The Annals of Statistics
1990, Vol. 18, No. 3, 1139–1157

A COMPARISON OF GENERALIZED CROSS VALIDATION AND MODIFIED MAXIMUM LIKELIHOOD FOR ESTIMATING THE PARAMETERS OF A STOCHASTIC PROCESS[1]

BY MICHAEL L. STEIN

The University of Chicago

Wahba compared the performance of generalized cross validation (GCV) and modified maximum likelihood (MML) procedures for choosing the smoothing parameter of a smoothing spline. This work makes a more careful study of the two procedures when the stochastic model motivating the modified maximum likelihood estimate is correct. In particular, it is shown that in the case of the linear smoothing spline with equally spaced observations, both estimates are asymptotically normal with the GCV estimate having twice the asymptotic variance of the MML estimate. The impact of using these estimates on the subsequent predictions is also calculated. Conjectures on how these results should generalize to higher order smoothing splines are developed. These conjectures suggest that the penalty for using GCV instead of MML when the stochastic model is correct is greater for higher order smoothing splines, both in terms of the efficiency in estimating the smoothing parameter and the impact on subsequent predictions.

1. Introduction. Wahba (1985) has compared the performance of generalized cross validation (GCV) and modified maximum likelihood (MML) (what Wahba calls generalized maximum likelihood) for choosing the smoothing parameter of a smoothing spline. Up to the order of approximation considered in her paper, Wahba (1985) found that in terms of the performance of subsequent predictions, GCV does at least as well as MML in a variety of situations and does much better in some situations. In particular, Wahba (1985) considered the behavior of predictions based on GCV and MML estimates when the Gaussian stochastic model that motivates the MML estimate is true. This circumstance might be considered the most favorable for MML and yet Wahba (1985) showed that, to first order, predictions based on a simplified version of the GCV do asymptotically as well as a similarly simplified version of the MML. A special case of her results is obtained by assuming the underlying process is Brownian motion and observations are taken at equally spaced intervals subject to iid Gaussian errors. The optimal linear predictors form a linear smoothing spline in this case. The covariance structure of the observations is particularly simple and this special structure will be exploited here to obtain the asymptotic distributions of the MML and GCV estimates of the smoothing parameter, which is essentially the ratio of the variance of the

Received March 1989; revised September 1989.

[1] Work supported by a NSF Mathematical Sciences Postdoctoral Research Fellowship. Manuscript prepared using computer facilities supported in part by NSF Grants DMS 86-01732 and DMS 87-03942 to the Department of Statistics at The University of Chicago.

AMS 1980 *subject classifications*. Primary 62M20; secondary 60G15, 62G05.

Key words and phrases. Brownian motion, spline smoothing, kriging.

increments of the Brownian motion to the variance of the observational errors. In particular, in Section 2, for such a process observed at $0, n^{-1}, 2n^{-1}, \dots, 1$, the MML and GCV estimates of the smoothing parameter are shown to be asymptotically normal with variance of order $n^{-1/2}$ instead of the usual n^{-1} and the asymptotic variance of the GCV estimate is twice that of the MML estimate. This result is used to approximate the density and the mean squared error of predictions based on the GCV and MML estimates. Stein (1989) gives rigorous proofs of these approximations when the MML estimate is used. In Section 4, I make conjectures as to how these results should generalize to higher order smoothing splines. For example, when the underlying process is integrated Brownian motion, which corresponds to using the cubic smoothing spline, the asymptotic variance of the GCV estimate of the smoothing parameter is conjectured to be $10/3$ that of the MML estimate.

The linear smoothing spline has been largely ignored in the splining literature, probably because it is not differentiable at the observation locations and thus does not yield a sufficiently smooth regression surface for most problems in nonparametric regression. There are two reasons for studying this model despite this problem. First, it is easy to study and we might hope that the results obtained will give insights into how estimates of the smoothing parameter in the higher order splines should behave. Furthermore, if the independent variable represents time or geographic location rather than the values of a covariate for some population, then this model is not at all inappropriate. Indeed, in the geostatistical literature, covariance structures for mineral deposits that are like the covariance structure of Brownian motion plus noise over at least relatively short distances are perhaps the most frequently used type of model [Journel and Huijbregts (1978), Chapter 4].

2. Estimating the smoothing parameter.

Suppose $W(x)$ is a Gaussian process on $\mathbb{R}$ satisfying $EW(x) = \mu$ and $\gamma(h) = \frac{1}{2}E(W(x+h) - W(x))^2 = \sigma^2|h|$ for all x. The function $\gamma(\cdot)$ is known as the semivariogram in the geostatistical literature [Journel and Huijbregts (1978)]. By describing the covariance structure of $W(\cdot)$ in terms of $\gamma(\cdot)$, the distribution of $W(\cdot)$ is not quite specified, as $\operatorname{var}(W(x))$ is not defined. If we set $\operatorname{var}(W(0)) = 0$, then $\operatorname{cov}(W(x), W(x')) = 2\sigma^2 \min(|x|, |x'|) I_{\{xx' > 0\}}$; that is, $W(\cdot)$ is just Brownian motion with $\operatorname{var}(W(1)) = 2\sigma^2$. However, all of the results in this and the next section depend on the covariance structure only through $\gamma(\cdot)$, so we will leave $\operatorname{var}(W(x))$ undefined.

Suppose we observe

$$y_j = W(\delta j) + \varepsilon_j, \quad \text{for } j = 0, \dots, n, \tag{2.1}$$

where the ε_j's are iid $N(0, \tau^2)$ and independent of $W(\cdot)$ and δ is the distance between neighboring observations. We will mainly consider $\delta = n^{-1}$ in this paper. Taking $\delta = n^{-1}$ would be standard when considering problems in nonparametric regression and $\delta = 1$ would be standard in time series analysis. Based on these observations, the best linear unbiased predictor of $W(x)$ has the form $\sum \lambda_j y_j$, where the mean square prediction error is minimized subject to the unbiasedness constraint $E\Sigma\lambda_j y_j = EW(x)$ for all μ. The unbiasedness constraint is just $\Sigma\lambda_j = 1$, and the mean square error of a linear unbiased predictor can be evaluated in terms of $\gamma(\cdot)$ and τ^2; in the present case it is given by

$$E\left(\sum_{j=0}^{n}\lambda_j y_j - W(x)\right)^2 = \tau^2\sum_{j=0}^{n}\lambda_j^2 + 2\sigma^2\sum_{j=0}^{n}\lambda_j|x-\delta j| - \delta\sigma^2\sum_{jk=0}^{n}\lambda_j\lambda_k|j-k|.$$

The best linear unbiased predictor, or kriging predictor, minimizes this expression subject to the unbiasedness constraint. We see that the minimizing λ_j's depend on σ^2 and τ^2 only through $\theta = \sigma^2/\tau^2$.

This kriging predictor is in fact identical to a linear smoothing spline. In the present setting, the smoothing spline $f(\cdot)$ of degree $2m - 1$ is defined as the minimizer of

$$(2.2)\qquad \frac{1}{n+1}\sum_{j=0}^{n}\left(f(\delta j) - y_j\right)^2 + \lambda\int_0^1\left(f^{(m)}(x)\right)^2 dx.$$

By taking $\lambda = \frac{1}{2}\theta/(n+1)$ and $m = 1$, $f(x)$ is identical to the best linear unbiased predictor defined above [Kimeldorf and Wahba (1970); Wahba and Wendelberger (1980)].

For purposes of obtaining the kriging or splining predictor, the essential parameter that needs to be estimated is θ, or equivalently, λ. Wahba (1985) compares two methods for estimating θ, generalized cross validation (GCV) and what she calls generalized maximum likelihood. Since her generalized maximum likelihood estimate is in fact identical to the modified maximum likelihood (MML) estimate of θ as defined by Patterson and Thompson (1971) in a different context, I will refer to this estimate as the modified maximum likelihood estimate. Denoting these estimates by $\hat{\theta}_{\mathrm{GCV}}$ and $\hat{\theta}_{\mathrm{MML}}$, respectively, in this section both estimates are shown to be asymptotically normal with the asymptotic variance of $\hat{\theta}_{\mathrm{GCV}}$ twice that of $\hat{\theta}_{\mathrm{MML}}$. This result for $\hat{\theta}_{\mathrm{GCV}}$ assumes that its asymptotic distribution can be determined by taking a linear approximation of the equation obtained by setting the derivative of the GCV criterion equal to zero.

Let us consider $\hat{\theta}_{\mathrm{MML}}$. The modified maximum likelihood estimates of θ and τ^2 are determined by maximizing the likelihood of the contrasts of the observations; that is, the largest set of linear combinations of the observations that are linearly independent and have mean zero [Patterson and Thompson (1971)]. The modified maximum likelihood estimates do not depend on the particular definition of the contrasts; for simplicity, we will consider $Z = (y_1 - y_0, \dots, y_n - y_{n-1})'$. By construction the distribution of Z does not depend on μ and the log likelihood of τ^2 and θ given Z is

$$l(\tau^2, \theta; Z) = -\frac{n}{2}\log 2\pi - n\log\tau - \frac{1}{2}\log|T + 2\theta\delta I| - \frac{1}{2\tau^2}Z'(T + 2\theta\delta I)^{-1}Z,$$

where T is the $n \times n$ tridiagonal matrix

$$T = \begin{pmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}.$$

Now T has the eigenvalue decomposition $S'\Lambda S$, where Λ is a diagonal $n \times n$ matrix with ith value $2\{1 - \cos(\pi i/(n+1))\}$ and S is an orthogonal $n \times n$ matrix with ijth element $(2/(n+1))^{1/2} \sin(\pi ij/(n+1))$. Letting $X = SZ$, the log likelihood is

$$(2.3) \qquad \begin{aligned} l(\tau^2, \theta; X) = &-\frac{n}{2}\log 2\pi - n \log \tau - \frac{1}{2}\log|\Lambda + 2\theta\delta I| \\ &- \frac{1}{2\tau^2} X'(\Lambda + 2\theta\delta I)^{-1} X, \end{aligned}$$

where $X \sim N(0, \tau^2(\Lambda + 2\theta\delta I))$. An additional simplification is possible in that $|T + 2\theta\delta I| = |\Lambda + 2\theta\delta I|$ can be explicitly evaluated. Let $U_n(a, b)$ be the $n \times n$ symmetric tridiagonal matrix with a on the main diagonal and b on the off-diagonals, where $a \geq |2b|$. By row reductions,

$$|U_n(a,b)| = a|U_{n-1}(a,b)| - b^2|U_{n-2}(a,b)|.$$

We have a linear difference equation with initial conditions $|U_1(a,b)| = a$ and $|U_2(a,b)| = a^2 - b^2$; it follows that [Strang (1976), page 186]

$$|U_n(a,b)| = \frac{1}{c}\left(\frac{a+c}{2}\right)^{n+1} - \frac{1}{c}\left(\frac{a-c}{2}\right)^{n+1},$$

where $c = (a^2 - 4b^2)^{1/2}$. In the present case, $a = 2 + 2\theta\delta$ and $b = -1$.

Calculating X takes $O(n^2)$ operations, although if $n + 1$ is composite, the fast Fourier transform can be used to reduce the computations. Once $X_1^2, \dots, X_n^2$ has been calculated, the log likelihood given in (2.3) requires $n + O(1)$ multiplications/divisions for each value of θ. Order n algorithms are available for both the modified likelihood function [Kohn and Ansley (1987)] and the GCV criterion function [O'Sullivan (1985) and Hutchinson and de Hoog (1985)] for all positive integers m. Specializing the algorithm of Kohn and Ansley (1987) to the present case, the modified likelihood function takes $4n + O(1)$ multiplications/divisions, which can be reduced to $3n + O(1)$ by using the explicit formula for the determinant given here. Thus, if evaluation of the modified likelihood function for many values of θ is required, it may be preferable to use (2.3) despite the initial effort needed to compute $X_1^2, \dots, X_n^2$.

The modified maximum likelihood estimates of θ and τ^2, denoted by $\hat{\theta}_{\mathrm{MML}}$ and $\hat{\tau}^2_{\mathrm{MML}}$, are given by the solution to

$$0 = \frac{\partial}{\partial\theta} l(\tau^2, \theta; X) = -\frac{1}{n}\mathrm{tr}(\Lambda + 2\theta\delta I)^{-1} + \frac{\delta}{\tau^2} X'(\Lambda + 2\theta\delta I)^{-2} X$$

and

$$0 = \frac{\partial}{\partial \tau^2} l(\tau^2, \theta; X) = -\frac{n}{2\tau^2} + \frac{1}{2\tau^4} X'(\Lambda + 2\theta\delta I)^{-1} X.$$

Wahba (1985), page 1384, notes that her generalized maximum likelihood estimate of θ is in fact the same as that obtained by maximizing the likelihood of the contrasts. The Fisher information matrix for θ and τ^2 based on X is

$$\begin{pmatrix} n/(2\tau^4) & \delta\tau^{-2}\operatorname{tr}(\Lambda + 2\theta I)^{-1} \\ \delta\tau^{-2}\operatorname{tr}(\Lambda + 2\theta I)^{-1} & 2\delta^2 \operatorname{tr}(\Lambda + 2\theta I)^{-2} \end{pmatrix}.$$

If $\delta = n^{-1}$, for fixed $\theta > 0$, as $n \to \infty$,

$$\begin{aligned} \operatorname{tr}(\Lambda + 2\theta n^{-1} I)^{-j} &= \sum_{k=1}^{n} \left\{ 2\left(1 - \cos\frac{\pi k}{n+1} + \frac{\theta}{n}\right)\right\}^{-j} \\ &= \frac{n}{2^j \pi} \int_0^{\pi} \left(1 + \frac{\theta}{n} - \cos y\right)^{-j} dy + O(n^j) \\ &= \frac{n}{2^{2j-1}(2 + n^{-1}\theta)^{j-1}\left(2n^{-1}\theta + (\theta/n)^2\right)^{1/2}} \\ &\quad \times \sum_{k=0}^{j-1} \frac{(2j-2k-3)!!(2k-1)!!}{k!(j-k-1)!} \left(\frac{2 + n^{-1}\theta}{n^{-1}\theta}\right)^k + O(n^j) \\ &= \frac{(2j-3)!!}{2^{2j-1/2}(j-1)!\theta^{j-1/2}} n^{j+1/2} + O(n^j), \end{aligned} \tag{2.4}$$

where $(2j+1)!! = 1 \cdot 3 \cdots (2j+1)$ and $(-1)!! = 1$ [Gradshteyn and Ryzhik (1980), page 383]. Since the components of X are independent, it follows from standard results on maximum likelihood estimates that for $\delta = n^{-1}$, $\hat\theta_{\text{MML}}$ and $\hat\tau^2_{\text{MML}}$ are asymptotically efficient and

$$\begin{pmatrix} n^{1/4}(\hat\theta_{\text{MML}} - \theta) \\ n^{1/2}(\hat\tau^2_{\text{MML}} - \tau^2) \end{pmatrix} \to_{\mathscr{L}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2^{5/2}\theta^{3/2} & 0 \\ 0 & 2\tau^4 \end{pmatrix}\right).$$

It is important to note that $\hat\theta_{\text{MML}} - \theta$ is $O_p(n^{-1/4})$, whereas $\hat\tau^2_{\text{MML}} - \tau^2$ is the more usual $O_p(n^{-1/2})$. it can be shown that the asymptotic covariance matrix of $\hat\sigma^2_{\text{MML}}$ and $\hat\tau^2_{\text{MML}}$ is the same as that for the minimum variance quadratic unbiased invariant estimates of σ^2 and ,τ^2 as found by Stein (1987).

If we instead set $\delta = 1$, by approximating $\operatorname{tr}(\Lambda + 2\theta I)^{-j}$ by an integral as in (2.4), we can obtain

$$n^{1/2}\begin{pmatrix} \hat\theta_{\text{MML}} - \theta \\ \hat\tau^2_{\text{MML}} - \tau^2 \end{pmatrix} \to_{\mathscr{L}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 2\left(1 + \theta - (2\theta + \theta^2)^{1/2}\right)^{-1} \times \begin{pmatrix} (2\theta + \theta^2)^{3/2} & -\tau^2(2\theta + \theta^2) \\ -\tau^2(2\theta + \theta^2) & \tau^4(2 + \theta) \end{pmatrix}\right).$$

Now that the distance between neighboring observations does not depend on n, we get the more familiar asymptotic results in which both $\hat{\theta}_{\mathrm{MML}} - \theta$ and $\hat{\tau}^2_{\mathrm{MML}} - \tau^2$ are $O_p(n^{-1/2})$.

Now let us consider the asymptotic behavior of $\hat{\theta}_{\mathrm{GCV}}$. For the setup given by (2.1), it can be easily shown that the expression $I - A(\lambda)$ as defined by Wahba (1985) is equal to $C'(T + 2\theta\delta I)^{-1}C$, where C is the $n \times (n+1)$ matrix

$$C = \begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix},$$

so that $\hat{\theta}_{\mathrm{GCV}}$ is the minimizer of [see (1.2), Wahba (1985)]

$$\begin{aligned} \frac{\left\| C'(T + 2\theta\delta I)^{-1} Z \right\|^2}{\left[\operatorname{tr} C'(T + 2\theta\delta I)^{-1} C\right]^2} &= \frac{Z'(T + 2\theta\delta I)^{-1} T (T + 2\theta\delta I)^{-1} Z}{\left[\operatorname{tr} T(T + 2\theta\delta I)^{-1}\right]^2} \\ &= \frac{X'(\Lambda + 2\theta\delta I)^{-1} \Lambda (\Lambda + 2\theta\delta I)^{-1} X}{\left[n - 2\theta\delta \operatorname{tr}(\Lambda + 2\theta\delta I)^{-1}\right]^2} \\ &= \frac{X' Q_1(\theta) X}{(\operatorname{tr} Q_0(\theta))^2}, \end{aligned} \tag{2.5}$$

where $Q_j(\theta) = (\Lambda + 2\theta\delta I)^{-j} - 2\theta\delta(\Lambda + 2\theta\delta I)^{-(j+1)}$. Furthermore, $(d/d\theta)Q_j(\theta) = -2(j+1)\delta Q_{j+1}(\theta)$. Thus, any local minimum of (2.5) satisfies

$$X'Q_1(\theta)X \operatorname{tr} Q_1(\theta) - X'Q_2(\theta)X \operatorname{tr} Q_0(\theta) = 0. \tag{2.6}$$

Assuming that the asymptotic distribution of $\hat{\theta}_{\mathrm{GCV}}$ can be obtained by taking a first order Taylor series of (2.6) about the true value of θ, we obtain the approximation

$$\begin{aligned} \hat{\theta}_{\mathrm{GCV}} - \theta &\approx \frac{1}{2\delta} \frac{X'Q_2X \operatorname{tr} Q_0 - X'Q_1X \operatorname{tr} Q_1}{3X'Q_3X \operatorname{tr} Q_0 - X'Q_2X \operatorname{tr} Q_1 - 2X'Q_1X \operatorname{tr} Q_2} \\ &\approx \frac{1}{2\delta\tau^2} \frac{X'Q_2X \operatorname{tr} Q_0 - X'Q_1X \operatorname{tr} Q_1}{\operatorname{tr} Q_2 \operatorname{tr} Q_0 - (\operatorname{tr} Q_1)^2}, \end{aligned} \tag{2.7}$$

where we have suppressed the dependence of Q_j on θ and the second approximation follows by replacing the denominator of the previous line by its expected value. Applying Liapunov's theorem [Chung (1974), page 200] to $X'[Q_2 \operatorname{tr} Q_0 - Q_1 \operatorname{tr} Q_1]X$, which is a sum of independent random variables, it follows from (2.4) and (2.7) that when $\delta = n^{-1}$,

$$n^{1/4}\left(\hat{\theta}_{\mathrm{GCV}} - \theta\right) \to_{\mathscr{L}} N\left(0, 2^{7/2}\theta^{3/2}\right),$$

and when $\delta = 1$,

$$n^{1/2}\left(\hat{\theta}_{\mathrm{GCV}} - \theta\right) \to_{\mathscr{L}} N\left(0, 4(2\theta + \theta^2)^{3/2} / \left\{1 + \theta - (2\theta + \theta^2)^{1/2}\right\}\right).$$

In both cases, the asymptotic variance is twice that of $\hat{\theta}_{\mathrm{MML}}$.

3. Predicting with an estimated parameter. In this section, we consider the effect of using an estimated value of θ on subsequent predictions. As a first step, we give a very accurate approximation to the best linear unbiased predictor by deriving an exact expression for the best linear unbiased predictor in a modified scenario. Consider the situation described by (2.1) except that now j ranges over the integers instead of from just 0 to n. For $x \in [0, \delta)$, the best linear unbiased predictor of $W(x)$ based on this doubly infinite sequence of observations that are δ apart and where $\theta = \sigma^2/\tau^2$ is

$$\begin{aligned}\hat{W}_\infty(x; \delta, \theta) &= \left[\frac{\delta\theta(1 + x\theta)}{\{\delta\theta(2 + \delta\theta)\}^{1/2}} - x\theta\right] \sum_{j=0}^{\infty} a(\delta\theta)^j y_{-j} \\ &\quad + \left[\frac{\delta\theta(1 + (\delta - x)\theta)}{\{\delta\theta(2 + \delta\theta)\}^{1/2}} - (\delta - x)\theta\right] \sum_{j=1}^{\infty} a(\delta\theta)^{j-1} y_j,\end{aligned} \tag{3.1}$$

where $a(t) = 1 + t - \{t(2 + t)\}^{1/2}$. We see that the predictor depends on x, δ and θ only through $\delta\theta$ and $x\theta$, so we can write

$$\hat{W}_\infty(x; \delta, \theta) = \sum_{j=-\infty}^{\infty} b_j(\delta\theta, x\theta) y_j.$$

The optimality of this predictor can be demonstrated by showing that $\hat{W}_\infty(x; \delta, \theta)$ is unbiased and satisfies the orthogonality condition that $\hat{W}_\infty(x; \delta, \theta) - W(x)$ is uncorrelated with all contrasts of the observations. The unbiasedness condition can be verified by straightforward calculation. To obtain the orthogonality condition, it suffices to show that $\hat{W}_\infty(x; \delta, \theta) - W(x)$ is uncorrelated with $y_k - y_1$ for all $k > 1$, $y_k - y_0$ for all $k < 0$ and $y_1 - y_0$. For example, for $k > 1$, we have

$$\begin{aligned}&\operatorname{cov}\left(\hat{W}_\infty(x; \delta, \theta) - W(x), y_k - y_1\right) \\ &\quad = \tau^2\{b_k(\delta\theta, x\theta) - b_1(\delta\theta, x\theta)\} \\ &\quad\quad - 2\delta\sigma^2\left\{\sum_{j=2}^{k} b_j(\delta\theta, x\theta)(j - 1) + (k - 1)\sum_{j=k+1}^{\infty} b_j(\delta\theta, x\theta)\right\} \\ &\quad = 0,\end{aligned}$$

by elementary calculations. Furthermore,

$$E\left(\hat{W}_\infty(x; \delta, \theta) - W(x)\right)^2 = \frac{\tau^2\theta\{\delta + 2x\theta(\delta - x)\}}{\{\delta\theta(2 + \delta\theta)\}^{1/2}}. \tag{3.2}$$

Muth (1960) gives a one-sided version of (3.1) in which y_j is observed only for $j \leq 0$.

For the remainder of the paper, set $\delta = n^{-1}$ and let $\hat{W}_\infty(x; \theta) = \hat{W}_\infty(x; n^{-1}, \theta)$. Again suppose y_j is only observed for $j = 0, \dots, n$, and consider predicting

$W(x)$, where $k/n \le x < (k+1)/n$, by

$$\tilde{W}_n(x;\theta) = S^{-1} \sum_{j=0}^{n} b_{j-k}\big(\theta n^{-1}, (x - kn^{-1})\theta\big) y_j,$$

where

$$S = \sum_{j=0}^{n} b_{j-k}\big(\theta n^{-1}, (x - kn^{-1})\theta\big).$$

$\tilde{W}_n(x;\theta)$ should provide a good approximation to the best linear unbiased predictor of $W(x)$ when n is large and x is not too close to 0 or 1 because of the geometric decay of the weights in (3.1). Specifically, let $\hat{W}_n(x;\theta)$ be the best linear unbiased predictor of $W(x)$ based on $y_0, \dots, y_n$. Letting $z = \min(x, 1-x)$, it can be shown that for all n sufficiently large,

$$E\big(\tilde{W}_n(x;\theta) - \hat{W}_\infty(x;\theta)\big)^2 = O(e^{-cn^{1/2}}),$$

where c is any constant less than z. Furthermore, using the orthogonality property of best linear unbiased predictors, it follows that

$$\begin{aligned} &E\big(\tilde{W}_n(x;\theta) - \tilde{W}_\infty(x;\theta)\big)^2 \\ &\qquad = E\big(\hat{W}_\infty(x;\theta) - \hat{W}_n(x;\theta)\big)^2 + E\big(\hat{W}_n(x;\theta) - \tilde{W}_n(x;\theta)\big)^2, \end{aligned}$$

so that

$$E\big(\hat{W}_n(x;\theta) - \hat{W}_\infty(x;\theta)\big)^2 = O(e^{-cn^{1/2}}), \tag{3.3}$$

for any $c < z$. Thus, as long as x is not too near 0 or 1, $\hat{W}_n(x;\theta)$ and $\hat{W}_\infty(x;\theta)$ are very similar.

To evaluate the effect of estimating θ on predicting $W(x)$, consider

$$\begin{aligned} &\hat{W}_n(x;\hat{\theta}) - W(x) \\ &\qquad = \{\hat{W}_n(x;\theta) - W(x)\} + \{\hat{W}_n(x;\hat{\theta}) - \hat{W}_n(x;\theta)\}, \end{aligned} \tag{3.4}$$

where $\hat{\theta}$ is some estimator for θ. If $\hat{\theta}$ is a function of the contrasts of the observations, which is the case for $\hat{\theta}_{\mathrm{MML}}$ and $\hat{\theta}_{\mathrm{GCV}}$, then the two terms on the right-hand side of (3.4) are independent since the first is independent of the contrasts and the second is a function of the contrasts. Thus, the penalty for estimating θ on prediction is to add an additional independent error to the error of the optimal predictor. For a consistent estimate of θ, we should have

$$\hat{W}_n(x;\hat{\theta}) - \hat{W}_n(x;\theta) \approx (\hat{\theta} - \theta)\frac{d}{d\theta}\hat{W}_n(x;\theta).$$

When $\hat{\theta}$ is $\hat{\theta}_{\mathrm{MML}}$ or $\hat{\theta}_{\mathrm{GCV}}$, then $\hat{\theta}$ and $(d/d\theta)\hat{W}_n(x;\theta)$ are asymptotically independent normal random variables, which allows us to derive the asymptotic distribution of $\hat{W}_n(x;\hat{\theta}_{\mathrm{MML}}) - \hat{W}_n(x;\theta)$. For fixed $x \in (0,1)$,

$$n^{1/2}\{\hat{W}_n(x;\hat{\theta}_{\mathrm{MML}}) - \hat{W}_n(x;\theta)\} \rightarrow_{\mathscr{L}} 2^{-1/2}\tau Z_1 Z_2, \tag{3.5}$$

$$n^{1/2}\{\hat{W}_n(x;\hat{\theta}_{\mathrm{GCV}}) - \hat{W}(x;\theta)\} \rightarrow_{\mathscr{L}} \tau Z_1 Z_2, \tag{3.6}$$

where Z_1 and Z_2 are iid $N(0,1)$. In addition, $n^{1/4}\{\hat{W}_n(x;\hat{\theta}_{\mathrm{MML}}) - W(x)\}$ has a density $p_n(y)$ satisfying

$$\begin{aligned}(3.7)\quad p_n(y) = {} & \left\{2\pi\left(\tau^2\left(\frac{\theta}{2}\right)^{1/2} + \frac{\tau^4}{2n^{1/2}}(1+s_n)\right)\right\}^{-1/2} \\ & \times \exp\left[-\frac{1}{2}y^2\left\{\tau^2\left(\frac{\theta}{2}\right)^{1/2} + \frac{\tau^4}{2n^{1/2}}(1+s_n)\right\}^{-1}\right] + n^{-1}r_n(y),\end{aligned}$$

where $r_n(y)$ is uniformly bounded in n and y for all n sufficiently large and $s_n \to 0$ as $n \to \infty$. Furthermore,

$$(3.8)\quad E\left(\hat{W}_n(x;\hat{\theta}_{\mathrm{MML}}) - W(x)\right)^2 = \tau^2\left\{\left(\frac{\theta}{2n}\right)^{1/2} + \frac{1}{2n} + o(n^{-1})\right\}.$$

Outlines of the proofs of (3.5)–(3.8) are in the Appendix; for details, see Stein (1989). Similarly, I would conjecture that $n^{1/4}\{\hat{W}_n(x;(\hat{\theta}_{\mathrm{GCV}}) - W(x)\}$ has a density $q_n(y)$ satisfying

$$\begin{aligned}(3.9)\quad q_n(y) = {} & \left\{2\pi\left(\tau^2\left(\frac{\theta}{2}\right)^{1/2} + \frac{\tau^4}{n^{1/2}}(1+s_n)\right)\right\}^{-1/2} \\ & \times \exp\left[-\frac{1}{2}y^2\left\{\tau^2\left(\frac{\theta}{2}\right)^{1/2} + \frac{\tau^4}{n^{1/2}}(1+s_n)\right\}^{-1}\right] + n^{-1}r_n(y)\end{aligned}$$

and

$$E\left(\hat{W}_n(x;\hat{\theta}_{\mathrm{GCV}}) - W(x)\right)^2 = \tau^2\left\{\left(\frac{\theta}{2n}\right)^{1/2} + \frac{1}{n} + o(n^{-1})\right\}.$$

Prediction of time series with estimated parameters has been studied by many authors, including Bhansali (1981), Fuller and Hasza (1981), Kunitomo and Yamamoto (1985), Lewis and Reinsel (1985) and Toyooka (1982). These studies all use the usual time series framework of a fixed distance between observations as the sample size increases [$\delta = 1$ in (2.1)]. They find that the increase in mean square prediction error caused by estimating parameters is $O(n^{-1})$ as found here, but since the mean square prediction error is $O(1)$ in their studies, the relative increase in mean square prediction error is also $O(n^{-1})$, whereas it is $O(n^{-1/2})$ here.

4. Conjectures for higher order splines. In this section, we consider how the results from the previous two sections should generalize when higher order smoothing splines are used; i.e., m is taken to be greater than 1 in (2.2). It is well known that higher order smoothing splines also correspond to optimal linear predictions under appropriate stochastic models [Kimeldorf and Wahba (1970)]. In fact, the smoothing spline of order m is identical to a problem in universal kriging of order $m - 1$ [Matheron (1973)]. Specifically, suppose

$$EW(x) = \sum_{j=0}^{m-1} \beta_j x^j \tag{4.1}$$

and that $W(\cdot)$ has the generalized covariance function

$$G(x) = \sigma^2(-1)^m |x|^{2m-1}, \tag{4.2}$$

by which we mean that if

$$\sum_{i=1}^{p} \lambda_i s_i^j = \sum_{i=1}^{q} \omega_i t_i^j = 0, \quad \text{for } j = 0, \dots, m-1,$$

then $\sum \lambda_i W(s_i)$ and $\sum \omega_i W(t_i)$ are contrasts and

$$\operatorname{cov}\left(\sum_{i=1}^{p} \lambda_i W(s_i), \sum_{i=1}^{q} \omega_i(t_i) \right) = \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_i \omega_j G(s_i - t_j).$$

As in the case $m = 1$, it will not be necessary to define the covariances of noncontrasts. Setting $\delta = n^{-1}$ throughout this section and again defining $\theta = \sigma^2/\tau^2$, I would expect the following generalizations of the results given previously in this paper to hold:

$$\begin{pmatrix} (n\theta)^{1/(4m)}(\hat{\theta}_{\mathrm{MML}} - \theta) \\ n^{1/2}(\hat{\tau}^2_{\mathrm{MML}} - \tau^2) \end{pmatrix} \to_{\mathscr{L}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} a_m\theta^2 & 0 \\ 0 & 2\tau^4 \end{pmatrix} \right) \tag{4.3}$$

and

$$(n\theta)^{1/(4m)}(\hat{\theta}_{\mathrm{GCV}} - \theta) \to_{\mathscr{L}} N(0, b_m\theta^2), \tag{4.4}$$

where

$$a_m = \frac{8m^2 \sin(\pi/2m)}{(2m-1)\{2(2m-1)!\}^{1/(2m)}}$$

$$b_m = \frac{16m^2(2m+1)\sin(\pi/2m)}{3(2m-1)\{2(2m-1)!\}^{1/(2m)}}.$$

For predicting $W(x)$ with x not too near 0 or 1,

$$E\big(\hat{W}_n(x;\theta) - W(x)\big)^2 \approx \frac{\{2\theta(2m-1)!\}^{1/(2m)}\tau^2}{2m\sin(\pi/2m)} n^{-(2m-1)/(2m)}, \tag{4.5}$$

$$\frac{E\big(\hat{W}_n(x;\hat{\theta}_{\mathrm{MML}}) - W(x)\big)^2}{E\big(\hat{W}_n(x;\theta) - W(x)\big)^2} \approx 1 + \sin(\pi/2m)\{2(2m-1)!\theta n\}^{-1/(2m)}, \tag{4.6}$$

and

$$\frac{E\big(\hat{W}_n(x;\hat{\theta}_{\mathrm{GCV}}) - W(x)\big)^2}{E\big(\hat{W}_n(x;\theta) - W(x)\big)^2} \approx 1 + \frac{2(2m+1)\sin(\pi/2m)}{3}\{2(2m-1)!\theta n\}^{-1/(2m)}. \tag{4.7}$$

I would also expect results analogous to (3.7) and (3.9) to hold. If these results are true, then they imply that the relative increase in mean square prediction error due to having to estimate θ tends to 0 at a slower rate, $n^{-1/(2m)}$, as m increases. In addition, the penalty for using $\hat{\theta}_{\mathrm{GCV}}$ rather than $\hat{\theta}_{\mathrm{MML}}$ also becomes more severe as m increases.

To see why (4.3) and (4.4) are plausible, we will consider the behavior of $\hat{\theta}_{\mathrm{MML}}$ and $\hat{\theta}_{\mathrm{GCV}}$ under a model for $W(\cdot)$ for which it exhibits similar local behavior as it does under (4.1) and (4.2). Specifically, suppose $W(\cdot)$ is a stationary Gaussian process on [0, 1] with unknown constant mean and covariance function

$$K(x) = \sigma^2(-1)^{m+1}m^{-1}B_{2m}(|x|), \tag{4.8}$$

for $x \in [-1,1]$, where B_{2m} is the Bernoulli polynomial of order $2m$ (Abramowitz and Stegun (1965), page 804]. $W(\cdot)$ is in fact a homogeneous process on the circle with perimeter 1, which follows from the general form of a homogeneous covariance function on the circle [(12) on page 74 of Yadrenko (1983)] and the Fourier series for B_{2m} [(23.1.18) in Abramowitz and Stegun (1965)]. Furthermore, $K(x)$ can be written in the form

$$K(x) = \sigma^2\left\{\sum_{j=0}^{m-1} a_j x^{2j} + (-1)^m\big(|x|^{2m-1} - m^{-1}x^{2m}\big)\right\}.$$

The local behavior of $W(\cdot)$ under (4.2) and (4.8) will be very similar, which can be seen by noting that $W^{(m-1)}(x)$ is Brownian motion under (4.2) and is Brownian bridge on [0, 1] under (4.8). Thus, it is plausible that the asymptotic behavior of $\hat{\theta}_{\mathrm{MML}}$ or $\hat{\theta}_{\mathrm{GCV}}$ should be the same under the two models.

We now outline a derivation of (4.3) and (4.4) when $W(\cdot)$ obeys (4.8). For $j = 0,\dots,n-1$, define y_j as in (2.1) and set $y_{n+j} = y_j$ for $j = 0,\dots,m-1$. Let Δ be a forward difference operator so that $\Delta y_j = y_{j+1} - y_j$. Define $Z = (Z_0,\dots,Z_{n-1})'$, where $Z_j = \Delta^m y_j$. Then Z is a set of n contrasts and since there are only n linearly independent observations, $y_0,\dots,y_{n-1}$, the distribution of Z must be singular. The covariance matrix of Z can be written as $\sigma^2 A + \tau^2 B$, where A and B are circulant matrices [Good (1950)]; specifically, for $i - j = k \pmod n$, where $|k| \le n/2$, the ij'th element of A and B are given by

$$a_k = -2(2m-1)!n^{-2m} + n^{-2m+1}g_k$$

and

$$b_k = \binom{2m}{m-k}(-1)^k,$$

respectively, where

$$g_k = (-1)^m \sum_{ij=0}^{m} \binom{m}{i}\binom{m}{j}(-1)^{i+j}|i-j+k|^{2m-1}.$$

Note that $b_k = 0$ for $|k| > m$ and $a_k = 2(2m-1)!n^{-2m}$ for $|k| > m-1$. All circulant matrices of order n share a common set of eigenvectors [Good (1950)], and if we let S be the matrix whose rows are these eigenvectors, then $X = SZ$ has a diagonal covariance matrix of the form $\sigma^2 U + \tau^2 V$. For $n > 2m$, the k'th elements of U and V are, respectively,

$$u_k = n^{-2m+1}\left\{\sum_{j=-m+1}^{m-1} g_j \cos\frac{2\pi jk}{n} - 2(2m-1)!I_{\{k=n\}}\right\}$$

and

$$v_k = (-1)^m \sum_{j=-m}^{m} \binom{2m}{m+j}(-1)^j \cos\frac{2\pi jk}{n} = \left(2\sin\frac{\pi k}{n}\right)^{2m},$$

using the formula for the eigenvalues of a circulant matrix [Good (1950)] and (1.1) in Oberhettinger (1973). Note that $v_n = 0$; it can also be shown that

$$\sum_{j=-m+1}^{m-1} g_j = 2(2m-1)!,$$

so that $u_n = 0$. As noted earlier, the fact that X has a singular distribution is expected. Let $D = V(U + \theta V)^+$, where the k'th diagonal element of $(U + \theta V)^+$ is $(u_k + \theta v_k)^{-1}$ for $k < n$ and the n'th element is zero. The MML estimate of θ and τ^2 is just the maximum likelihood estimate based on X and it can be shown that the information matrix for θ and τ^2 based on X is

$$\begin{pmatrix} \frac{1}{2}\tau^{-4}(n-1) & \frac{1}{2}\tau^{-2}\operatorname{tr} D \\ \frac{1}{2}\tau^{-2}\operatorname{tr} D & \frac{1}{2}\operatorname{tr} D^2 \end{pmatrix}. \tag{4.9}$$

For k small relative to n, the k'th diagonal element of D is

$$d_k \approx \frac{2(2m-1)!n^{-2m+1}}{(2\pi k/n)^{2m} + 2(2m-1)!\theta n^{-2m+1}}.$$

Furthermore, $d_k = d_{n-k}$ for $1 \le k \le n/2$, and for any fixed $\varepsilon > 0$, $d_k =$

$O(n^{-2m+1})$ for $\varepsilon n < k < (1-\varepsilon)n$. These results can be used to show that

$$\begin{aligned}\operatorname{tr} D^j &= \sum_{k=1}^{n} d_k^j \sim 2\sum_{k=1}^{\infty}\left\{\frac{2(2m-1)!n^{2m+1}}{(2\pi k/n)^{2m} + 2(2m-1)!\theta n^{-2m+1}}\right\}^j \\ &\sim 2\frac{n}{2\pi}\int_0^{\infty}\left\{\frac{2(2m-1)!n^{-2m+1}}{x^{2m}+2(2m-1)!\theta n^{-2m+1}}\right\}^j dx \\ &= \frac{n}{2m\pi\theta^j}\{2(2m-1)!\theta n^{-2m+1}\}^{1/(2m)}\frac{\Gamma(1/2m)\Gamma(j-1/2m)}{\Gamma(j)} \\ &= \frac{\{2(2m-1)!n\theta\}^{1/(2m)}}{2m\theta^j\sin(\pi/(2m))(j-1)!}\prod_{k=1}^{j-1}\left(k-\frac{1}{2m}\right),\end{aligned} \tag{4.10}$$

where the integral is evaluated using (3.241.4) from Gradshteyn and Ryzhik (1980) and the last line follows from properties of the gamma function [Abramowitz and Stegun (1965), pages 255–256]. (4.3) then follows from (4.9) and (4.10) when $W(\cdot)$ obeys (4.8).

To obtain the asymptotic distribution of $\hat{\theta}_{\mathrm{GCV}}$ under (4.8), analogous to (2.7), we have the approximation

$$\hat{\theta}_{\mathrm{GCV}} - \theta \approx \frac{X'\{(\operatorname{tr} C)(U+\theta V)^{+}CD - (\operatorname{tr} CD)(U+\theta V)^{+}C\}X}{\tau^2\{\operatorname{tr} C \operatorname{tr} CD^2 - (\operatorname{tr} CD)^2\}}, \tag{4.11}$$

where $C = U(U+\theta V)^{+}$. Using (4.10), it can be shown that the asymptotic variance of the right-hand side of (4.11) is given by

$$\frac{2\operatorname{tr}(D-\theta D^2)^2}{\{\operatorname{tr}(D^2-\theta D^3)\}^2} \sim \frac{16\theta^2 m^2(2m+1)\sin(\pi/(2m))}{3(2m-1)\{2(2m-1)!\theta n\}^{1/(2m)}}. \tag{4.12}$$

Using Liapunov's theorem [Chung (1974), page 200], it follows that the right-hand side of (4.11) is asymptotically normal with asymptotic variance given by (4.12). (4.4) then follows from (4.11) and (4.12) when $W(\cdot)$ obeys (4.8). Note that the ratio of the asymptotic variances of $\hat{\theta}_{\mathrm{GCV}}$ to $\hat{\theta}_{\mathrm{MML}}$ is $(4m+2)/3$, so that the loss of efficiency in using $\hat{\theta}_{\mathrm{GCV}}$ rather than $\hat{\theta}_{\mathrm{MML}}$ increases with m. Furthermore, (4.3) and (4.4) agree with the results from Section 2 when $m = 1$, lending further support to the conjecture that they are valid for $m > 1$ when $W(\cdot)$ obeys (4.2).

Let us next consider (4.5). Silverman (1984) showed that smoothing splines can be approximated by kernel smoothers, so it is plausible that these kernel smoothers will have mean square error very near to the mean square error of the smoothing spline under the stochastic model for which the smoothing spline is the best linear unbiased predictor. Specifically, following Silverman (1984), suppose $\kappa(x)$ satisfies $(-1)^m\kappa^{(2m)}(x) + \kappa(x) = \delta_x$, where δ_x is the

Dirac delta function. This differential equation can be solved using Fourier transforms [Silverman (1984)] and

$$\kappa(x) = \frac{1}{2\pi}\int \frac{e^{i\omega x}}{1+\omega^{2m}}\,d\omega$$
$$= \frac{1}{2m}\sum_{k=1}^{m} \exp\left\{-|x|\sin\frac{(2k-1)\pi}{2m}\right\}$$
$$\times \sin\left\{\frac{(2k-1)\pi}{2m} + |x|\cos\frac{(2k-1)\pi}{2m}\right\}.$$

Then $\kappa(\cdot)$ is an even function and satisfies

$$\begin{aligned} &\int \kappa(x)\,dx = 1, \\ &\int x^k \kappa(x)\,dx = 0, \quad \text{for } k = 1,\dots,m-1. \end{aligned} \tag{4.13}$$

For $x \in (0,1)$ and not too near zero or one, consider predicting $W(x)$ by

$$h_n \sum_{j=0}^{n} \kappa(h_n|nx-j|)y_j,$$

which considering the results of Silverman (1984), is likely to be asymptotically optimal for optimal h_n. For n large and h_n small, because $\kappa(\cdot)$ satisfies (4.13), the prediction error will be nearly a contrast. If we treated it as if it were exactly a contrast, then using (4.2), we would have

$$\begin{aligned} &E\left(W(x) - h_n\sum_{j=0}^{n}\kappa(h_n|nx-j|)y_j\right)^2 \\ &\quad = h_n^2\tau^2\sum_{j=0}^{n}\kappa(h_n|nx-j|)^2 \\ &\qquad - 2h_n\sigma^2\sum_{j=0}^{n}\kappa(h_n|nx-j|)(-1)^m|x-jn^{-1}|^{2m-1} \\ &\qquad + h_n^2\sigma^2\sum_{ij=0}^{n}\kappa(h_n|nx-i|)\kappa(h'_n|nx-j|)(-1)^m|n^{-1}(i-j)|^{2m-1} \\ &\quad \approx h_n\tau^2 b_0 + \sigma^2(nh_n)^{-(2m-1)}b_1, \end{aligned} \tag{4.14}$$

for n large and h_n small, where

$$b_0 = \int \kappa(x)^2\,dx = \frac{2m-1}{4m^2\sin(\pi/2m)},$$

by Parseval's relation and

$$b_1 = (-1)^m\left\{\int\int\kappa(x)\kappa(y)|x-y|^{2m-1}\,dx\,dy - 2\int\kappa(x)|x|^{2m-1}\,dx\right\}$$
$$= \frac{(2m-1)!}{2m^2\sin(\pi/2m)},$$

by direct calculation. The value of h_n minimizing (4.14) is given by

$$h_{\text{opt}} = \{2\theta(2m-1)!\}^{1/(2m)}n^{-(2m-1)/(2m)}$$

and (4.5) follows. When $m = 1$, (4.5) yields $\tau^2\theta/(2n)^{1/2}$, in agreement with (3.2) for $\delta = n^{-1}$ and n large. If we let $h_n = h_{\text{opt}}(1 + \alpha n^{-1/(4m)})$ where α is a constant, then (4.14) yields

$$E\left(W(x) - h_{\text{opt}}(1+\alpha n^{-1/(4m)})\sum_{j=0}^{n}\kappa\big(h_{\text{opt}}(1+\alpha n^{-1/(4m)})|nx-j|\big)y_j\right)^2$$
$$\approx \{2\theta(2m-1)!\}^{1/(2m)}\frac{\tau^2}{2m\sin(\pi/2m)} \tag{4.15}$$
$$\times n^{-(2m-1)/(2m)}\left(1+\frac{(2m-1)\alpha^2}{2n^{1/(2m)}}\right).$$

If we assume that $\hat{\theta}_{\text{MML}}$ and $\hat{\theta}_{\text{GCV}}$ are independent of $(d/d\theta)\hat{W}_n(x; n^{-1}, \theta)$, then (4.6) and (4.7) follow from (4.3), (4.4) and (4.15).

5. Discussion. Even assuming that all of the conjectures in Section 4 are true and that (4.6) and (4.7) are indicative of the advantage $\hat{\theta}_{\text{MML}}$ has over $\hat{\theta}_{\text{GCV}}$ in sample sizes occurring in practice, one can argue that this is a small price to pay for the potentially vastly superior performance of $\hat{\theta}_{\text{GCV}}$ over $\hat{\theta}_{\text{MML}}$ when the stochastic model for $W(\cdot)$ is incorrect and $W(\cdot)$ is in fact a deterministic function with square integrable $2m$'th derivative [Wahba (1985)]. It is not surprising that maximum likelihood procedures can perform poorly if no element in the stochastic model on which the likelihood is based is even roughly similar to the underlying truth. If we want likelihood methods to yield good predictions for a large class of functions, we need to choose a broad class of stochastic models. One way to obtain a more flexible model would be to use the model defined by (4.1) and (4.2) where m and σ^2 are both considered unknown. However, some modification of both GCV and MML would be needed if m were to be allowed to be any positive integer since neither criterion is well defined when m is greater than the number of observations. If in fact $W^{(m-1)}(\cdot)$ behaves at least locally like Brownian motion, then I would expect suitably modified versions of both GCV and MML estimates of m would equal m with probability tending to 1 as $n \to \infty$, and that the first order asymptotic behavior of the estimates of θ would be as given in Section 4 where m is assumed known. For general $W(\cdot)$, the asymptotic behavior of estimates of θ and m might be rather complicated. In particular, if $W(\cdot)$ is analytic on $[0, 1]$, then any estimate of m that yields predictions with good large sample properties would presumably tend to infinity as $n \to \infty$, although the rate of increase may be very slow.

APPENDIX

Outlines of the proofs of (3.5)–(3.8) are given; see Stein (1989) for details. We first derive an exact expression for $(d/d\theta)\hat{W}_n(x;\theta)$. For $x \in [kn^{-1}, (k+1)n^{-1})$, define $W_n^*(x) = (k+1-nx)y_k + (nx-k)y_{k+1}$. Now, $\hat{W}_n(x;\theta)$ can be obtained by finding the optimal linear predictor of $W(x) - W_n^*(x)$ based on Z. By straightforward calculation,

$$g_j(x) \triangleq \tau^{-2}\operatorname{cov}\left(W(x) - W_n^*(x), y_j - y_{j-1}\right)$$
$$= \begin{cases} nx - k - 1, & j = k, \\ 2k + 1 - 2nx, & j = k+1, \\ nx - k, & j = k + 2, \\ 0, & \text{otherwise}. \end{cases}$$

Let $g(x) = (g_1(x), \dots, g_n(x))'$. Then

$$\hat{W}_n(x;\theta) = W_n^*(x) + g(x)'(T + 2\theta n^{-1}I)^{-1}Z$$

and $W_n^*(x)$ and $g(x)$ do not depend on θ, so

$$\frac{d}{d\theta}\hat{W}_n(x;\theta) = -\frac{2}{n}g(x)'(T + 2\theta n^{-1}I)^{-2}Z$$
$$= -\frac{2}{n}g(x)'S(\Lambda + 2\theta n^{-1}I)^{-2}X,$$

where S and X are defined in Section 2. Calculating $Sg(x)$, we obtain

$$\frac{d}{d\theta}\hat{W}_n(x;\theta)$$
$$= -\frac{4}{n}\left(\frac{2}{n+1}\right)^{1/2}\sum_{j=1}^{n}\sin\frac{\pi j}{2(n+1)}$$
$$\times\left\{\cos\frac{\pi j(2k+1)}{2(n+1)} - 2(nx-k)\sin\frac{\pi j}{2(n+1)}\sin\frac{\pi j(k+1)}{n+1}\right\}$$
$$\times\frac{X_j}{\left(\lambda_j + 2\theta n^{-1}\right)^2},$$

where $X = (X_1, \dots, X_n)'$. Thus, $(d/d\theta)\hat{W}_n(x;\theta)$ is normally distributed with mean 0 and (Stein 1989)

$$\text{(A.1)}\qquad \operatorname{var}\left(\frac{d}{d\theta}\hat{W}_n(x;\theta)\right) \sim \frac{\tau^2}{2^{7/2}\theta^{3/2}n^{1/2}},$$

for fixed $x \in (0,1)$.

For any estimate $\hat{\theta}$ of θ, we can write

$$\text{(A.2)}\qquad \hat{W}_n(x;\hat{\theta}) - \hat{W}_n(x;\theta) = (\hat{\theta}-\theta)\frac{d}{d\theta}\hat{W}_n(x;\theta) + \frac{4}{n^2}(\hat{\theta}-\theta)^2 g(x)'S\left(\lambda + 2n^{-1}D(\hat{\theta})\right)^{-3}X,$$

where $D(\hat{\theta})$ is a diagonal matrix with each diagonal element between θ and $\hat{\theta}$. Let us next show that $\hat{\theta}_{\text{MML}} - \theta$ and $(d/d\theta)\hat{W}_n(x;\theta)$ are asymptotically independent and jointly normal. We will need the following lemma, proven in Stein (1989):

LEMMA 1. *If for each n, Z_{in}, $i = 1,\dots,n$ are iid $N(0,1)$,*

$$\sum_{i=1}^{n} a_{in}^2 = 1, \qquad \sum_{i=1}^{n} b_{in}^2 = \tfrac{1}{2} \quad \textit{and} \quad \lim_{n\to\infty} \max_{1\le i\le n} |b_{in}| = 0,$$

then

$$\left(\sum_{i=1}^{n} a_{in}Z_{in}, \sum_{i=1}^{n} b_{in}(Z_{in}^2 - 1)\right) \to_{\mathscr{L}} N(0,I).$$

By linearizing the likelihood equations, we can show

$$\hat{\theta}_{\text{MML}} - \theta = \frac{V'(\Lambda + 2\theta n^{-1}I)^{-1}V - \operatorname{tr}(\Lambda + 2\theta n^{-1}I)^{-1}}{2n^{-1}\operatorname{tr}(\Lambda + 2\theta n^{-1}I)^{-2}} + o_p(n^{-1/4}),$$

where $V = \tau^{-1}(\Lambda + 2\theta n^{-1}I)^{-1/2}X \sim N(0,I)$. Then Lemma 1 can be applied to

$$V'(\Lambda + 2\theta n^{-1}I)^{-1}V - \operatorname{tr}(\Lambda + 2\theta n^{-1}I)^{-1}$$

and

$$(d/d\theta)\hat{W}(x;\theta) = -2n^{-1}g(x)'S(\Lambda + 2\theta n^{-1}I)^{-3/2}V$$

and the asymptotic independence and joint normality of $\hat{\theta}_{\text{MML}} - \theta$ and $(d/d\theta)\hat{W}(x;\theta)$ follow. Finally, we can show that the remainder term in (A.2) is $O_p(n^{-1})$, proving (3.5). Using (2.7), (3.6) similarly follows.

A heuristic justification of (3.8) can be obtained using (3.4) and (A.2) and ignoring all lower order terms. A rigorous proof is given in Stein (1989) and depends crucially on the tail probabilities for $n^{1/4}(\hat{\theta}_{\text{MML}} - \theta)$ being sufficiently small for large n, which can be shown using Theorem 5.1 of Ibragimov and Has'minskii (1981). Furthermore, it can be shown that the fourth moment of $n^{1/2}\{\hat{W}_n(x;\hat{\theta}_{\text{MML}}) - \hat{W}_n(x;\theta)\}$ is bounded for all n sufficiently large, and by symmetry considerations, its first and third moments are zero. From Theorem

6.4.2 of Chung [(1974), page 168], it follows that $\phi_n(t)$, the characteristic function of $n^{1/2}\{\hat{W}_n(x;\hat{\theta}_{\mathrm{MML}}) - \hat{W}_n(x;\theta)\}$, is, for all n sufficiently large, of the form

$$\phi_n(t) = 1 - \tfrac{1}{4}\tau^2(1 + r_n)t^2 + \alpha_n(t)t^4,$$

where $r_n \to 0$ as $n \to \infty$ and $|\alpha_n(t)|$ is uniformly bounded in both n and t. Using (3.2), (3.3) and the independence of the two terms on the right-hand side of (3.4), it follows that $n^{1/4}\{\hat{W}_n(x;\hat{\theta}_{\mathrm{MML}}) - W(x)\}$ has characteristic function

$$\exp\left[-\left\{\frac{\tau^2\theta^{1/2}}{2^{3/2}} + O(n^{-1})\right\}t^2\right]\left\{1 - \frac{1}{4}\tau^2(1 + r_n)n^{-1/2}t^2 + \alpha_n(n^{-1/4}t)n^{-1}t^4\right\}$$
$$= \exp\left[-\left\{\frac{\tau^2\theta^{1/2}}{2^{3/2}} - \frac{\tau^2(1 + s_n)}{4n^{1/2}}\right\}t^2\right]\{1 + \beta_n(n^{-1/4}t)n^{-1}t^4\},$$

where $s_n \to 0$ and $\beta_n(\cdot)$ is uniformly bounded for all t and all n sufficiently large. Since this characteristic function is in L^1, $n^{1/4}\{\hat{W}_n(x;\hat{\theta}_{\mathrm{MML}}) - W(x)\}$ has the density

$$p_n(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\left\{-iyt - \tau^2\left(\frac{\theta^{1/2}}{2^{3/2}} + \frac{1 + s_n}{4n^{1/2}}\right)t^2\right\}$$
$$\times\{1 + \beta_n(n^{-1/4}t)n^{-1}t^4\}\,dt,$$

and (3.9) follows.

REFERENCES

Abramowitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions*, 9th ed. Dover, New York.

Bhansali, R. J. (1981). Effects of not knowing the order of an autoregression on the mean squared error of prediction I. *J. Amer. Statist. Assoc.* **78** 588–597.

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

Fuller, W. A. and Hasza, D. P. (1981). Properties of predictors for autoregressive time series. *J. Amer. Statist. Assoc.* **76** 155–161.

Good, I. J. (1950). On the inversion of circulant matrices. *Biometrika* **37** 185–186.

Gradshteyn, I. S. and Ryzhik, I. M. (1980). *Table of Integrals, Series, and Products*. Academic, Orlando.

Hutchinson, M. F. and de Hoog, F. R. (1985). Smoothing data with spline functions. *Numer. Math.* **31** 377–403.

Ibragimov, I. V. and Has'minskii, R. K. (1981). *Statistical Estimation—Asymptotic Theory* (S. Kotz, transl.). Springer, New York.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic, New York.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41** 495–502.

Kohn, R. and Ansley, C. F. (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J. Sci. Statist. Comput.* **8** 33–48.

Kunitomo, N. and Yamamoto, T. (1985). Properties of predictors in misspecified autoregressive time series models. *J. Amer. Statist. Assoc.* **80** 941–950.

Lewis, R. and Reinsel, G. C. (1985). Prediction of multivariate time series by autoregressive model fitting. *J. Multivariate Anal.* **16** 393–411.

MATHERON, G. (1973). The intrinsic random functions and their applications. *J. Appl. Probab.* **5** 439–468.

MUTH, J. F. (1960). Optimal properties of exponentially weighted forecasts. *J. Amer. Statist. Assoc.* **55** 299–306.

OBERHETTINGER, F. (1973). *Fourier Expansions—A Collection of Formulas*. Academic, New York.

O'SULLIVAN, F. (1985). Comments on "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B. Silverman. *J. Roy. Statist. Soc. Ser. B* **47** 39–40.

PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58** 545–554.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York.

SILVERMAN, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12** 898–916.

STEIN, M. L. (1987). Minimum norm quadratic estimation of spatial variograms. *J. Amer. Statist. Assoc.* **82** 765–772.

STEIN, M. L. (1989). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. Technical Report No. 249, Dept. Statist., Univ. Chicago.

STRANG, G. (1976). *Linear Algebra and Its Applications*. Academic, New York.

TOYOOKA, Y. (1982). Prediction error in a linear model with estimated parameters. *Biometrika* **69** 453–459.

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. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Rev.* **108** 36–57.

YADRENKO, M. I. (1983). *Spectral Theory of Random Fields*. Optimization Software, New York.

DEPARTMENT OF STATISTICS
THE UNIVERSITY OF CHICAGO
CHICAGO, IL 60637