Grant DMS-89-02188.
AMS 1980 *subject classification*. Primary 65D10.
Key words and phrases. Minimax estimates, Bayes estimates, nonparametric regression, smoothing.

calculating eigenvalues and eigenvectors of a matrix, and is thus easily accomplished with existing computer programs.

Although the conditions that we place on Ω are general, they are motivated by a particular case: $\mathbf{t}_i = t_i$ in $\mathscr{R}$ and priors which force the function to be smooth. Typically, the smoothness of a function is characterised by its kth derivative. For instance, in spline fitting, the smoothness of f is quantified by the integral of the square of its kth derivative; see, for instance, Silverman (1985). Here, the smoothness of f is characterised by divided differences. Let $\Delta^{(k)}(\mathbf{f})$ be the $n-k$ vector of kth divided differences of f, based upon the $f(t_i)$'s. Assuming that $\Delta^{(k)}(\mathbf{f})$ has mean zero and small covariance matrix, reflects a belief that the kth derivative of f is small. Thus, a reasonable Ω consists of all n-variate distributions under which the mean of $\Delta^{(k)}(\mathbf{f})$ is zero and the expected value of $\|\Delta^{(k)}(\mathbf{f})\|^2$ is bounded by ρ. The user chooses k and ρ. These numbers have their analogues in the usual methods of smoothing regression: k is analogous to the order of kernel or spline and $1/\rho$ is analogous to the smoothing parameter (in spline smoothing) or the bandwidth (in kernel smoothing); see Eubank (1988) for a discussion of these smoothing regression techniques.

For a particular choice of Ω, the minimax Bayes estimate of $\mathbf{f}$ is the same as the frequentist minimax estimate given by Speckman (1985) and Nussbaum (1985). This fact is discussed in Section 4.

The assumptions and general form of the minimax estimator are given in Section 2. In Section 3, these results are applied to Ω's which are based on kth divided differences. The Bayesian and frequentist asymptotic minimax mean squared errors of these estimators are given, along with the asymptotic normality of linear functions of $\hat{\mathbf{f}}$.

2. Minimax estimates for a general class of priors. Assume that (1) holds. Let A be an $L \times n$ matrix of rank $n-k$, some k, $0 < k < n$. Define $\Omega \equiv \Omega(A, \rho)$ to be the class of all priors on $\mathbf{f}$ such that the covariance of $\mathbf{f}$ is defined, $E(A\mathbf{f}) = \mathbf{0}$ and $E\|A\mathbf{f}\|^2 \le \rho$. Thus, other than the finite second moment assumption, the only restrictions placed upon the prior of $\mathbf{f}$ involve the mean and covariance of $A\mathbf{f}$.

Let the principal value decomposition of A be PDQ^t, where P is $L \times (n-k)$, Q is $n \times (n-k)$, D is diagonal with $0 \ne D_{11}^2 \le D_{22}^2 \le \cdots \le D_{n-k,n-k}^2$ and $P^tP = I_{n-k} = Q^tQ$. The columns of Q span the row space of A and are equal to the eigenvectors of A^tA which correspond to the nonzero eigenvalues D_{jj}^2, $j = 1, \ldots, n-k$.

THEOREM 1.

$$\min_{C} \max_{\text{priors in } \Omega} E\|C\mathbf{Y} - \mathbf{f}\|^2$$

$$= \sigma^2 k + \sigma^2 \sum s_{ii}/(s_{ii} + \sigma^2) = \sigma^2 k + \sigma^2 \sum (1 - |D_{ii}|\sigma^2/\alpha)_+,$$

where

$$s_{ii} = s_{ii}(\rho) = \left(\alpha/|D_{ii}| - \sigma^2\right)_+,$$

x_+ *is the maximum of* 0 *and* x, *and* α *is chosen so that* $\sum D_{ii}^2 s_{ii} = \rho$. *Let* P_1 *denote the* $n \times n$ *projection matrix onto the* k*-dimensional null space of* A. *Then the minimizing* C *is given by*

$$\hat{C} = P_1 + Q\hat{C}^*Q^t,$$

where $\hat{C}^*$ *is diagonal with*

$$\hat{C}_{ii}^* = \hat{C}_{ii}^*(\rho) = \left(1 - |\dot{D}_{ii}|\sigma^2/\alpha\right)_+.$$

Let $\mathbf{f}^* = Q^t\mathbf{f}$ *and* $\mathbf{Y}^* = Q^t\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}^*$. *Then* $\hat{C}^*\mathbf{Y}^*$ *is the Bayes estimate of* $\mathbf{f}^*$ *for the problem in which* $\mathbf{f}^*$ *and* $\boldsymbol{\varepsilon}^*$ *are independent normal random vectors,* $E(\mathbf{f}^*) = \mathbf{0}$ *and the covariance of* $\mathbf{f}^*$ *is* $\hat{\Sigma}^* = \operatorname{diag}(s_{11}, \ldots, s_{n-k,n-k})$.

Thus, the minimax Bayes estimate of $\mathbf{f}$ *may be thought of as the Bayes estimate with the above prior on* $\mathbf{f}^*$ *and a diffuse prior on* $P_1\mathbf{f}$.

PROOF. Let $P_2 = QQ^t$, the projection matrix onto the row space of A, so that $P_1 + P_2 = I_n$. We first show that the minimizing C must satisfy $CP_1 = P_1$ and $P_1CP_2 = 0$.

$$E\|C\mathbf{Y} - \mathbf{f}\|^2 \geq \|(C - I_n)E(\mathbf{f})\|^2 = \|(C - I_n)P_1E(\mathbf{f})\|^2$$

which is unbounded unless $(C - I_n)P_1 = 0$. So $CP_1 = P_1$ for the minimizing C. For such C, $C = P_1 + P_1CP_2 + P_2CP_2$ and therefore

$$\begin{aligned} E\|C\mathbf{Y} - \mathbf{f}\|^2 &= E\|P_1\mathbf{Y} - P_1\mathbf{f} + P_1CP_2\mathbf{Y} + P_2CP_2\mathbf{Y} - P_2\mathbf{f}\|^2 \\ &= \sigma^2 k + E\|P_1CP_2\mathbf{Y}\|^2 + E\|P_2CP_2\mathbf{Y} - P_2\mathbf{f}\|^2, \end{aligned}$$

since $P_1\mathbf{Y}$ and $P_2\mathbf{Y}$ are uncorrelated, given $\mathbf{f}$. It follows that the minimizing C must satisfy $P_1CP_2 = 0$, for otherwise C may be replaced by P_2C.

Let $C^* = Q^tCQ$, $\mathbf{f}^* = Q^t\mathbf{f}$ and $\mathbf{Y}^* = Q^t\mathbf{Y}$. Then

$$\min_C \max_{\text{priors in } \Omega} E\|C\mathbf{Y} - \mathbf{f}\|^2 = \sigma^2 k + \min_{C^*} \max_{\text{priors in } \Omega^*} E\|C^*\mathbf{Y}^* - \mathbf{f}^*\|^2,$$

where Ω^* consists of all priors on $\mathbf{f}^*$ with $E(\mathbf{f}^*) = \mathbf{0}$ and trace $D^2 \operatorname{cov}(\mathbf{f}^*) \leq \rho$. Given $\mathbf{f}$, $\mathbf{Y}^*$ has mean $\mathbf{f}^*$ and covariance matrix $\sigma^2 I_{n-k}$ and there is no loss of generality in supposing that $\mathbf{Y}^*$ and $\mathbf{f}^*$ are jointly normal, since only first and second moments enter. Since $E\|C^*\mathbf{Y}^* - \mathbf{f}^*\|^2$ is linear in Σ^*, the covariance of $\mathbf{f}^*$ and quadratic in C^*, the min and max may be reversed [Karlin (1959), page 281]. Then, for a fixed prior, the $\hat{\mathbf{f}}^*$ that minimizes $E\|\hat{\mathbf{f}}^* - \mathbf{f}^*\|^2$ is simply the Bayes estimate $\Sigma^*(\Sigma^* + \sigma^2 I_{n-k})^{-1}\mathbf{Y}^*$. Thus

$$\begin{aligned} \min_{C^*} \max_{\text{priors in } \Omega^*} E\|C^*\mathbf{Y}^* - \mathbf{f}^*\|^2 &= \sigma^2 \max_{\Sigma^*} \operatorname{trace} \Sigma^*\left(\Sigma^* + \sigma^2 I_{n-k}\right)^{-1} \\ &\equiv \sigma^2 \max_{\Sigma^*} h(\Sigma^*). \end{aligned}$$

Here the maximum is taken over all symmetric nonnegative definite Σ^* for which trace $D^2\Sigma^* \le \rho$.

We will show that the Σ^* that maximizes h is $\hat{\Sigma}^*$, as defined in Theorem 1. Since h is a concave function of Σ^*, symmetric nonnegative definite, it suffices to show that if R is symmetric, with

$$\text{trace } D^2(\hat{\Sigma}^* + \delta R) \le \rho \tag{2a}$$

and

$$\hat{\Sigma}^* + \delta R \ge 0 \tag{2b}$$

for positive δ sufficiently small, then

$$\lim_{\delta \to 0} \delta^{-1}\big(h(\hat{\Sigma}^* + \delta R) - h(\hat{\Sigma}^*)\big) \le 0.$$

Fix R satisfying (2). Then

$$\begin{aligned}
&\lim_{\delta \to 0} \delta^{-1}\big(h(\hat{\Sigma}^* + \delta R) - h(\hat{\Sigma}^*)\big) \\
&\quad = \sigma^2 \,\text{trace}\big(\hat{\Sigma}^* + \sigma^2 I_{n-k}\big)^{-2} R \\
&\quad = \sigma^2 \sum_{i:\, \alpha > \sigma^2 |D_{ii}|} R_{ii} D_{ii}^2/\alpha^2 + \sigma^2 \sum_{i:\, \alpha \le \sigma^2 |D_{ii}|} R_{ii}/\sigma^4 \\
&\quad = \sigma^2 \,\text{trace}\, D^2 R/\alpha^2 + \sum_{i:\, \alpha \le \sigma^2 |D_{ii}|} R_{ii}\big(\alpha^2 - \sigma^4 D_{ii}^2\big)/(\alpha^2 \sigma^2) \\
&\quad \le 0,
\end{aligned}$$

since (2a) implies that trace $D^2R \le 0$ and (2b) implies that $R_{ii} \ge 0$ for all i with $\alpha \le \sigma^2 |D_{ii}|$.

The theorem follows immediately, since $\hat{C}^* = \hat{\Sigma}^*(\hat{\Sigma}^* + \sigma^2 I_{n-k})^{-1}$ and $\hat{C} = P_1 + P_2 \hat{C} P_2 = P_1 + Q\hat{C}^* Q^t$. □

The minimax estimate of $\hat{\mathbf{f}}^*$ is the sum of the projection of $\mathbf{Y}$ onto the null space of A and a damped projection of $\mathbf{Y}$ onto the row space of A. Since $\hat{C}_{11}^* \ge \hat{C}_{22}^* \ge \cdots \ge \hat{C}_{n-k,n-k}^*$, the damping occurs in the direction of eigenvectors corresponding to large eigenvalues of A^tA. Consider the case in which $\mathbf{t}_i \in \mathscr{R}$ and $A\mathbf{f}$ is the $n-2$ vector of second divided differences of $\mathbf{f}$. Let $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{n-2}$ (the columns of Q) denote the eigenvectors of A^tA corresponding to eigenvalues $D_{11}^2, \ldots, D_{n-2,n-2}^2$. Since $\mathbf{q}_{n-2}$ maximizes (over unit length vectors) $\|A\mathbf{q}\|$, the norm of second divided differences of $\mathbf{q}$, $\mathbf{q}_{n-2}$ can be considered the highest frequency or roughest direction in $\mathscr{R}^n$. Similarly, $\mathbf{q}_{n-3}$ may be thought of as the second roughest direction in $\mathbf{R}^n$. Thus the minimax estimate of $\mathbf{f}$ is based on a $\mathbf{Y}$ that has had high frequencies lessened or removed. Under the prior for which $\hat{\mathbf{f}}^*$ is Bayes, the linear part of $\mathbf{f}$ is diffuse and the $\mathbf{q}_j^t\mathbf{f}$'s, $j = 1, \ldots, n-k$, are independent with means $\mathbf{0}$ and variances s_{jj} with $s_{11} \ge s_{22} \ge \cdots \ge s_{n-k,n-k}$.

For general A, two extreme cases: no damping of high frequency directions (ρ approaches infinity) and complete damping ($\rho \to 0$) are easily studied. As ρ

approaches 0, $\sum D_{ii}^2 s_{ii} = \rho \to 0$, so $s_{ii} \to 0$ and $\hat{C}_{ii}^* = s_{ii}/(s_{ii} + \sigma^2) \to 0$. Thus, $\hat{\mathbf{f}}$ approaches $P_1\mathbf{Y}$, the least squares predictor of $\mathbf{f}$ in the linear model $Q^t\mathbf{f} = \mathbf{0}$. As ρ approaches infinity, by Lemma 1, $J = n - k$ and $\alpha = (\rho + \sigma^2 \sum_1^{n-k} D_{ii}^2)/\sum_1^{n-k} |D_{ii}| \to \infty$. Thus $\hat{C}_{ii}^* \to 1$ and $\hat{\mathbf{f}}$ approaches $\mathbf{Y}$.

LEMMA 1. *Let*

$$g(j) = \sum_1^j D_{ii}^2 - |D_{jj}| \sum_1^j |D_{ii}|.$$

Then $g(1) = 0$ *and* $g(j) \geq g(j+1)$ *with strict inequality if* $D_{jj}^2 < D_{j+1,j+1}^2$. *Then* $J = J_n(\rho)$ *is the largest* $i \leq n - k$ *such that* $\hat{C}_{ii}^*(\rho) > 0$ *if and only if* J *is the largest* $i \leq n - k$ *such that* $\rho + \sigma^2 g(i) > 0$.

PROOF. To prove the monotonicity of g, write

$$g(j) - g(j+1) = \left(|D_{j+1,j+1}| - |D_{jj}|\right) \sum_1^J |D_{ii}| \geq 0$$

with strict inequality if $|D_{j+1,j+1}| > |D_{jj}|$.

Let J be equal to the largest $i \leq n - k$ such that $\hat{C}_{ii}^* > 0$ and consider the case $J < n - k$. The case that $J = n - k$ is similar. Then

$$\sigma^2 |D_{JJ}| < \alpha \leq \sigma^2 |D_{J+1,J+1}|,$$

where α satisfies

$$\sum_1^J D_{ii}^2 \left(\alpha/|D_{ii}| - \sigma^2\right) = \rho.$$

Solving for α and substituting into the inequality yields

$$\rho + \sigma^2 g(J+1) \leq 0 < \rho + \sigma^2 g(J).$$

A reversal of the argument completes the proof. □

Theorem 2 gives sufficient conditions for the asymptotic normality of linear combinations of the $\hat{f}(\mathbf{t}_i)$'s, conditional on f. Studying linear combinations of parameter estimates is often used when the number of parameters to be estimated approaches infinity [see, e.g., Portnoy (1984)].

THEOREM 2. *Suppose that* (1) *holds and that the* ε_i*'s are independent and identically distributed. Let* $\Delta = \Delta(\rho)$ *be an* $(n - k) \times (n - k)$ *diagonal matrix with* $\Delta_{ii} = 1$ *if* $s_{ii} > 0$ *and zero otherwise and let* $\mathbf{e}_i$ *be an* n*-vector of zeroes, with a* 1 *in the* i*th position. Let* σ_b^2 *by the conditional variance of* $\mathbf{b}^t\hat{\mathbf{f}}$ *given* f. *If*

$$\lim_{n\to\infty} \max_{1 \leq i \leq n} \|P_1\mathbf{e}_i\| + \|\Delta Q^t\mathbf{e}_i\| = 0,$$

then

$$\lim_{n\to\infty} \sup_{\mathbf{b}:\,\sigma_b>0} \sup_{z} \left| P\left\{ \frac{\mathbf{b}^t\left(\hat{\mathbf{f}} - E\left(\hat{\mathbf{f}}\middle| f\right)\right)}{\sigma_b} \le z \middle| f \right\} - \Phi(z) \right| = 0,$$

where Φ is the standard normal cumulative distribution function.

REMARK. Note that in the case that $Q^t\mathbf{f} = \mathbf{0}$, $E(\hat{\mathbf{f}}|f) = \mathbf{f}$. If $Q^t\mathbf{b} = \mathbf{0}$, then $\mathbf{b}^tE(\hat{\mathbf{f}}|f) = \mathbf{b}^t\mathbf{f}$.

PROOF. Let $\mathbf{b}$ be an n-vector. Then

$$\mathbf{b}^t\left(\hat{\mathbf{f}} - E\left(\hat{\mathbf{f}}\middle| f\right)\right) = \mathbf{b}^t\left(P_1 + Q\hat{C}^*Q^t\right)\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix} \equiv \mathbf{c}^t\boldsymbol{\varepsilon}$$

and

$$\sigma_b^2/\sigma^2 = \|\mathbf{c}\|^2 = \|P_1\mathbf{b}\|^2 + \left\|Q\hat{C}^*Q^t\mathbf{b}\right\|^2 = \|P_1\mathbf{b}\|^2 + \left\|\hat{C}^*Q^t\mathbf{b}\right\|^2.$$

Now

$$\begin{aligned}|c_i| &= \left|\mathbf{e}_i^t\mathbf{c}\right| \le \left|\mathbf{e}_i^tP_1^tP_1\mathbf{b}\right| + \left|\mathbf{e}_i^tQ\hat{C}^*Q^t\mathbf{b}\right|\\ &= \left|(P_1\mathbf{e}_i)^t(P_1\mathbf{b})\right| + \left|(\Delta Q^t\mathbf{e}_i)^t\left(\hat{C}^*Q^t\mathbf{b}\right)\right|\\ &\le \|P_1\mathbf{e}_i\|\,\|P_1\mathbf{b}\| + \|\Delta Q^t\mathbf{e}_i\|\left\|\hat{C}^*Q^t\mathbf{b}\right\|\\ &\le \|\mathbf{c}\|\{\|P_1\mathbf{e}_i\| + \|\Delta Q^t\mathbf{e}_i\|\}.\end{aligned}$$

The theorem now follows from the Lindeberg–Feller theorem [cf. Hájek and Šidák (1965), pages 153–154]. □

3. Minimax estimates when f is smooth. Throughout this section, assume that $\mathbf{t}_i = t_i = i/n$. To apply the results of Section 2 to priors with the kth derivative of f small, A is the $(n-k)\times n$ matrix with $A\mathbf{f} = \boldsymbol{\Delta}^{(k)}(\mathbf{f})$, the vector of kth divided differences based upon $\mathbf{f}$. For instance

$$\boldsymbol{\Delta}^{(1)} = \boldsymbol{\Delta}^{(1)}(\mathbf{f}) = \begin{pmatrix} \dfrac{f(t_2)-f(t_1)}{t_2-t_1}\\ \vdots\\ \dfrac{f(t_n)-f(t_{n-1})}{t_n-t_{n-1}}\end{pmatrix}$$

and

$$\Delta^{(2)}(\mathbf{f}) = \begin{pmatrix} \dfrac{\Delta_3^{(1)} - \Delta_2^{(1)}}{t_3 - t_1} \\ \vdots \\ \dfrac{\Delta_n^{(1)} - \Delta_{n-1}^{(1)}}{t_n - t_{n-2}} \end{pmatrix}.$$

THEOREM 3. *Suppose that* $\rho = \rho_n$ *satisfies*

$$0 < \lim_{n\to\infty} \rho_n/n = \rho^* < 1.$$

Then

$$J = J_n = \max\{i \le n - k : \hat{C}_{ii}^* > 0\} \sim (\beta^* n)^{1/(2k+1)},$$

$$\min\max E(\|\hat{\mathbf{f}} - \mathbf{f}\|)^2 \sim \frac{k}{k+1}\sigma^2(n\beta^*)^{1/(2k+1)},$$

where

$$\beta^* = \rho^* \frac{(k+1)!(k-1)!(2k+1)}{\pi^{2k}\sigma^2}.$$

Furthermore, if $F_n = \{f : \|A(f(t_1), \ldots, f(t_n))^t\|^2 \le \rho_n\}$, *then*

$$\max_{f\in F_n} E\left(\|\hat{\mathbf{f}} - \mathbf{f}\|^2 \middle| f\right) \sim \frac{k}{k+1}\sigma^2(n\beta^*)^{1/(2k+1)}.$$

PROOF. Fix k and let $\gamma_i = 2\{1 - \cos(i\pi/(n-k+1))\}$. Since D_{ii}^2 is an eigenvalue of AA^t, by Lemma 6 of Utreras (1983), for $2k+1 \le i \le n-2k$,

$$(k!)^{-2} n^{2k} \gamma_{i-2k}^k \le D_{ii}^2 \le (k!)^{-2} n^{2k} \gamma_{i+2k}^k.$$

Using these bounds and the fact that $D_{11}^2 \le D_{22}^2 \le \cdots$, one can show that for all K_0 there exists K_1, not depending on i such that

$$\left| D_{ii}^2 - (k!)^{-2}\left(\frac{n}{n-k+1}\right)^{2k} (i\pi)^{2k} \right| \le K_1 i^{2k-1}$$

and

$$\left| |D_{ii}| - (k!)^{-1}\left(\frac{n}{n-k+1}\right)^{k} (i\pi)^{k} \right| \le K_1 i^{k-1}$$

for all $i \le K_0 n^{2/3}$.

Suppose that $j = j_n$ satisfies $j^{2k+1}/n \sim \beta$, $0 < \beta < 1$. Then

$$n^{-1}\sum_1^j D_{ii}^2 \to \frac{\beta\pi^{2k}(k!)^{-2}}{(2k+1)},$$

$$n^{-(k+1)/(2k+1)}\sum_1^j |D_{ii}| \to \frac{\beta^{(k+1)/(2k+1)}\pi^k}{(k+1)!}, \tag{3}$$

$$n^{-k/(2k+1)}|D_{jj}| \to \beta^{k/(2k+1)}\pi^k(k!)^{-1},$$

$$n^{-1}g(j) \to -\beta\frac{\pi^{2k}k}{(k!)^2(2k+1)(k+1)},$$

where g is as defined in Lemma 1. Furthermore, convergence is uniform in β in compact subsets of $(0, 1)$.

To study the asymptotics of J_n, by Lemma 1, J_n satisfies $\sigma^2 g(J_n + 1) + \rho_n \le 0 < \sigma^2 g(J_n) + \rho_n$. The first statement of Theorem 3 follows from (3) and the monotonicity of g.

To calculate the asymptotic minimax Bayesian mean squared error, write

$$\rho_n = \sum_1^J D_{ii}^2\left(\frac{\alpha_n}{|D_{ii}|} - \delta^2\right)$$

and thus

$$\alpha_n = \frac{\rho_n + \sigma^2\Sigma_1^J D_{ii}^2}{\Sigma_1^J|D_{ii}|} \sim (n\beta^*)^{k/(2k+1)}\frac{\pi^k\sigma^2}{k!}.$$

The asymptotic expression is then calculated by applying (3) and the asymptotic value of J_n to

$$E\|\hat{C}\mathbf{Y} - \mathbf{f}\|^2 = \sigma^2\left\{k + J - \frac{\sigma^2}{\alpha}\sum_1^J |D_{ii}|\right\}.$$

To calculate the maximum conditional mean squared error, fix f. Then

$$E\left(\|\hat{\mathbf{f}} - \mathbf{f}\|^2 \middle| f\right) = \sigma^2 k + \sigma^2 \operatorname{trace}(\hat{C}^*)^2 + \|(\hat{C}^* - I_{n-k})Q^t\mathbf{f}\|^2.$$

By the asymptotic value of J_n and by (3),

$$\operatorname{trace}(\hat{C}^*)^2 \sim (n\beta^*)^{1/(2k+1)}\frac{2k^2}{(k+1)(2k+1)}.$$

Also, writing ρ for ρ_n,

$$\begin{aligned}
\max_{\|A\mathbf{f}\|^2 \le \rho} \left\|(\hat{C}^* - I_{n-k})Q^t\mathbf{f}\right\|^2 &= \max_{\|D\mathbf{f}^*\|^2 \le \rho} \left\|(\hat{C}^* - I_{n-k})\mathbf{f}^*\right\|^2 \\
&= \max_{\|\mathbf{v}\|^2 \le \rho} \left\|(\hat{C}^* - I_{n-k})D^{-1}\mathbf{v}\right\|^2 \\
&= \rho \max_{1 \le i \le n-k} \frac{\left(\hat{C}^*_{ii} - 1\right)^2}{D^2_{ii}} \\
&= \rho \max\left\{\sigma^4\alpha^{-2}, \max_{i>J} D_{ii}^{-2}\right\} \\
&= \rho \max\left\{\sigma^4\alpha^{-2}, D^{-2}_{J+1,J+1}\right\} \\
&\sim (n\beta^*)^{1/(2k+1)}\sigma^2 \frac{k}{(2k+1)(k+1)}.
\end{aligned}$$

The last statement of the theorem follows. □

COMMENTS. Theorem 3's assumption that $\rho = O(n)$ is a reasonable one since, if one believes that the variance of the ith component of $\mathbf{f}^*$ is bounded, then one would assume that the trace of $\operatorname{cov}(\mathbf{f}^*) = O(n)$.

Note that the Bayesian and conditional rates of convergence of the mean squared error are equal. Furthermore, a rate of $n^{1/(2k+1)}$ is that which a frequentist would expect for $n\int(\hat{f}(t) - f(t))^2\,dt$ if the function f had k continuous derivatives; see Stone (1982).

If $k = 1$, then the minimax estimate of $\mathbf{f}$ can be easily computed. Let

$$\begin{aligned}
\mathbf{v}_j &= \left(\sin(j\pi/n), \sin(2j\pi/n), \ldots, \sin((n-1)j\pi/n)\right)^t, \\
D^2_{jj} &= 2n^2(1 - \cos(j\pi/n)).
\end{aligned} \tag{4}$$

Then $\mathbf{v}_j$ is an eigenvector of AA^t with corresponding eigenvalue D^2_{jj}. Therefore, the principal value decomposition of A is VDQ^t, where $Q = A^tVD^{-1}$ and the jth column of V is $\mathbf{v}_j/\|\mathbf{v}_j\|$. The projection matrix P_1 is $(P_1)_{ij} = 1/n$.

THEOREM 4. *Suppose that $k = 1$. Then*

$$\lim_{n\to\infty} \sup_{\mathbf{b}:\,\sigma_b>0} \sup_z \left| P\left\{ \frac{\mathbf{b}^t\left(\hat{\mathbf{f}} - E\left(\hat{\mathbf{f}}\middle| f\right)\right)}{\sigma_b} \le z \middle| f \right\} - \Phi(z) \right| = 0.$$

PROOF. By Theorem 2, we must show that

$$\max_{1 \le i \le n} \|P_1\mathbf{e}_i\| \to 0$$

and

$$\max_{1\le i\le n} \left\| \Delta Q^t \mathbf{e}_i \right\| \to 0$$

as n approaches infinity. The first statement is immediate, since $[P_1]_{ij} = n^{-1}$. For J as defined in Theorem 3,

$$\left\| \Delta Q^t \mathbf{e}_i \right\|^2 = \sum_{j=1}^{J} Q_{ij}^2.$$

Let $\mathbf{v}_j$ and D_{jj}^2 be as defined in (4). Then

$$Q_{ij} = \begin{cases} -\dfrac{n(\mathbf{v}_j)_1}{|D_{jj}|\,\|\mathbf{v}_j\|}, & \text{if } i = 1, \\ \dfrac{n\left((\mathbf{v}_j)_{i-1} - (\mathbf{v}_j)_i\right)}{|D_{jj}|\,\|\mathbf{v}_j\|}, & \text{if } 1 < i < n, \\ \dfrac{n(\mathbf{v}_j)_{n-1}}{|D_{jj}|\,\|\mathbf{v}_j\|}, & \text{if } i = n. \end{cases}$$

It is easily shown that

$$\lim_{n\to\infty} \max_{1\le j\le J} \left| \frac{\|\mathbf{v}_j\|^2}{n} - \int_0^1 \sin^2(j\pi x)\,dx \right| = 0$$

and $\int_0^1 \sin^2(j\pi x)\,dx = 1/2$. Since $2(1-\cos u)/u^2$ approaches 1 as u approaches 0,

$$\lim_{n\to\infty} \max_{1\le j\le J} \left| \frac{D_{jj}^2}{n^2(j\pi/n)^2} - 1 \right| = 0.$$

Thus

$$\lim_{n\to\infty} \max_{1\le j\le J} \left| \frac{1}{D_{jj}^2\|\mathbf{v}_j\|^2} - \frac{2}{\pi^2 j^2 n} \right| = 0.$$

Since there exists K, not depending on i, j or n, such that $|(\mathbf{v}_j)_1| \le Kj/n$, $|(\mathbf{v}_j)_{i-1} - (\mathbf{v}_j)_i| \le Kj/n$, $1 < i < n$, and $|(\mathbf{v}_j)_{n-1}| \le Kj/n$, $Q_{ij}^2 = O(1/n)$ uniformly in $i \le n$ and $j \le J$. Therefore,

$$\max_{1\le i\le n} \sum_{j=1}^{J} Q_{ij}^2 = O(J/n) \to 0. \qquad \square$$

4. Comments. Our approach is inspired by several techniques in nonparametric (or smoothing) regression.

In standard Bayesian methods, a prior on the entire function f is chosen and thus the resulting estimate is best for a particular prior, rather than for a whole class of priors. Wahba (1978) shows that the usual spline function estimates are Bayes estimates. In a more classical Bayesian vein, O'Hagan

(1978) develops a general framework for Bayesian regression under both proper and vague priors. Weerahandi and Zidek (1988) use Taylor series expansions of f to justify a particular form of f's prior, which depends upon a few hyperparameters. Since these hyperparameters can be estimated from the data, this method may produce estimates which perform well over a class of priors.

Frequentist approaches to nonparametic regression can be divided into two classes. In the first, one finds the $\hat{f}$ that fits the data well, but is not too rough. For instance, the estimate $\hat{f}$ that minimizes $\|\hat{\mathbf{f}} - \mathbf{Y}\|$ subject to $\int(\hat{f}^{(k)}(t))^2\,dt \le \rho$ is a spline; see, for example, Eubank (1988). In the second approach, a minimax method, one seeks $\hat{f}$ close to all f in a class of smooth functions. For evaluation of f at a fixed point t, one might seek $\hat{f}(t)$ linear, to minimize the maximum (over all f's in a particular class) $E(\hat{f}(t) - f(t))^2$. In the approximately linear model [see Legostaeva and Shiryayev (1971), or Sacks and Ylvisaker (1978)] one considers all f with, roughly, $|f^{(k)}(t)| \le \rho$, a known constant. Li (1982) considers all f with $\int(f^{(k)}(x))^2\,dx \le \rho$. To estimate the entire function, Speckman (1985) and Nussbaum (1985) find the linear estimator which minimizes the maximum of $E\|\hat{\mathbf{f}} - \mathbf{f}\|^2$. The maximum is taken over all f with $\int(f^{(k)}(t))^2\,dt \le \rho$.

Our Bayesian minimax approach places seemingly softer smoothness constraints on f. However, the Speckman and Nussbaum estimate is a particular case of our estimate. Specifically, let $\phi_1, \ldots, \phi_n$ be the basis for the space of natural splines of degree $2k - 1$ with knots at $t_1, \ldots, t_n$, proposed by Demmler and Reinsch (1975). Then $\int\phi_i^{(k)}\phi_j^{(k)} = \lambda_i I\{i = j\}$, where $0 = \lambda_1 = \cdots = \lambda_k < \lambda_{k+1} < \cdots < \lambda_n$. Let Λ be an $(n - k) \times (n - k)$ diagonal matrix with iith element equal to λ_{k+i} and let Q be $n \times (n - k)$ with $Q_{ij} = \phi_{k+j}(t_i)$. If, in Theorem 1, we let $A = \Lambda^{1/2}Q^t$, the resulting minimax Bayes estimator of $\mathbf{f}$ is the same as the Speckman and Nussbaum estimate.

The Bayesian minimax approach has also been employed in other contexts. Leamer (1982), Polasek (1984) and DasGupta and Studden (1989) estimate a normal mean $\boldsymbol{\theta}$, but consider classes of priors on $\boldsymbol{\theta}$ different from those considered here. The Bayesian minimax approach is used in ranking and selection, where it is often called the Γ minimax approach; see, for example, Gupta and Hwang (1977).

Acknowlegement. The authors would like to thank an anonymous reviewer, whose suggestions led to a shortening of the proof of Theorem 1.

REFERENCES

DasGupta, A. and Studden, W. J. (1989). Frequentist behavior of robust Bayes estimates of normal means. *Statist. Decisions* **7** 333–361.

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

Eubank, R, (1988). *Spline Smoothing and Nonparametric Regression*. North-Holland, Amsterdam.

GUPTA, S. S. and HWANG, D.-Y. (1977). On some Γ-minimax selection and multiple comparison procedures. In *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. S. Moore, eds.) 139–155. Academic, New York.

HÁJEK, J. and ŠIDÁK, Z. (1965). *Theory of Rank Tests*. Academic, New York.

KARLIN, S. (1959). *Mathematical Methods and Theory in Games Programming and Economics* **1**. Addison-Wesley, Reading, Mass.

LEAMER, E. E. (1982). Sets of posterior means with bounded variance prior. *Econometrica* **50** 725–736.

LEGOSTAEVA, I. L. and SHIRYAEV, A. N. (1971). Minimax weights in a trend detection problem of a random process. *Theory Probab. Appl.* **16** 344–349.

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

NUSSBAUM, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13** 984–997.

O'HAGAN, A. (1978). Curve fitting and optimal design for prediction. *J. Roy. Statist. Soc. Ser. B* **40** 1–42.

POLASEK, W. (1984). Multivariate regression systems: Estimation and sensitivity analysis of two-dimensional data. In *Robustness of Data Analysis* (J. B. Kadane, ed.) 229–309. North-Holland, Amsterdam.

PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. *Ann. Statist.* **12** 1298–1309.

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

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–50.

SPECKMAN, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13** 970–983.

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

UTRERAS, F. (1983). Natural spline functions, their associated eigenvalue problem. *Numer. Math.* **42** 107–117.

WAHBA, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

WEERAHANDI, S. and ZIDEK, J. V. (1988). Bayesian nonparametric smoothers for regular processes. *Canad. J. Statist.* **16** 61–74.

DEPARTMENT OF STATISTICS
2021 WEST MALL
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, BRITISH COLUMBIA
CANADA V6T 1W5

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109