The Annals of Statistics
1993, Vol. 21, No. 3, 1522–1544

SPLINE SMOOTHING WITH AN ESTIMATED ORDER PARAMETER[1]

BY MICHAEL L. STEIN

University of Chicago

Smoothing splines of a fixed order are commonly used as nonparametric regression estimates. The only parameter, then, that needs to be estimated is the smoothing parameter, which is often estimated using some form of cross validation. This work allows the order of the smoothing spline to be estimated using a model in which the order parameter is continuous. Within this setting, generalized cross validation and modified maximum likelihood estimates of the order and smoothing parameters are compared. I show that there are both stochastic and fixed regression functions for which modified maximum likelihood yields asymptotically better estimates of the regression function than generalized cross validation. These results are supported by a small simulation study, although there are functions for which the asymptotic results can be misleading even for fairly large sample sizes.

1. Introduction. Smoothing splines are a popular approach to nonparametric estimation of regression functions [Eubank (1987) and Wahba (1989)]. The smoothing spline estimate minimizes a sum of the residual sum of squares and an integral of the square of the mth derivative of the estimate [see (1.1)]. The constant that determines the relative contribution of these two terms is known as the smoothing parameter and m is the order of the spline. Most work on using smoothing splines for nonparametric regression assumes that the order of the spline is fixed and only the smoothing parameter is estimated, commonly by cross validation or generalized cross validation [Craven and Wahba (1979)]. While it is possible to use cross validation to select the order of the spline in addition to the smoothing parameter [Wahba and Wendelberger (1980) and Gamber (1979)], the asymptotic analysis is muddied by the discrete nature of the order parameter. Viewing the smoothing spline as an optimal linear predictor under a certain stochastic model for the regression function provides a natural way to let the order parameter be continuous. Now the class of nonparametric estimates of the regression function depends on two continuous parameters, and it is possible to use more standard methods to obtain asymptotic results.

Received January 1990; revised June 1992.

[1]Research partially supported by NSF Grant DMS-89-02667. This work was done using computer facilities supported in part by NSF Grants DMS-86-01732, DMS-87-03942 and DMS-89-05292 awarded to the Department of Statistics at the University of Chicago, and by the University of Chicago Block Fund.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62M20.

Key words and phrases. Nonparametric regression, Gaussian process, kriging, generalized cross validation, modified maximum likelihood.

The central problem this paper addresses is the effect of estimating these two parameters on the subsequent estimate of the regression function. Under weak assumptions, past research has shown that selecting the smoothing parameter by cross validation can yield estimates of the regression function that do as well asymptotically as using the optimal values of these parameters [Wahba (1985), Speckman (1985) and Li (1986)]. Such results do not allow us to distinguish between two procedures that both satisfy this optimality property. This work studies the properties of regression function estimates based on generalized cross validation and modified maximum likelihood estimates of the smoothing and order parameters. Specifically, for certain classes of fixed and stochastic regression functions, I derive the second term in the asymptotic mean square error of the regression function estimate based on approximations to the modified maximum likelihood and generalized cross validation estimates. While I would expect that all of the results given in this paper for the approximations to these estimates also apply to the estimates themselves, I have not been able to prove that this is the case. These results suggest that there are both fixed and stochastic regression functions under which modified maximum likelihood estimates of these parameters yield asymptotically better estimates of the regression function than generalized cross validation.

We will restrict attention to evenly spaced observations and estimates of the regression function that have an interpretation as optimal linear predictors assuming the regression function is a stationary process on the circle. When the order m is an integer, these estimates correspond to what are known as periodic smoothing splines [Cogburn and Davis (1974)]. Wahba (1975) and Rice and Rosenblatt (1981) have also used periodic smoothing splines to obtain results that are much more difficult to derive in a more general setting.

Suppose we observe

$$Y_i = f((i-1)n^{-1}) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the ε_i's are iid $N(0, \theta_3)$ and f and θ_3 are unknown. The periodic smoothing spline estimate of f of order m is the function $\hat{f}$ that minimizes

$$n^{-1} \sum_{i=1}^{n} \{Y_i - g((i-1)n^{-1})\}^2 + \lambda(2\pi)^{-2m} \int_0^1 \{g^{(m)}(t)\}^2 \, dt \tag{1.1}$$

among those functions g with an absolutely continuous $(m-1)$th derivative satisfying the periodic boundary condition

$$g^{(j)}(0) = g^{(j)}(1) \quad \text{for } j = 0, \ldots, m-1.$$

Common practice would be to select a value of m a priori, often $m = 2$, and then estimate λ using cross validation.

The function minimizing (1.1) also can be interpreted as the best linear unbiased predictor of f under a stochastic model for the regression function. Suppose $Ef(x) = \mu$, μ unknown,

$$\operatorname{cov}(f(x), f(x')) = \theta_1\theta_3 \sum_{j=1}^{\infty} j^{-2m} \cos\{2\pi j(x - x')\},$$

and the ε_i's are uncorrelated with f. This stochastic process is stationary on the circle of unit perimeter. The best linear unbiased predictor of $f(x)$ is of the form $\Sigma c_i y_i$, where the c_i's are chosen to minimize $\operatorname{var}(f(x) - \Sigma c_i y_i)$ subject to the unbiasedness constraint $\Sigma c_i = 1$. For $\lambda = 2/(\theta_1 n)$, the best linear unbiased predictor of $f(x)$ equals $\hat{f}(x)$, where $\hat{f}$ is the function minimizing (1.1) subject to the periodic boundary condition. Viewed in this stochastic framework, there is no compelling reason to restrict m to being an integer, so we will consider the model

$$\operatorname{cov}(f(x), f(x')) = \theta_1 \theta_3 \sum_{j=1}^{\infty} j^{-\theta_2} \cos\{2\pi j(x - x')\}, \tag{1.2}$$

where $\theta_2 > 1$. Whether or not we take this stochastic model seriously, we can study the class of estimates of f we get as θ_1 and θ_2 vary as a generalization of the class we get from (1.1) where $\lambda > 0$ and m is a positive integer.

Given that we want to use an estimate of f from the class defined by the stochastic model in (1.2), we need a method for choosing θ_1 and θ_2. One standard possibility is generalized cross validation (GCV). Another is the modified maximum likelihood (MML) estimates under the assumptions that f is a Gaussian process with mean μ, covariance function as in (1.2) and the ε_i's are iid $N(0, \theta_3)$ and independent of f. Wahba (1985) and Stein (1990) have compared GCV and MML estimates of θ_1 when $\theta_2 = 2m$ is fixed. Wahba (1985) showed that GCV appears to select nearly optimal values of θ_1 for large sample sizes in a much broader set of circumstances than MML. Stein (1990) showed that under the Gaussian model for the regression function and assuming that the selected value of m is correct, while both methods of estimation yield asymptotically optimal predictions of the stochastic regression function, the likelihood method does have a large sample advantage in the second term in an asymptotic expansion for the mean square prediction error.

Sections 2 and 3 investigate the properties of the MML and GCV estimates assuming the stochastic model for the regression function is valid. Section 2 gives the asymptotic distribution for $\hat{\theta}$, the MML estimate of $\theta = (\theta_1, \theta_2, \theta_3)'$. I have been unable to obtain rigorous proofs for the properties of $\tilde{\theta}$, the GCV estimate of (θ_1, θ_2). Instead, I consider the properties of an approximation to $\tilde{\theta}$, denoted by $\tilde{\theta}^*$, based on linearizing the estimating equations obtained from the GCV criterion function. While it appears reasonable that $\tilde{\theta}$ and $\tilde{\theta}^*$ should have similar asymptotic behavior, that remains to be proven. Both $\hat{\theta}$ and $\tilde{\theta}^*$ are asymptotically normal and the rates of convergence for the estimates of $(\theta_1, \theta_2)'$ are the same. However, the MML estimates have much higher asymptotic efficiency, particularly for large θ_2. An interesting feature of these results is that the asymptotic correlation for either $\hat{\theta}_1$ and $\hat{\theta}_2$ or $\tilde{\theta}_1^*$ and $\tilde{\theta}_2^*$ is 1.

Section 3 studies the mean square error of predictions of $f(0)$ for Gaussian f based on estimated values of θ_1 and θ_2. Define $\hat{f}(0; \theta')$ to be the best linear unbiased predictor of $f(0)$ for a particular θ' and $\hat{\theta}^*$ to be the approximation to $\hat{\theta}$ we obtain by linearizing the likelihood equations. Predictions based on either $\hat{\theta}^*$ or $\tilde{\theta}^*$ are asymptotically optimal and the leading term in the relative

increase in mean square prediction error due to using either estimator of θ rather than the true θ is of order n^{-1/θ_2}. However, the constant multiplying this n^{-1/θ_2} term can be much smaller for $\hat{\theta}^*$ than for $\tilde{\theta}^*$, especially for large θ_2.

Sections 4 and 5 investigate the asymptotic properties of these procedures for a class of fixed regression functions. Suppose

$$f(x) = a_0 + \sum_{j=1}^{\infty} \{a_j \cos(2\pi jx) + b_j \sin(2\pi jx)\},$$

where, roughly,

$$\sum_{j=m}^{m+k} (a_j^2 + b_j^2) \approx 2c\theta_3 \sum_{j=m}^{m+k} j^{-\nu},$$

for some $c > 0$ whenever m is large and k is not negligible relative to m. In Section 4, $\theta_1 = c$ and $\theta_2 = \nu$ are shown to be the asymptotically optimal values for θ_1 and θ_2 in terms of estimating f. However, using values of θ_2 very different than ν can yield estimates of f that are asymptotically only slightly suboptimal if θ_1 is chosen appropriately. Under the much stronger assumption $a_j^2 + b_j^2 = 2c\theta_3 j^{-\nu} + O(j^{-\eta})$ for some $\eta > \nu + 1/2$, the approximations to the MML and GCV estimates of (θ_1, θ_2) are asymptotically normal with limiting mean (c, ν). The asymptotic variances are of the same order of magnitude as in the stochastic setting but with considerably smaller constants for both approximate estimates. Somewhat surprisingly, the asymptotic advantage of the approximate MML estimate over the approximate GCV estimate is even stronger than in the stochastic setting. This advantage in estimating ν and c translates into an advantage in estimating f similar to that obtained in the stochastic setting.

Section 6 describes the results of a series of simulations for three closely related fixed regression functions. For a fixed regression functions satisfying $a_j^2 + b_j^2 = 9j^{-4}$ for all $j \geq 1$, MML yields clearly better estimates of the regression function than GCV, in line with the asymptotic theory. However, by setting $a_1^2 + b_1^2 = 0$ but leaving the other Fourier coefficients of f unchanged, GCV now yields better estimates of f than MML for $n = 100$, about the same for $n = 400$, and only for $n = 1600$ does MML yield clearly better estimates. In contrast, if $a_1^2 + b_1^2$ is increased rather than decreased, MML produces better estimates of f than GCV for $n = 100$ and $n = 400$. Since changing a single Fourier coefficient does not affect the asymptotic results, we see that the asymptotic results do not tell the whole story.

2. **Parameter estimation.** Throughout Sections 2 and 3 we will assume that f is a mean μ Gaussian process with homogeneous covariance function

$$\text{cov}(f(x), f(x+h)) = \sigma^2 \sum_{k=1}^{\infty} k^{-\theta_2} \cos(2\pi kh), \tag{2.1}$$

where $\theta_2 > 1$. All probability statements and expectations in these two sections will be unconditional over the distribution for f. We observe

$$Y_j = f((j-1)/n) + \varepsilon_j, \qquad j = 1, \ldots, n,$$

where the ε_j's are iid $N(0, \theta_3)$ and independent of f. Then $Y_n = (Y_1, \ldots, Y_n)$ has a circulant covariance matrix with ijth element

$$\theta_3 \left\{ I_{\{i=j\}} + \theta_1 \sum_{k=1}^{\infty} k^{-\theta_2} \cos(2\pi k(i-j)/n) \right\},$$

where $\theta_1 = \sigma^2/\theta_3$. All symmetric circulant matrices of order n share a common set of real eigenvectors [Brockwell and Davis (1987), page 130], and by taking S to be the matrix whose rows are these eigenvectors, we get that $\mathbf{Z}_n = (Z_1, \ldots, Z_n)' = S\mathbf{Y}_n$ has independent Gaussian components with $EZ_j = 0$ for $j < n$ and $EZ_n = n^{1/2}\mu$. Furthermore, the variance of Z_j can be shown to be

$$\theta_3 + \tfrac{1}{2}\theta_1\theta_3 n^{1-\theta_2} H(\theta_2, j/n) =_{\text{def}} \theta_3 w_j(\theta_1, \theta_2),$$

where

$$H(\theta_2, x) = \sum_{p=-\infty}^{\infty} |p + x|^{-\theta_2}.$$

Here and elsewhere define $0^{-\theta_2} = 0$.

The MML estimate $\hat{\theta}$ of θ maximizes the likelihood of $(Z_1, \ldots, Z_{n-1})$ with respect to θ. Let $\mathscr{I}$ be the expected Fisher information matrix for θ in $(Z_1, \ldots, Z_{n-1})$, and $\mathscr{I}_{ij}$ the ijth element of $\mathscr{I}$. The following proposition is proven in Appendix A.

PROPOSITION 2.1. *For* $0 < \theta_1 < \infty$, $1 < \theta_2 < \infty$, $\theta_3 > 0$ *and* $\beta = \theta_2^{-1}\log(\theta_1/2)$, *as* $n \to \infty$,

$$\mathscr{I}_{11} = n^{1/\theta_2}\frac{1}{4}\left(\frac{\theta_1}{2}\right)^{1/\theta_2 - 2} I_0(1 + O(n^{-\varepsilon})),$$

$$\mathscr{I}_{12} = -n^{1/\theta_2}\frac{1}{2}\left(\frac{\theta_1}{2}\right)^{1/\theta_2 - 1} (\theta_2^{-1} I_0 \log n + \beta I_0 + I_1)(1 + O(n^{-\varepsilon})),$$

$$\mathscr{I}_{22} = n^{1/\theta_2}\left(\frac{\theta_1}{2}\right)^{1/\theta_2} \{\theta_2^{-2} I_0 \log^2 n + 2\theta_2^{-1}(\beta I_0 + I_1)\log n + \beta^2 I_0 + 2\beta I_1 + I_2\}(1 + O(n^{-\varepsilon})),$$

$$\mathscr{I}_{33} = n/(2\theta_3^2),$$

$$\mathscr{I}_{13} = O(n^{1/\theta_2})$$

and

$$\mathscr{I}_{23} = O(n^{1/\theta_2}\log n),$$

where $\varepsilon = \min(\theta_2^{-1}, (2\theta_2 - 1)(\theta_2 - 1)/(3\theta_2 - 1))$, *and*

$$I_j = \int_0^\infty \frac{\log^j y\, dy}{(1 + y^{\theta_2})^2}.$$

From Proposition 2.1, it follows that

$$|\mathscr{I}| = n^{1+2/\theta_2} \frac{1}{8\theta_3^2} (\theta_1/2)^{2/\theta_2 - 2} (I_0 I_2 - I_1^2)(1 + O(n^{-\delta} \log^2 n)),$$

where

$$\delta = \min\left(\frac{1}{\theta_2}, 1 - \frac{1}{\theta_2}, \frac{(\theta_2 - 1)(2\theta_2 - 1)}{3\theta_2 - 1} \right).$$

It can be shown that

$$I_0 I_2 - I_1^2 = \frac{\pi^2}{\theta_2^4 \sin^2(\pi/\theta_2)} \left[\left\{ \frac{\pi(\theta_2 - 1)}{\theta_2 \sin(\pi/\theta_2)} \right\}^2 - 1 \right],$$

which is positive for $\theta_2 > 1$. Defining $\mathscr{I}^{ij}$ to be the ijth element of $\mathscr{I}^{-1}$, we have the following corollary to Proposition 2.1:

COROLLARY 2.1. *Under the same conditions as Proposition* 2.1, *as* $n \to \infty$,

$$\mathscr{I}^{11} = \frac{n^{-1/\theta_2} 4 (\theta_1/2)^{-1/\theta_2 + 2}}{I_0 I_2 - I_1^2} \{\theta_2^{-2} I_0 \log^2 n + 2\theta_2^{-1}(\beta I_0 + I_1) \log n + \beta^2 I_0 + 2\beta I_1 + I_2\}(1 + O(n^{-\delta} \log^2 n)),$$

$$\mathscr{I}^{12} = \frac{n^{-1/\theta_2} 2 (\theta_1/2)^{-1/\theta_2 + 1}}{I_0 I_2 - I_1^2} (\theta_2^{-1} I_0 \log n + \beta I_0 + I_1)(1 + O(n^{-\delta} \log^2 n)),$$

$$\mathscr{I}^{13} = O(n^{-1} \log^2 n),$$

$$\mathscr{I}^{22} = \frac{n^{-1/\theta_2} (\theta_1/2)^{-1/\theta_2}}{I_0 I_2 - I_1^2} (I_0 + O(n^{-\delta} \log^2 n)),$$

$$\mathscr{I}^{23} = O(n^{-1} \log n)$$

and

$$\mathscr{I}^{33} = 2\theta_3^2 n^{-1}(1 + O(n^{-\delta} \log^2 n)).$$

If we had only derived the highest order term for each $\mathscr{I}_{ij}$ in Proposition 2.1, we would not have been able to obtain the highest order terms in $\mathscr{I}^{-1}$. Using this corollary, we can obtain:

COROLLARY 2.2. *Under the same conditions as Proposition* 2.1, *there exists a sequence of local maxima of the likelihood equations with asymptotic distribution* $N(\theta, \mathscr{I}^{-1})$.

The proof of this corollary is an easy application of Proposition 2.1 and Corollary 2.1 to Theorem 1 of Mardia and Marshall (1984), which is in turn an application of a general theorem on maximum likelihood estimates due to Sweeting (1980).

Assuming that the global maximum to the likelihood function has the same asymptotic behavior as given in Corollary 2.2, we have that the asymptotic correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$ is 1. Furthermore, for any function h from $\mathbb{R}^2$ to $\mathbb{R}^2$ that is smooth in a neighborhood of (θ_1, θ_2), the components of h will also have asymptotic correlation of 1. By taking a transformation that depends on n, we can get a nondegenerate limiting covariance matrix. Specifically, for $t = (t_1, t_2)$, define $u(t) = (u_1(t), u_2(t))'$, where

$$u_1(t) = \frac{t_1 - \theta_1}{\theta_1} - \log n \frac{t_2 - \theta_2}{\theta_2}$$

and $u_2(t) = t_2 - \theta_2$, the dependence of u on n and θ being suppressed. Then

$$\begin{pmatrix} n^{1/(2\theta_2)} u(\hat{\theta}) \\ n^{1/2}(\hat{\theta}_3 - \theta_3) \end{pmatrix} \tag{2.3}$$

$$\to_{\mathscr{L}} N\left(\mathbf{0}, \begin{pmatrix} \frac{(\theta_1/2)^{-1/\theta_2}}{I_0 I_2 - I_1^2} \begin{pmatrix} \beta^2 I_0 + 2\beta I_1 + I_2 & \beta I_0 + I_1 & 0 \\ \beta I_0 + I_1 & I_0 & 0 \\ 0 & 0 & 2\theta_3^2 \end{pmatrix} \end{pmatrix}\right),$$

where it is understood that $u(\hat{\theta})$ means $u(\hat{\theta}_1, \hat{\theta}_2)$.

This paper makes no attempt to prove similar results for GCV estimates or for MML estimates when the regression function is fixed. Instead, I will only consider the properties of approximations to these estimates obtained by linearizing the estimating equations that are obtained by setting derivatives of the GCV criterion function or the likelihood to 0. Under certain regularity conditions [Crowder (1986)], these linearized approximations are asymptotically equivalent to the actual estimates. However, even if we could verify the conditions in Theorem 3.3 of Crowder (1986), we would still not be able to obtain rigorous results on the mean square errors of estimates of the regression function with estimated θ, as this theorem says nothing about the convergence of moments of estimates of θ.

Let us now consider the GCV estimates $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$ of θ_1 and θ_2 under the stochastic model for f. The GCV estimate minimizes

$$\frac{n^{-1}\sum_{j=1}^{n-1}\left(Z_j / w_j(\theta_1, \theta_2)\right)^2}{\left\{n^{-1}\sum_{j=1}^{n-1} w_j(\theta_1, \theta_2)^{-1}\right\}^2}. \tag{2.4}$$

By taking derivatives of this equation with respect to θ_1 and θ_2, linearizing the resulting equations and setting certain quadratic forms equal to their expected values (see Appendix A), we obtain an approximation to $\tilde{\theta}$ which we will denote by $\tilde{\theta}^*$. The MML has a similar linearized approximation which we will denote

TABLE 1
Constants for comparing asymptotic properties for MML and GCV estimates of θ under the stochastic model for the regression function given in (2.1)

θ_2	I_0/d_M	I_1/d_M	I_2/d_M	L_{00}/d_C^2	L_{10}/d_C^2	L_{11}/d_C^2	γ_M	γ_C
2	0.868	−0.868	2.14	5.45	−1.82	2.55	0.318	0.637
4	1.07	−1.20	2.54	32.6	−10.4	6.01	0.225	0.750
6	1.08	−1.20	2.47	94.4	−23.6	9.31	0.159	0.743
8	1.08	−1.18	2.40	2.05×10^2	−41.4	12.5	0.122	0.731
10	1.07	−1.15	2.34	3.78×10^2	−63.8	15.7	0.109	0.726
20	1.04	−1.09	2.19	2.67×10^3	−246.	31.2	0.0498	0.697

See (2.3) and (2.5) for the relationship between columns 2–7 and the asymptotic covariance matrix of the estimates. See (3.3) and (3.4) for the definitions of γ_M and γ_C. $d_M = I_0 I_2 - I_1^2$ and $d_C = I_{01} I_{21} - I_{11}^2$.

by $\hat{\theta}^*$; (2.3) still holds if $u(\hat{\theta})$ is replaced by $u(\hat{\theta}^*)$. Define

$$I_{jk} = \int_0^\infty \frac{y^{k\theta_2} \log^j y \, dy}{\left(1 + y^{\theta_2}\right)^{k+2}},$$

so that $I_{j0} = I_j$ and let

$$L_{pq} = I_{p+1,1} I_{q+1,1} I_{02} - \left(I_{p1} I_{q+1,1} + I_{p+1,1} I_{q1}\right) I_{12} + I_{p1} I_{q1} I_{22}.$$

Then (see Appendix A)

$$n^{1/(2\theta_2)} u(\tilde{\theta}^*) \tag{2.5}$$
$$\to_{\mathcal{L}} N\left(\mathbf{0}, \frac{(\theta_1/2)^{-1/\theta_2}}{\left(I_{01} I_{21} - I_{11}^2\right)^2} \begin{pmatrix} \beta^2 L_{00} + 2\beta L_{10} + L_{11} & \beta L_{00} + L_{10} \\ \beta L_{00} + L_{10} & L_{00} \end{pmatrix}\right),$$

which we can compare to the analogous result for the MML estimates given by (2.3). The rates of convergence are the same in each case but the constants are different. Some comparisons of the covariance matrices (2.3) and (2.5) are given in Table 1. We see that as θ_2 increases, so does the relative superiority of MML to GCV. Even for $\theta_2 = 4$, which corresponds to f behaving locally like integrated Brownian motion, the asymptotic variance of $\tilde{\theta}_2^*$ is roughly 30 times that of $\hat{\theta}_2$.

3. Predicting with estimated parameters. Let us first consider predicting $f(0)$ based on the observations $\mathbf{Y}_n$ as defined in the previous section in the case where θ is known. The best linear predictor of $f(0)$ based on $\mathbf{Y}_n$ is

$$\hat{f}_b(0) = w_n^{-1}\mu + n^{-1/2} \sum_{j=1}^{n} \left(1 - w_j^{-1}\right) Z_j. \tag{3.1}$$

The best linear unbiased predictor is $\hat{f}(0) = \hat{f}_b(0) + w_n^{-1}(n^{-1/2} Z_n - \mu)$. By symmetry, the mean square error of the best linear unbiased predictor of

$f(j/n)$ is the same for each j, so the following result, proven in Appendix B, applies to $f(j/n)$ for any integer $0 \le j \le n$.

PROPOSITION 3.1.

$$\text{(3.2)}\quad \operatorname{var}\bigl(f(0) - \hat{f}(0)\bigr) = \frac{\pi\theta_2\theta_1^{1/\theta_2}}{\theta_2 \sin(\pi/\theta_2)}(n/2)^{1/\theta_2 - 1} + O(n^{1-\theta_2} + n^{-1-1/\theta_2}).$$

When θ_2 is an even integer, (3.2) agrees with (4.7) in Stein (1990).

We next consider the effect of using estimates of θ on prediction. Let $\hat{f}(0;\theta')$ be the predictor of $f(0)$ obtained using the estimate θ' in place of θ in the expression for $\hat{f}(0)$. Then

$$\hat{f}(0;\theta') - f(0) = \bigl\{\hat{f}(0) - f(0)\bigr\} + \bigl\{\hat{f}(0;\theta') - \hat{f}(0)\bigr\}$$

and the two terms on the right-hand side are exactly independent when θ' is $\hat{\theta}$, $\tilde{\theta}$ or their approximations, which follows from the independence of $\hat{f}(0) - f(0)$ and $Z_1, \ldots, Z_{n-1}$. Using a Taylor series in θ to approximate $\hat{f}(0;\theta') - \hat{f}(0)$, it follows (see Appendix B) that

$$\text{(3.3)}\quad \frac{E\bigl(\hat{f}(0;\hat{\theta}^*) - f(0)\bigr)^2}{E\bigl(\hat{f}(0;\theta) - f(0)\bigr)^2} = 1 + \gamma_M(\theta_2)\left(\frac{2}{\theta_1 n}\right)^{1/\theta_2} + o(n^{-1/\theta_2})$$

and

$$\text{(3.4)}\quad \frac{E\bigl(\hat{f}(0;\tilde{\theta}^*) - f(0)\bigr)^2}{E\bigl(\hat{f}(0;\theta) - f(0)\bigr)^2} = 1 + \gamma_C(\theta_2)\left(\frac{2}{\theta_1 n}\right)^{1/\theta_2} + o(n^{-1/\theta_2}),$$

where

$$\gamma_M(\theta_2) = \frac{\theta_2 \sin(\pi/\theta_2)(I_2 I_{01} - 2I_1 I_{11} + I_0 I_{21})}{\pi(I_0 I_2 - I_1^2)}$$

and

$$\gamma_C(\theta_2) = \frac{\theta_2 \sin(\pi/\theta_2)(L_{11} I_{01} - 2L_{10} I_{11} + L_{00} I_{21})}{\pi(I_{11}^2 - I_{21} I_{01})^2}.$$

While the relative increase in mean square error is $O(n^{-1/\theta_2})$ in each case, the constant multiplying this term is much larger for the approximate GCV estimate when θ_2 is not near 1 (see the last two columns of Table 1).

4. Deterministic regression functions. In this section and the next, we will consider estimating a fixed function f on $[0, 1]$ with absolutely convergent Fourier series of the form

$$\text{(4.1)}\quad f(x) = a_0 + \sum_{k=1}^{\infty} \{a_k \cos(2\pi kx) + b_k \sin(2\pi kx)\}.$$

Let $\mathbf{f}_n = (f(0), f(1/n), \ldots, f((n-1)/n))'$ and $\mathbf{g}_n = (g_1, \ldots, g_n)' = S\mathbf{f}_n$. By

straightforward calculation,

$$g_j = \begin{cases} (n/2)^{1/2} \sum_{k=0}^{\infty} (a_{kn+j} + a_{(k+1)n-j}), & \text{for } j \le [n/2], \\ (n/2)^{1/2} \sum_{k=0}^{\infty} (b_{kn+j} + b_{(k+1)n-j}), & \text{for } [n/2] < j \le n-1, \\ n^{1/2} \sum_{k=0}^{\infty} a_{kn}, & \text{for } j = n. \end{cases} \tag{4.2}$$

Defining $\mathbf{Y}_n$ and $\mathbf{Z}_n$ as in Section 2 (except here f is fixed), we have $Z_j = g_j + e_j$, where the e_j's are iid $N(0, \theta_3)$. Let $A(\theta)\mathbf{Y}_n$ be the best linear unbiased predictor of $\mathbf{f}_n$ as a function of θ under the stochastic model used in Section 2. Taking expectations over the distribution of the e_j's, the expected average squared error (EASE) of this estimator of $\mathbf{f}_n$ is

$$\begin{aligned} \text{EASE}(\theta) &= \frac{1}{n} E\|\mathbf{f}_n - \mathbf{Y}_n\|^2 \\ &= \frac{1}{n} \sum_{j=1}^{n-1} \left(\frac{g_j}{w_j(\theta_1, \theta_2)} \right)^2 + \frac{\theta_3}{n} \sum_{j=1}^{n-1} \left(1 - \frac{1}{w_j(\theta_1, \theta_2)} \right)^2 + \frac{\theta_3}{n}. \end{aligned} \tag{4.3}$$

For some $c > 0$, suppose

$$\lim_{m \to \infty} \sup_{p > rm} \left| \frac{\sum_{j=m}^{m+p} \left(a_j^2 + b_j^2 - 2c\theta_3 j^{-\nu} \right)}{\sum_{j=m}^{m+p} j^{-\nu}} \right| = 0 \tag{4.4}$$

for all $r > 0$. Note that the assumption that f has an absolutely convergent Fourier series implies that $\nu > 2$. The following proposition describes the asymptotic behavior of EASE(θ) under (4.4) (proof available from the author).

PROPOSITION 4.1. *Suppose* (4.4) *is satisfied with* $\nu > 2$ *and* $\theta_2 > 1$ *and* $a_j^2 + b_j^2 < Cj^{-\nu}$ *for all* j *and some constant* C. *Let*

$$R_\nu(\theta_2) = \left[\frac{(\theta_2 - \nu + 1)(\nu - 1)\sin(\pi/\theta_2)}{2(\theta_2 - 1)\sin\{\pi(\nu - 1)/\theta_2\}} \right]^{1/\nu} \frac{2\pi\nu(\theta_2 - 1)}{(\nu - 1)\theta_2^2 \sin(\pi/\theta_2)},$$

$R_\nu(\nu - 1)$ *being defined by continuity. For* $\theta_2 > (\nu - 1)/2$,

$$\inf_{\theta_1} \text{EASE}(\theta) \sim R_\nu(\theta_2)\theta_3 c^{1/\nu} n^{-1+1/\nu}.$$

For $\nu = 2\theta_2 + 1$ *and* $\theta_2 > 1$ (*so* $\nu > 3$),

$$\begin{aligned} \inf_{\theta_1} \text{EASE}(\theta) \sim\ & 2^{2(\nu-3)/\nu(\nu-1)} \left\{ \frac{(\nu - 1)^2 \sin(2\pi/(\nu - 1))}{2\pi(\nu - 3)} \right\}^{-1+1/\nu} \\ & \times \left(\frac{2^{(3-\nu)/(\nu-1)}}{\nu - 1} + 2^{(\nu-3)/(\nu-1)} \right) \theta_3 (c \log n)^{1/\nu} n^{-1+1/\nu}. \end{aligned}$$

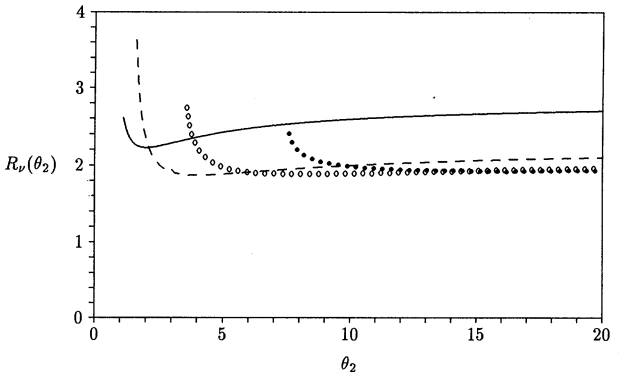

FIG. 1. *Plots of* $R_\nu(\theta_2)$ [*see* (4.3) *for definition*] *for various values of* ν; *solid line for* $\nu = 2$, *dashed line for* $\nu = 4$, *open circles for* $\nu = 8$ *and filled circles for* $\nu = 16$.

For $\nu > 2\theta_2 + 1$ *and* $\theta_2 > 1$,

$$\liminf_{n\to\infty} n^{1-1/(2\theta_2+1)} \inf_{\theta_1} \mathrm{EASE}(\theta) > 0.$$

Furthermore,

$$\inf_{\theta_1,\theta_2} \mathrm{EASE}(\theta) \sim \mathrm{EASE}(c,\nu,\theta_3) \sim \frac{2\pi\theta_3}{\nu \sin(\pi/\nu)} \left(\frac{c}{2}\right)^{1/\nu} n^{-1+1/\nu}. \tag{4.5}$$

Equation (4.4) is closely related to the last equation in Section 3 of Wahba (1985), in which she notes that maximum likelihood appears to work well when this sort of condition is satisfied. The connection between (4.4) and the stochastic model of the previous sections can be seen by noting that under the stochastic model with $\theta_1 = c$ and $\theta_2 = \nu$, $E(a_j^2 + b_j^2) = 2c\theta_3 j^{-\nu}$. Plots of $R_\nu(\theta_2)$ as a function of θ_2 for various values of ν are given in Figure 1. We see that it is possible to use values of θ_2 very different than ν and still obtain estimators that are nearly optimal within the class $A(\theta)\mathbf{Y}_n$. Since the GCV statistic is supposed to approximate $\mathrm{ASE}(\theta)$, it is apparent that GCV will not provide much information about ν. However, if we are only interested in estimating $\mathbf{f}_n$ with an estimate of the form $A(\theta)\mathbf{Y}_n$, then Figure 1 suggests that using θ_2 near ν is not essential as long as a good value of θ_1 for the particular θ_2 selected is used. Hall and Marron (1988) give similar results for the density estimation problem. Note that the right-hand sides of (4.5) and (3.2) are the same if we identify θ_1 with c and θ_2 with ν.

5. Predicting a deterministic regression function with estimated parameters. For fixed regression functions with algebraically decaying Fourier coefficients, we describe results analogous to those in Sections 2 and 3 for Gaussian regression functions. Suppose

$$a_j^2 + b_j^2 = 2c\theta_3 j^{-\nu} + O(j^{-\eta}) \tag{5.1}$$

for some $\eta > \nu + 1/2$. This condition is surely much stronger than necessary to obtain the results in this section; the point here is to show that there is some class of fixed regression functions for which results similar to those for stochastic regression functions can be obtained. Throughout this section, all integrals, sums and functions, such as I_j, that implicitly depended on (θ_1, θ_2) in Sections 2 and 3 will now depend on (c, ν). In particular, now define $u(t)$ by

$$u_1(t) = \frac{t_1 - c}{c} - \log n \frac{t_2 - \nu}{\nu}$$

and $u_2(t) = t_2 - \nu$. The analogous result to (2.3) is

$$n^{1/(2\nu)} u(\hat{\theta}^*) \to_{\mathcal{L}} N\left(\mathbf{0}, \frac{(c/2)^{-1/\nu}}{(I_1^2 - I_0 I_2)^2} \begin{pmatrix} \beta^2 M_{002} + 2\beta M_{102} + M_{112} & \beta M_{002} + M_{102} \\ \beta M_{002} + M_{102} & M_{002} \end{pmatrix}\right), \tag{5.2}$$

where $\beta = \nu^{-1} \log(c/2)$,

$$M_{pqr} = I_{p+1,1} I_{q+1,1} J_{0r} - (I_{p1} I_{q+1,1} + I_{p+1,1} I_{q1}) J_{1r} + I_{p1} I_{q1} J_{2r}$$

and

$$J_{jk} = \int_0^\infty \frac{y^{k\nu} \log^j y}{(1 + y^\nu)^{k+3}} \, dy.$$

For the approximate GCV estimates,

$$n^{1/(2\nu)} u(\tilde{\theta}^*) \to_{\mathcal{L}} N\left(\mathbf{0}, \frac{(c/2)^{-1/\nu}}{(I_{11}^2 - I_{01} I_{21})^2} \begin{pmatrix} \beta^2 M_{003} + 2\beta M_{103} + M_{113} & \beta M_{003} + M_{103} \\ \beta M_{003} + M_{103} & M_{003} \end{pmatrix}\right). \tag{5.3}$$

Proofs are essentially the same as for (2.5). We see that we have the same rates of convergence as in the stochastic case. Some values for the asymptotic covariances in (5.2) and (5.3) are given in Table 2. Not surprisingly, the estimates of ν and c in this setting are considerably less variable than those of

TABLE 2

Constants for comparing asymptotic properties of MML and GCV estimates of c and ν for fixed regression functions satisfying (4.4)

ν	M_{002}/d_M^2	M_{102}/d_M^2	M_{112}/d_M^2	M_{003}/d_C^2	M_{103}/d_C^2	M_{113}/d_C^2	δ_M	δ_C
2	0.0592	−0.131	0.363	0.536	−0.204	0.477	0.0488	0.0910
4	0.0581	−0.132	0.310	4.10	−1.57	1.10	0.0269	0.121
6	0.0457	−0.101	0.226	12.8	−3.79	1.73	0.0144	0.124
8	0.0367	−0.0793	0.173	28.8	−6.83	2.34	0.00875	0.124
10	0.0304	−0.0649	0.139	54.3	−10.7	2.94	0.00584	0.124
20	0.0161	−0.0333	0.0691	401.	−42.5	5.92	0.00157	0.122

See (5.2) and (5.3) for the relationship between columns 2–7 and the asymptotic covariance matrix of the estimates. See (5.4) and (5.5) for the definitions of δ_M and δ_C and Table 1 for the definitions of d_M and d_C.

θ_1 and θ_2 in the stochastic setting given in Table 1. Less obvious is that now the relative asymptotic advantage of $\hat{\theta}^*$ over $\tilde{\theta}^*$, in terms of estimating ν and c, is even greater than it was in the stochastic setting. Analogous to (3.3) and (3.4), we have

$$\frac{\text{EASE}(\hat{\theta}^*)}{\text{EASE}(c, \nu, \theta_3)} = 1 + \delta_M(\nu)\left(\frac{2}{cn}\right)^{1/\nu} + o(n^{-1/\nu}) \tag{5.4}$$

and

$$\times$$

$$\frac{\text{EASE}(\tilde{\theta}^*)}{\text{EASE}(c, \nu, \theta_3)} = 1 + \delta_C(\nu)\left(\frac{2}{cn}\right)^{1/\nu} + o(n^{-1/\nu}), \tag{5.5}$$

where

$$\delta_M(\nu) = \frac{\nu \sin(\pi/\nu)(M_{112}I_{01} - 2M_{102}I_{11} + M_{002}I_{21})}{\pi(I_1^2 - I_0 I_2)^2}$$

and

$$\times$$

$$\delta_C(\nu) = \frac{\nu \sin(\pi/\nu)(M_{113}I_{01} - 2M_{103}I_{11} + M_{003}I_{21})}{\pi(I_{11}^2 - I_{01}I_{21})^2}.$$

Appendix C outlines a proof of (5.5); the proof of (5.4) is similar. The last two columns of Table 2 show that $\delta_M(\nu)$ is considerably smaller than $\delta_C(\nu)$ when ν is large.

6. Simulations.

6.1. *Summary.* This section reports the results of simulations on three fixed regression functions that are identical except for their leading Fourier coefficient. Since the asymptotic results from the earlier sections do not distinguish between these three functions, it is instructive to see how the results for these functions differ for various sample sizes. For a fixed regression functions f satisfying (5.1) exactly for $j \geq 1$, MML yields clearly superior estimates of f. By setting $a_1^2 + b_1^2 = 0$, the small sample behavior of the MML estimates of f changes substantially, so that GCV yields better estimates of f for quite large sample sizes. In contrast, making $a_1^2 + b_1^2$ very large does not seriously effect the estimates of f based on the MML.

6.2. *Computational issues.* All computations reported on here were done in double precision FORTRAN on a SUN Sparcstation 1.

The MML estimate was computed by profiling the likelihood based on $Z_1, \ldots, Z_{n-1}$ over θ_3 and then maximizing the profiled log likelihood with respect to θ_1 and θ_2. This profiled log likelihood and the GCV criterion function were optimized using a double precision version of POWELL [Press, Flannery, Teukolsky and Vetterling (1986), page 299], a modified conjugate direction set method for minimizing a function of several variables. It is helpful to use $\log(1 + \theta_1)$ and θ_2 for the input variables to POWELL, especially when minimizing the GCV criterion function, as the contours of the GCV criterion function tend to be closer to ellipsoidal in this coordinate system. In

some cases, the estimates of θ_1 or θ_2 or both will tend to infinity. I put artificial upper bounds on θ_1 and θ_2 of 10^{19} and 50, respectively, to avoid computational problems. While these bounds are arbitrary, they appear to have essentially no effect on the estimates of the regression function.

The minimum of the GCV criterion function is often difficult to determine because there is commonly a curve of values of θ_1 and θ_2 giving very close to the minimum value of the GCV criterion function, which is not surprising in light of the results in Section 4. Fortunately, while $\tilde{\theta}_1$ and $\tilde{\theta}_2$ might sometimes be difficult to determine, the resulting predictions are rather insensitive to changes in $\tilde{\theta}_1$ and $\tilde{\theta}_2$ along the curve for which the GCV criterion function is nearly minimized. While the MML estimates are fairly insensitive to the choice of starting values for θ, the calculated GCV estimates do sometimes depend somewhat on the starting values, especially for $n = 100$ and f_0. These simulations used multiple starting values to minimize the effect of the starting values on the final estimates. Still, there is no guarantee that the calculated estimates are always the global optima of the criterion functions.

6.3. *Results.* We consider the following three fixed regression functions:

$$f_a(x) = 3\left\{a\cos(2\pi x) + \sum_{m=2}^{\infty} m^{-2}\cos(2\pi mx)\right\}$$

for $a = 0, 1, 2$. We have $f_1(x) = 3\pi^2(x^2 - x + 1/6)$. These functions are infinitely differentiable on $(0, 1)$ and satisfy $f(0) = f(1)$ but $f'(0) \neq f'(1)$. They all satisfy (5.1) with $\nu = 4$ and $c = 4.5$. One hundred simulations were run for each of these functions with $n = 100, 400, 1600$. The same $\mathbf{e}_n$'s were used for each function to facilitate comparisons between functions. The results of the simulations are summarized in Table 3.

Since all three functions satisfy (5.1), MML should yield better estimates of the regression function than GCV for sufficiently large n. The superiority of

TABLE 3
Average squared errors for fixed regression functions

		Median ASE		Mean ASE			
Function	n	**MML**	**GCV**	**MML**	**GCV**	**MML vs GCV***	**median($\hat{\theta}_2$)**
f_0	100	0.1541	0.1087	0.2939	0.2183	33	1.72
f_1	100	0.0917	0.1004	0.1003	0.1394	78	4.20
f_2	100	0.0951	0.0997	0.1019	0.1276	63	5.59
f_0	400	0.0345	0.0352	0.0364	0.0376	54	2.69
f_1	400	0.0315	0.0350	0.0332	0.0365	81	3.96
f_2	400	0.0318	0.0352	0.0334	0.0365	69	4.84
f_0	1600	0.0117	0.0125	0.0121	0.0131	70	3.18
f_1	1600	0.0113	0.0124	0.0117	0.0130	83	3.98
f_2	1600	0.0115	0.0125	0.0118	0.0129	74	4.69

*The number of times out of 100 that $\text{ASE}(\hat{\theta}) < \text{ASE}(\tilde{\theta})$.

MML over GCV for estimating f_1 is apparent even for $n = 100$, in agreement with the asymptotic results of Section 5. In contrast, MML does worse than GCV for f_0 when $n = 100$, about the same as GCV when $n = 400$, and only for $n = 1600$ is MML clearly superior. For f_2, MML does nearly as well as for f_1 even though both f_0 and f_2 have the same absolute difference from f_1. The results for GCV are much less sensitive to the differences between the three functions.

The trouble MML has in estimating f_0 is directly traceable to its tendency to choose small values of θ_2. Based on (5.1), it is reasonable to define the "true" value of the parameters θ_1 and θ_2 as $\theta_1 = 4.5$ and $\theta_2 = 4$ for all three functions. The last column in Table 3 shows that MML tends to underestimate θ_2 for f_0 and overestimate θ_2 for f_2. Considering the results in Section 4, it is not surprising that underestimating θ_2 is a greater problem than overestimating it.

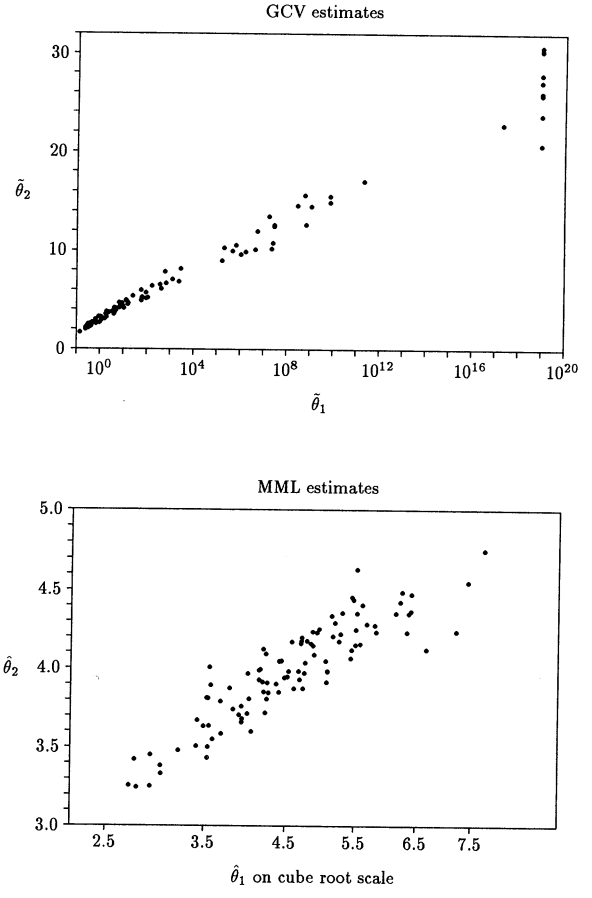

FIG. 2. *Estimates of θ_1 and θ_2 for f_1 with $n = 1600$.*

Let us now briefly consider the behavior of the estimates of θ. Figure 2 plots the MML and GCV estimates of θ_1 and θ_2 when the regression function is f_1 and $n = 1600$. $\hat{\theta}_1$ is plotted on a cube root scale, which helps to normalize its distribution. The GCV estimate of θ is very dispersed; there are many cases for which the estimates are only constrained by the artificial upper bound of 10^{19} I have imposed on $\tilde{\theta}_1$. These very large values of θ_2 do not lead to bad estimates of the regression function, nor does changing the bound have a substantial impact on the estimates. For f_0 and f_2, the GCV estimates tend not to change dramatically, while the MML estimates of θ are, as already noted, substantially biased for f_0 and f_2 even when $n = 1600$.

APPENDICES

Appendix A. Proof of the results from Section 2.

PROOF OF PROPOSITION 2.1. From (2.5) in Mardia and Marshall (1984), for $\lambda_j(\theta) = \theta_3 w_j(\theta_1, \theta_2)$,

$$\mathscr{I}_{kl} = \frac{1}{2}\sum_{j=1}^{n-1}\left(\frac{\partial}{\partial\theta_k}\lambda_j(\theta)\right)\left(\frac{\partial}{\partial\theta_l}\lambda_j(\theta)\right)/\lambda_j(\theta)^2.$$

So, for example,

$$\mathscr{I}_{11} = \tfrac{1}{8}n^{2-2\theta_2}\sum_{j=1}^{n-1} H(\theta_2, j/n)^2/w_j(\theta_1, \theta_2)^2$$

and $\times$

$$\mathscr{I}_{22} = \tfrac{1}{8}\theta_1^2 n^{2-2\theta_2}\sum_{j=1}^{n-1}\{\log n\, H(\theta_2, j/n) + J(\theta_2, j/n)\}^2/w_j(\theta_1, \theta_2)^2,$$

where

$$J(\theta_2, x) = \sum_{p=-\infty}^{\infty} |p + x|^{-\theta_2}\log|p + x|,$$

and $0^{-\theta_2}\log 0$ is taken to be 0. Consider $\mathscr{I}_{11}$. Using $H(\theta_2, x) = H(\theta_2, 1 - x)$, for $\alpha < 1$ and $\alpha\theta_2 > 1$,

$$\left|\sum_{j=1}^{n-1}\left\{\frac{H(\theta_2, j/n)}{w_j(\theta_1, \theta_2)}\right\}^2 - 2\sum_{j=1}^{\lfloor n^\alpha\rfloor}\left\{\frac{H(\theta_2, j/n)}{w_j(\theta_1, \theta_2)}\right\}^2\right| \le 2\sum_{j=\lfloor n^\alpha\rfloor}^{\lfloor (n+1)/2\rfloor}\left\{\frac{H(\theta_2, j/n)}{w_j(\theta_1, \theta_2)}\right\}^2.$$

For $0 < x \le 1/2$, there exists a constant C (depending on θ_2) such that $0 < H(\theta_2, x) < Cx^{-\theta_2}$, and it follows that

$$\sum_{j=\lfloor n^\alpha\rfloor}^{\lfloor (n+1)/2\rfloor}\left\{\frac{H(\theta_2, j/n)}{w_j(\theta_1, \theta_2)}\right\}^2 = O(n^{\alpha+2\theta_2-2\alpha\theta_2}).$$

Furthermore, for $j \le \lfloor n^\alpha \rfloor$, $H(\theta_2, j/n) = (n/j)^{\theta_2} + O(1)$, so

$$\sum_{j=1}^{\lfloor n^\alpha \rfloor} \left\{ \frac{H(\theta_2, j/n)}{w_j(\theta_1, \theta_2)} \right\}^2 = \left[\sum_{j=1}^{\lfloor n^\alpha \rfloor} \left\{ (j/n)^{\theta_2} + \frac{1}{2}\theta_1 n^{1-\theta_2} \right\}^{-2} \right] \{1 + O(n^{(\alpha-1)\theta_2})\}$$

and

$$\begin{aligned} 0 &\le n\int_0^\infty \left(x^{\theta_2} + \tfrac{1}{2}\theta_1 n^{1-\theta_2}\right)^{-2} dx - \sum_{j=1}^{\lfloor n^\alpha \rfloor} \left\{(j/n)^{\theta_2} + \tfrac{1}{2}\theta_1 n^{1-\theta_2}\right\}^{-2} \\ &\le n\int_0^{1/n} \left(x^{\theta_2} + \tfrac{1}{2}\theta_1 n^{1-\theta_2}\right)^{-2} dx + n\int_{\lfloor n^\alpha \rfloor/n}^\infty \left(x^{\theta_2} + \tfrac{1}{2}\theta_1 n^{1-\theta_2}\right)^{-2} dx \\ &= O(n^{-2+2\theta_2} + n^{\alpha - 2\alpha\theta_2 + 2\theta_2}). \end{aligned}$$

Finally,

$$n\int_0^\infty \left(x^{\theta_2} + \tfrac{1}{2}\theta_1 n^{1-\theta_2}\right)^{-2} dx = n^{2\theta_2 - 2 + 1/\theta_2} (\theta_1/2)^{1/\theta_2 - 2} I_0.$$

Thus,

$$\begin{aligned} \mathscr{I}_{11} &= \tfrac{1}{4} n^{1/\theta_2} (\theta_1/2)^{1/\theta_2 - 2} I_0 (1 + O(n^{(\alpha-1)\theta_2})) + O(n^{2+\alpha-2\alpha\theta_2} + 1) \\ &= \tfrac{1}{4} n^{1/\theta_2} (\theta_1/2)^{1/\theta_2 - 2} I_0 \{1 + O(n^{(\alpha-1)\theta_2} + n^{2+\alpha-2\alpha\theta_2 - 1/\theta_2} + n^{-1/\theta_2})\}. \end{aligned}$$

Setting $\alpha = (2 + \theta_2 - 1/\theta_2)/(3\theta_2 - 1)$ yields (2.2). The other elements of $\mathscr{I}$ can be handled similarly. Explicit expressions for the I_j's can be obtained using (3.241.4), (4.252.4) and (4.261.14) from Gradshteyn and Ryzhik (1980), respectively. □

PROOF OF (2.5). Differentiate (2.4) with respect to θ_1 and θ_2, set the resulting equations equal to 0, and rearrange terms to obtain

$$\begin{aligned} &\sum_{j=1}^{n-1} w_j(\tilde\theta_1, \tilde\theta_2)^{-1} \sum_{j=1}^{n-1} \left\{ Z_j^2 H(\tilde\theta_2, j/n) w_j(\tilde\theta_1, \tilde\theta_2)^{-3} \right\} \\ &\quad - \sum_{j=1}^{n-1} \left\{ H(\tilde\theta_2, j/n) w_j(\tilde\theta_1, \tilde\theta_2)^{-2} \right\} \sum_{j=1}^{n-1} \left\{ Z_j^2 w_j(\tilde\theta_1, \tilde\theta_2)^{-2} \right\} = 0 \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} &\sum_{j=1}^{n-1} w_j(\tilde\theta_1, \tilde\theta_2)^{-1} \sum_{j=1}^{n-1} \left\{ Z_j^2 J(\tilde\theta_2, j/n) w_j(\tilde\theta_1, \tilde\theta_2)^{-3} \right\} \\ &\quad - \sum_{j=1}^{n-1} \left\{ J(\tilde\theta_2, j/n) w_j(\tilde\theta_1, \tilde\theta_2)^{-2} \right\} \sum_{j=1}^{n-1} \left\{ Z_j^2 w_j(\tilde\theta_1, \tilde\theta_2)^{-2} \right\} = 0. \end{aligned} \tag{A.2}$$

Since the distribution of $(\tilde\theta_1, \tilde\theta_2)$ obviously does not depend on θ_3, let us take

$\theta_3 = 1$ for convenience. Taking first-order Taylor series about θ_1 and θ_2, $H(\tilde{\theta}_2, j/n) \approx H_j - u_2(\tilde{\theta})J_j$, $J(\tilde{\theta}_2, j/n) \approx J_j - u_2(\tilde{\theta})M_j$ and

$$w_j\left(\tilde{\theta}_1, \tilde{\theta}_2\right)^{-p} \approx w_j^{-p}\left[1 - \frac{p\theta_1 n^{1-\theta_2} H_j}{2w_j}\left\{u_1(\tilde{\theta}) - u_2(\tilde{\theta})(\alpha \log n + J_j/H_j)\right\}\right],$$

where $\alpha = 1 - 1/\theta_2$, $w_j = w_j(\theta_1, \theta_2)$, $H_j = H(\theta_2, j/n)$, $J_j = J(\theta_2, j/n)$ and $M_j = M(\theta_2, j/n)$, where

$$M(\theta_2, x) = \sum_{p=-\infty}^{\infty} |p + x|^{-\theta_2} \log^2|p + x|.$$

Define

$$S_{abcde} = \sum_{j=1}^{n-1} w_j^{-a} H_j^b J_j^c Z_j^{2d} M_j^e,$$

where a, b, c, d and e are integers, and trailing zeros are omitted, so $S_1 = S_{10000}$, for example. Using the above approximations to linearize (A.1) and (A.2) yields $u(\tilde{\theta}) \approx A^{-1}\mathbf{b}$, where $A = (a_{ij})$ is given by

$$a_{11} = \theta_1 n^{1-\theta_2}\left(\tfrac{3}{2}S_1S_{4201} - \tfrac{1}{2}S_{21}S_{3101} - S_{32}S_{2001}\right),$$

$$\begin{aligned} a_{12} = \log n\{&\alpha\theta_1 n^{1-\theta_2}\left(-\tfrac{3}{2}S_1S_{4201} + \tfrac{1}{2}S_{21}S_{3101} + S_{32}S_{2001}\right) \\ &+ S_1S_{3101} - S_{21}S_{2001}\} \\ &+ \theta_1 n^{1-\theta_2}\left(S_{311}S_{2001} - \tfrac{3}{2}S_1S_{4111} - \tfrac{1}{2}S_{201}S_{3101} + S_{21}S_{3011}\right) \\ &+ S_1S_{3011} - S_{201}S_{2001}, \end{aligned}$$

$$a_{21} = \theta_1 n^{1-\theta_2}\left(\tfrac{3}{2}S_1S_{4111} + \tfrac{1}{2}S_{21}S_{3011} - S_{201}S_{3101} - S_{311}S_{2001}\right),$$

$$\begin{aligned} a_{22} = \log n\{&\alpha\theta_1 n^{1-\theta_2}\left(S_{201}S_{3101} + S_{311}S_{2001} - \tfrac{3}{2}S_1S_{4111} - \tfrac{1}{2}S_{21}S_{3011}\right) \\ &+ S_1S_{3011} - S_{201}S_{2001}\} \\ &+ \theta_1 n^{1-\theta_2}\left(\tfrac{1}{2}S_{201}S_{3011} + S_{302}S_{2001} - \tfrac{3}{2}S_1S_{4021}\right) + S_1S_{30011} - S_{2001}S_{20001} \end{aligned}$$

and

$$\mathbf{b} = \begin{pmatrix} S_1S_{3101} - S_{21}S_{2001} \\ S_1S_{3011} - S_{201}S_{2001} \end{pmatrix}.$$

The variability in A is small relative to its expected value, so let us define

$$u(\tilde{\theta}^*) = \{E(A)\}^{-1}\mathbf{b}. \tag{A.3}$$

The usual approach to deriving the asymptotic distribution of $u(\tilde{\theta})$ would be to obtain the asymptotic distribution of $u(\tilde{\theta}^*)$ and to show that $u(\tilde{\theta}) \approx u(\tilde{\theta}^*)$ in an appropriate sense. This second step is normally quite difficult and will be left unproven.

To derive the asymptotic distribution of $u(\tilde{\theta}^*)$, note that for a a positive integer,

$$S_a = n + O(n^{1/\theta_2}), \tag{A.4}$$

and for a, b, c and e nonnegative integers satisfying $0 < b + c + e \leq a$, there exists $\varepsilon > 0$ such that

$$\begin{aligned} S_{abc0e} &= 2n^{1/\theta_2+(\theta_2-1)(b+c+e)}\left(\frac{1}{2}\theta_1\right)^{-b-c-e+1/\theta_2}(1 + O(n^{-\varepsilon})) \\ &\quad \times \int_0^\infty \frac{y^{(a-b-c-e)\theta_2}(-\alpha \log n + \beta + \log y)^{c+2e}\,dy}{(1+y^{\theta_2})^a}. \end{aligned} \tag{A.5}$$

Using (A.4) and (A.5), it can be shown that

$$E(A) = \tfrac{1}{2}\theta_1 n^{2-\theta_2}\begin{bmatrix} S_{32} & -S_{32}\alpha \log n - S_{311} \\ S_{311} & -S_{311}\alpha \log n - S_{302} \end{bmatrix}(1 + O(n^{-\varepsilon}))$$

for some $\varepsilon > 0$. Then from (A.5) we obtain

$$\det(E(A)) = 4n^{4\theta_2-4+2/\theta_2}\left(\tfrac{1}{2}\theta_1\right)^{2+2/\theta_2}(I_{11}^2 - I_{01}I_{21})(1 + O(n^{-\varepsilon})).$$

Thus,

$$\begin{aligned} u(\tilde{\theta}^*) &= \tfrac{1}{4}n^{2-3\theta_2-2/\theta_2}\left(\tfrac{1}{2}\theta_1\right)^{-3-2/\theta_2}(I_{11}^2 - I_{01}I_{21})^{-1}(1 + O(n^{-\varepsilon})) \\ &\quad \times \begin{bmatrix} \alpha \log n(S_{21}S_{311}S_{2001} + S_1S_{32}S_{3011} - S_1S_{311}S_{3101} - S_{201}S_{32}S_{2001}) \\ \quad + S_{21}S_{302}S_{2001} + S_1S_{311}S_{3011} - S_1S_{302}S_{3101} - S_{201}S_{311}S_{2001} \\ S_{21}S_{311}S_{2001} + S_1S_{32}S_{3011} - S_1S_{311}S_{3101} - S_{201}S_{32}S_{2001} \end{bmatrix}. \end{aligned}$$

Equation (2.5) follows using standard central limit theory. □

Appendix B. Proof of results from Section 3.

PROOF OF PROPOSITION 3.1. Since $f(0) - \hat{f}_b(0)$ and $\hat{f}_b(0) - \hat{f}(0)$ are uncorrelated,

$$\begin{aligned} \operatorname{var}(f(0) - \hat{f}(0)) &= \operatorname{var}(f(0)) - \operatorname{var}(\hat{f}_b(0)) + \operatorname{var}(\hat{f}_b(0) - \hat{f}(0)) \\ &= \theta_3\theta_1 \sum_{p=1}^{\infty} p^{-\theta_2} - \theta_3 n^{-1} \sum_{j=1}^{n} \frac{(w_j - 1)^2}{w_j} + \frac{\theta_3}{nw_n} \end{aligned}$$

and

$$n^{-1} \sum_{j=1}^{n} \frac{(w_j - 1)^2}{w_j} = 2n^{-1} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{(w_j - 1)^2}{w_j} + O(n^{1-2\theta_2}),$$

so

$$\operatorname{var}\bigl(f(0) - \hat{f}(0)\bigr) = \theta_3\theta_1 \sum_{p=1}^{\infty} p^{-\theta_2} - 2n^{-1}\theta_3 \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{(w_j - 1)^2}{w_j} + \frac{\theta_3}{nw_n} + O(n^{1-2\theta_2})$$

$$= \theta_3 \sum_{j=1}^{\lfloor n/2 \rfloor} \left\{ \theta_1 j^{-\theta_2} - 2n^{-1} \frac{(w_j - 1)^2}{w_j} \right\} + \frac{\theta_3}{n} + O(n^{1-\theta_2}).$$

For $1 \le j \le \lfloor n/2 \rfloor$, $H(\theta_2, j/n) = (n/j)^{\theta_2} + O(1)$, and it follows that

$$\operatorname{var}\bigl(f(0) - \hat{f}(0)\bigr)$$

$$= \theta_3 \sum_{j=1}^{\lfloor n/2 \rfloor} \left[\theta_1 j^{-\theta_2} - \frac{1}{2}\theta_1^2 n^{1-2\theta_2} \frac{\{(n/j)^{\theta_2} + O(1)\}^2}{1 + (1/2)\theta_1 n^{1-\theta_2}\{(n/j)^2 + O(1)\}} \right]$$

$$+ \frac{\theta_3}{n} + O(n^{1-\theta_2})$$

$$= \theta_3 \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{\theta_1 j^{-\theta_2}}{1 + (1/2)\theta_1 n j^{-\theta_2}} (1 + O(n^{1-\theta_2})) + \frac{\theta_3}{n} + O(n^{1-\theta_2}).$$

Furthermore, using the Euler–Maclaurin formula [Abramowitz and Stegun (1965), taking $n = 1$ in 23.1.30], it can be shown that

$$\sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{j^{\theta_2} + (1/2)\theta_1 n} = \int_1^{\lfloor n/2 \rfloor} \frac{dx}{x^{\theta_2} + (1/2)\theta_1 n} + \frac{1}{\theta_1 n} + O(n^{-\theta_2} + n^{-1-1/\theta_2}),$$

so that

$$\sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{j^{\theta_2} + (1/2)\theta_1 n}$$

$$= \int_0^{\infty} \frac{dx}{x^{\theta_2} + (1/2)\theta_1 n} + \frac{1}{\theta_1 n} - \int_0^1 \frac{dx}{x^{\theta_2} + (1/2)\theta_1 n}$$

$$- \int_{\lfloor n/2 \rfloor}^{\infty} \frac{dx}{x^{\theta 2} + (1/2)\theta_1 n} + O(n^{-1-1/\theta_2} + n^{-\theta_2})$$

$$= \frac{1}{\theta_2} \left(\frac{1}{2}\theta_1 n \right)^{1/\theta_2 - 1} \frac{\pi}{\sin(\pi/\theta_2)} - \frac{1}{\theta_1 n} + O(n^{1-\theta_2} + n^{-1-1/\theta_2}).$$

Proposition 3.1 follows. □

PROOF OF (3.3). We merely outline the proof as the technical details are similar to those leading to (5.5) (see Appendix C). The proof of (3.4) is

essentially the same. Taking a first-order Taylor series in θ,

$$\hat{f}(0;\hat{\theta}^*) - \hat{f}(0;\theta) = \tfrac{1}{2}\theta_1 n^{1/2-\theta_2} u(\hat{\theta}) S_{2101} - \tfrac{1}{2}\theta_1 n^{1/2-\theta_2}(\hat{\theta}_2 - \theta_2)(S_{2011} + \alpha \log n S_{2101})$$

plus a remainder whose second moment is $o(n^{-1})$. By lengthy but straightforward calculations, we obtain

$$E\big(\hat{f}(0;\hat{\theta}^*) - \hat{f}(0;\theta)\big)^2 = \frac{2\theta_3(I_2 I_{01} - 2I_1 I_{11} + I_0 I_{21})}{n(I_0 I_2 - I_1^2)} + o(n^{-1}).$$

Equation (3.3) follows using (3.2) and the independence of $\hat{f}(0;\hat{\theta}^*) - \hat{f}(0;\theta)$ and $\hat{f}(0;\theta) - f(0)$. □

Appendix C.

PROOF OF (5.5). This Appendix outlines the proof of (5.5), the proof of (5.4) being similar.

First, define $w_j(\theta_1,\theta_2) = 0$ whenever $\theta_1 < 0$ or $\theta_2 \le 1$. If we let $\bar{\theta} = \tilde{\theta}^*$ when $\tilde{\theta}_1^* \ge 0$ and $\tilde{\theta}_2^* > 1$ and $\bar{\theta} = (0,2)$ otherwise, then $\mathrm{EASE}(\tilde{\theta}^*) = \mathrm{EASE}(\bar{\theta})$. Let

$$d_j(\theta) = \frac{g_j - e_j}{w_j(\theta_1,\theta_2)} + e_j,$$

so that

$$\mathrm{ASE}(\theta) = n^{-1}\left(\sum_{j=1}^{n-1} d_j(\theta)^2 + e_n^2\right).$$

By taking a second-order Taylor series in θ about (c,ν) with remainder, it can be shown that $d_j(\theta) = d_{0j} + d_{1j}(\theta) + d_{2j}(\theta) + d_{3j}(\theta)$, where

$$d_{0j} = d_j(c,\nu),$$

$$d_{1j}(\theta) = \frac{cn^{1-\nu}(g_j - e_j)}{2w_j^2}\{-u_1(\theta)H_j + u_2(\theta)(\alpha \log n\, H_j + J_j)\},$$

$$d_{2j}(\theta) = (g_j - e_j)\sum_{k=0}^{2} a_{kj} u_1(\theta)^k u_2(\theta)^{2-k},$$

$$d_{3j}(\theta) = (g_j - e_j)\sum_{k=0}^{3} b_{kj}(\phi_j) u_1(\theta)^k u_2(\theta)^{3-k},$$

and ϕ_j is between (c,ν) and (θ_1,θ_2). Here, we define $w_j = w_j(c,\nu)$ and similarly define H_j and J_j. Furthermore, there exists a constant C such that for all n sufficiently large,

$$|a_{kj}| \le C \log^2 n\, n^{1-\nu} H_j / w_j^2 \tag{C.1}$$

and for any fixed $\alpha > 0$

$$\text{(C.2)} \qquad \sup_{|u(\theta)| \le n^{-\alpha}} \left| b_{kj}(\phi_j) \right| \le C \log^3 n \, n^{1-\nu} H_j / w_j^2.$$

For $\alpha < 1/\nu$, there exists $\beta > 0$ such that $P(|u(\bar{\theta})| > n^{-\alpha}) \le \exp\{-n^{\beta}\}$ for all n sufficiently large using the fact that all normalized quadratic forms in normal random variables have uniform exponential bounds on their tail behavior [Ponomarenko (1978)]. Using this fact and (C.2), it is possible to show that for $\alpha = 1/(2\nu)$, say,

$$E\left\{ \frac{1}{n} \sum_{j=1}^{n-1} d_{3j}(\bar{\theta})^2 \right\} = E\left[\left\{ \frac{1}{n} \sum_{j=1}^{n-1} d_{3j}(\bar{\theta})^2 \right\} \left\{ I_{\{|u(\bar{\theta})| \le n^{-\alpha}\}} + I_{\{|u(\bar{\theta})| > n^{-\alpha}\}} \right\} \right]$$

$$= o(n^{-1-1/\nu}).$$

Using (C.1) and (5.1), it can be shown that

$$\frac{1}{n} E \sum_{j=1}^{n-1} \left\{ d_{2j}(\bar{\theta})^2 + d_{2j}(\bar{\theta}) d_{0j} + d_{1j}(\bar{\theta}) d_{0j} \right\} = o(n^{-1}) \text{ and}$$

$$E\left\{ \frac{1}{n} \sum_{j=1}^{n-1} d_{1j}(\bar{\theta})^2 \right\} = \frac{2\theta_3 (M_{113} I_{01} - 2M_{103} I_{11} + M_{003} I_{21})}{n (I_{11}^2 - I_{01} I_{21})^2} + o(n^{-1}).$$

Equation (5.5) follows using Proposition 4.1. □

Acknowledgments. I would like to thank Mark Handcock and the referees for several helpful comments on both the substance and organization of this paper.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. (1965). *Handbook of Mathematical Functions*. Dover, New York.

BROCKWELL, P. J. and DAVIS, R. A. (1987). *Time Series: Theory and Methods*. Springer, New York.

COGBURN, R. and DAVIS, H. T. (1974). Periodic splines and spectral estimation. *Ann. Statist.* **2** 1108–1126.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of cross-validation. *Numer. Math.* **31** 377–403.

CROWDER, M. J. (1986). On consistency and inconsistency of estimating equations. *Econom. Theory* **2** 305–330.

EUBANK, R. L. (1987). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

GAMBER, H. (1979). Choice of optimal shape parameter when smoothing noisy data. *Comm. Statist. A* **8** 1425–1435.

GRADSHTEYN, I. S. and RYZHIK, I. M. (1980). *Table of Integrals, Series, and Products*. Academic, New York.

HALL, P. and MARRON, J. S. (1988). Choice of kernel order in density estimation. *Ann. Statist.* **16** 161–173.

LI, K. C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112.

MARDIA, K. V. and MARSHALL, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146.

PONOMARENKO, L. S. (1978). Inequalities for distributions of quadratic forms in normal random variables. *Theory Probab. Appl.* **23** 652–656.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1986). *Numerical Recipes*. Cambridge Univ. Press.

RICE, J. and ROSENBLATT, M. (1981). Integrated mean squared error of a smoothing spline. *J. Approx. Theory* **33** 353–369.

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

STEIN, M. L. (1990). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Statist.* **18** 1139–1157.

SWEETING, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.* **8** 1375–1381.

WAHBA, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.* **24** 383–393.

WAHBA, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13** 1378–1402.

WAHBA, G. (1989). *Spline Models for Observational Data*. SIAM, Philadelphia.

WAHBA, G. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review* **8** 1122–1143.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637-1514