The Annals of Statistics
1986, Vol. 14, No. 3, 1101–1112

ASYMPTOTIC OPTIMALITY OF C_L AND GENERALIZED CROSS-VALIDATION IN RIDGE REGRESSION WITH APPLICATION TO SPLINE SMOOTHING[1]

BY KER-CHAU LI

University of California, Los Angeles

The asymptotic optimality of Mallows' C_L and generalized cross-validation is demonstrated in the setting of ridge regression. An application is made to spline smoothing in nonparametric regression. A counterexample is given to help understand why sometimes GCV may not be asymptotically optimal. The coefficient of variation for the eigenvalues oi the information matrix must be large in order to guarantee the optimality of GCV. The proof is based on the connection between GCV and Stein's unbiased risk estimate.

1. Introduction. Suppose that we observe n independent normal random variables y_i, $i = 1, 2, \ldots, n$, each associated with p_n explanatory variables, $x_{i1}, x_{i2}, \ldots, x_{ip_n}$. In ridge regression, we may estimate the mean $\boldsymbol{\mu}_n = (\mu_1, \ldots, \mu_n)'$ of $\mathbf{y}_n = (y_1, \ldots, y_n)'$ by $\hat{\boldsymbol{\mu}}_n(h) = X_n(X_n'X_n + hI)^{-1}X_n'\mathbf{y}_n$ where X_n is the $n \times p_n$ design matrix (x_{ij}). The choice of ridge parameter h is crucial and many procedures have been proposed. Two of them, namely C_L (Mallows, 1973) and GCV (Craven and Wahba, 1979), will be studied here.

Let σ^2 be the common variance of y_i and put $M_n(h) = X_n(X_n'X_n + hI)^{-1}X_n'$. C_L selects h by minimizing

$$n^{-1}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2 + 2\sigma^2 n^{-1}\operatorname{tr} M_n(h). \tag{1.1}$$

Subtracting σ^2, (1.1) gives an unbiased estimate of the risk $R_n(h) = En^{-1}\|\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2$. If σ^2 is unknown, then it has to be replaced by an estimate $\hat{\sigma}^2$. Thus the stability of $\hat{\sigma}^2$ may influence the performance of C_L. Generalized cross-validation (GCV) does not need σ^2. It selects h by minimizing

$$\frac{n^{-1}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2}{\left(1 - n^{-1}\operatorname{tr} M_n(h)\right)^2}. \tag{1.2}$$

Let $\hat{h}_M$ and $\hat{h}_G$ denote the h selected by C_L and GCV, respectively. Put $L_n(h) = n^{-1}\|\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2$. We shall show that they are asymptotically optimal (a.o.) in the sense that as $n \to \infty$,

$$\frac{L_n(\hat{h})}{\inf_{h \geq 0} L_n(h)} \to 1, \quad \text{in probability}, \tag{1.3}$$

where $\hat{h} = \hat{h}_M, \hat{h}_G$.

Received October 1984; revised August 1985.

[1] This work was supported by the National Science Foundation under Grant No. MCS-8200631.

AMS 1980 *subject classifications*. 62G05, 62G99.

Key words and phrases. C_L, generalized cross-validation, ridge regression, smoothing splines, Stein estimates, Stein's unbiased risk estimates.

The following is the only condition needed for C_L to be a.o:

$$\text{(A.1)} \qquad \inf_{h \geq 0} nR_n(h) \to \infty.$$

There are no explicit restrictions on the sequence of design matrices X_n. But an implicit one more or less implied by (A.1) is that p_n tends to ∞ as $n \to \infty$. Without (A.1), it seems that no selection procedure can be a.o.; otherwise the resulting estimates may possess unattainably small risk.

The result for GCV requires, in addition to (A.1), a certain condition on the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p_n} \geq 0$ of the information matrix $X_n'X_n$. Roughly speaking, the coefficient of variation for the λ_i's should tend to infinity as $n \to \infty$ (see (A.2) of Section 3) and hence make the problem ill-posed. We also provide an example to show that if the spread of these eigenvalues does not tend to infinity, then GCV may not be a.o.

As an application, we consider the spline smoothing problem. Suppose $\mu_i = f(x_i)$, with the unknown function $f \in W_2^k[0,1] = \{f\colon f$ has absolutely continuous derivatives, $f', \dots, f^{(k-1)}$ and $\int_0^1 f^{(k)}(x)^2\,dx < \infty\}$ and $x_i \in [0,1]$. The smoothing spline estimate $\hat{f}_h$ of f is the solution of

$$\min_{f \in W_2^k[0,1]} \; n^{-1} \sum_{i=1}^{n} (y_i - f(x_i))^2 + h \int_0^1 f^{(k)}(x)^2\,dx.$$

It is well known that $\hat{\mu}_n(h) = (\hat{f}_h(x_1), \dots, \hat{f}_h(x_n))'$ takes the form of ridge regression with the first k eigenvalues being $+\infty$ (see, e.g., Li (1985)). We shall show that C_L and GCV are both a.o. if f is not a polynomial of degree $k - 1$ or less.

For spline smoothing, there have been some results in the literature that are related to the a.o. property of GCV, mostly due to Wahba and her collaborators. Let h_G be the minimizer of the expectation of (1.2) over $h \geq 0$. It was shown in Craven and Wahba (1979) that

$$\text{(1.4)} \qquad R_n(h_G)/\inf R_n(h) \to 1.$$

See Wahba (1985) for more information. However, it is clear that the results of this type do not necessarily lead to the a.o. of (1.3). For example, if f is a polynomial of degree $k - 1$, then (1.3) cannot hold for any selection procedure; but it is easy to see that (1.4) holds. The big gap between (1.3) and (1.4) was closed significantly by Speckman (1982) who established (1.3) under the assumption that $\hat{h}_G$ is selected by minimizing (1.2) over h in some closed interval that converges to 0 in some fashion as n tends to infinity. Speckman's result was derived (by Cox (1983)) without the normality assumption. Erdal (1983) and Golub, Heath, and Wahba (1979) discussed the properties of GCV for general ridge regression.

Our method in proving (1.3) for GCV is based on the connection between Stein's unbiased risk estimate (Stein, 1981) and GCV. This connection has been used to demonstrate the consistency of GCV in many settings (Li, 1985). The a.o. of C_L and GCV in the discrete index set case, such as model-selection or nearest neighbor nonparametric regression, was established in Li (1984).

2. Mallows' C_L.

In this section we shall prove the following theorem.

THEOREM 1. *Under* (A.1), (1.3) *holds for* $\hat{h} = \hat{h}_M$.

We shall assume that X_n is diagonal, i.e., $x_{ij} = 0$ for $i \neq j$ and $x_{ii} = \lambda_{ii}^{1/2}$. This is without loss of generality because after a suitable orthogonal transformation we can reduce any X_n to a diagonal form without changing the error distribution (due to normality). Now $M_n(h)$ is simply an $n \times n$ diagonal matrix with $\lambda_i(h + \lambda_i)^{-1}$ as the ith diagonal element. Here we put $\lambda_i = 0$ for $i > p_n$. Note that λ_i may depend on n.

Let $\mathbf{e}_n = (e_1, e_2, \dots, e_n)' = \mathbf{y}_n - \boldsymbol{\mu}_n$ and $A_n(h) = I - M_n(h)$. Clearly,

$$\begin{aligned} n^{-1}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2 &+ 2\sigma^2 n^{-1}\operatorname{tr} M_n(h) \\ &= n^{-1}\|\mathbf{e}_n\|^2 + L_n(h) + 2n^{-1}\langle \mathbf{e}_n, A_n(h)\boldsymbol{\mu}_n\rangle \\ &\quad + 2n^{-1}\big(\sigma^2 \operatorname{tr} M_n(h) - \langle \mathbf{e}_n, M_n(h)\mathbf{e}_n\rangle\big). \end{aligned}$$

Therefore, it is enough to show that in probability,

$$\sup_{h \ge 0} \left| n^{-1}\langle \mathbf{e}_n, A_n(h)\boldsymbol{\mu}_n\rangle \right| \Big/ R_n(h) \to 0, \tag{2.1}$$

$$\sup_{h \ge 0} n^{-1}\left| \sigma^2 \operatorname{tr} M_n(h) - \langle \mathbf{e}_n, M_n(h)\mathbf{e}_n\rangle \right| \Big/ R_n(h) \to 0, \tag{2.2}$$

and

$$\sup_{h \ge 0} | L_n(h)/R_n(h) - 1 | \to 0. \tag{2.3}$$

The following useful lemma is recalled from Li (1985). It first appeared in Speckman (1982, 1985).

LEMMA 2.1. *Assume that W_i, $i = 1, 2, \dots, n$, are independent random variables with means* 0 *and finite second moments. Then for any $\delta > 0$, we have*

$$P\left\{ \sup_{0 \le c_1 \le \cdots \le c_n \le a} \left| \sum_{i=1}^n c_i W_i \right| \ge \delta \right\} \le \delta^{-2} a^2 E\left(\sum_{i=1}^n W_i \right)^2.$$

If the W_i's have finite fourth moments, then

$$P\left\{ \sup_{0 \le c_1 \le \cdots \le c_n \le a} \left| \sum_{i=1}^n c_i W_i \right| \ge \delta \right\} \le \delta^{-4} a^4 E\left(\sum_{i=1}^n W_i \right)^4.$$

We begin to prove (2.1). Put $B_n(h) = \sum_{i=1}^n \mu_i^2 (\lambda_i + h)^{-1}$. Since $nR_n(h) \ge h^2 B_n(h)$, it is enough to show that

$$\sup_{h \ge 0} \left| \sum_{i=1}^n e_i \mu_i (\lambda_i + h)^{-1} \right| \Big/ B_n(h)^{1/2} (nR_n(h))^{1/2} \to 0. \tag{2.1$'$}$$

For each n, let $I_1(j) = \{1, 2, \dots, j\}$ and $I_2(j) = \{j + 1, \dots, n\}$. Define $\bar{n}$ to be the largest i such that $\lambda_i \neq 0$. Put $Q_n = \inf_{h \ge 0} nR_n(h)$ and $V_n(h) = \sum_{i=1}^n \lambda_i^2 (\lambda_i + h)^{-2}$. Clearly (2.1$'$) will hold if we can show that for any natural number k and for $l = 1, 2,$

$$(2.4)\qquad \sup_{h \geq \lambda_k} \left| \sum_{i \in I_l(k)} e_i \mu_i (\lambda_i + h)^{-1} \right| \Big/ B_n(h)^{1/2} Q_n^{1/2} \to 0,$$

and that for any $\varepsilon > 0$, there exist constants $c_1(\varepsilon), c_2(\varepsilon)$ such that

$$(2.5)\qquad \begin{aligned} &P\left\{ \sup_{j=k,\dots,\bar{n}} \sup_{\lambda_{j+1} \leq h \leq \lambda_j} \left| \sum_{i \in I_l(j)} e_i \mu_i (\lambda_i + h)^{-1} \right| \Big/ B_n(h)^{1/2} V_n(h)^{1/2} > \varepsilon \right\} \\ &\quad \leq c_l(\varepsilon) \sum_{j=k}^{\infty} j^{-2}, \end{aligned}$$

for $l = 1, 2.$

Proof of (2.4). When $l = 1$, the left side of (2.4) does not exceed

$$Q_n^{-1/2} \max_{1 \leq i \leq k} |e_i| \sum_{i=1}^{k} \sup_{h \geq \lambda_k} |\mu_i| (\lambda_i + h)^{-1} / B_n(h)^{1/2} \leq Q_n^{-1/2} \max_{1 \leq i \leq k} |e_i| k,$$

which tends to 0 because of (A.1), as desired.

When $l = 2$, it suffices to show that for any $\varepsilon > 0$,

$$(2.6)\qquad P\left\{ \sup_{h \geq \lambda_k} \left| \sum_{i=k+1}^{n} e_i \mu_i h (\lambda_i + h)^{-1} \right| \Big/ h B_n(h)^{1/2} Q_n^{1/2} \geq \varepsilon \right\} \to 0.$$

Since $h^2 B_n(h)$ is nondecreasing in h, the left side of (2.6) does not exceed

$$\begin{aligned} &P\left\{ \sup_{h \geq \lambda_k} \left| \sum_{i=k+1}^{n} e_i \mu_i h (\lambda_i + h)^{-1} \right| \Big/ \lambda_k B_n(\lambda_k)^{1/2} Q_n^{1/2} \geq \varepsilon \right\} \\ &\quad \leq P\left\{ \left| \sum_{i=k+1}^{n} e_i \mu_i \lambda_k (\lambda_i + \lambda_k)^{-1} \right| \geq \tfrac{1}{2} \varepsilon \lambda_k B_n(\lambda_k)^{1/2} Q_n^{1/2} \right\} \\ &\qquad + P\left\{ \sup_{h \geq \lambda_k} \left| \sum_{i=k+1}^{n} e_i \mu_i \left(h(\lambda_i + h)^{-1} - \lambda_k (\lambda_i + \lambda_k)^{-1} \right) \right| \right. \\ &\qquad\qquad \left. \geq \tfrac{1}{2} \varepsilon \lambda_k B_n(\lambda_k)^{1/2} Q_n^{1/2} \right\}. \end{aligned}$$

By Chebyshev's inequality, the first term of the last expression does not exceed

$$\left(\tfrac{1}{2} \varepsilon B_n^{1/2}(\lambda_k) Q_n^{1/2} \right)^{-2} E\left(\sum_{i=k+1}^{n} e_i \mu_i (\lambda_i + \lambda_k)^{-1} \right)^2 \leq 4 \varepsilon^{-2} Q_n^{-1} \sigma^2 \to 0,$$

because of (A.1). The second term is also no greater than $4\varepsilon^{-2} Q_n^{-1} \sigma^2$ due to Lemma 2.1. To see this, observe that

$$h(\lambda_i + h)^{-1} - \lambda_k (\lambda_i + \lambda_k)^{-1} = (\lambda_i + \lambda_k)^{-1} (h - \lambda_k) \lambda_i (\lambda_i + h)^{-1},$$

and that for $i \geq k + 1$ and $h \geq \lambda_k$, $(h - \lambda_k)\lambda_i(\lambda_i + h)^{-1}$ is nonincreasing in i and is no greater than λ_k. Now set $W_i = e_i \mu_i (\lambda_i + \lambda_k)^{-1}$, $a = \lambda_k$, and $\delta = (\varepsilon/2) \lambda_k B_n(\lambda_k)^{1/2} Q^{1/2}$ in Lemma 2.1 to yield the desired bound. Therefore (2.4) is proved. □

Proof of (2.5). For $l = 1$, since $B_n(h)$ and $V_n(h)$ are both nonincreasing in h, the left side of (2.5) does not exceed

$$\sum_{j=k}^{\bar n} P\left\{ \sup_{\lambda_{j+1}\le h\le\lambda_j} \left|\sum_{i=1}^{j} e_i\mu_i(\lambda_i+h)^{-1}\right| \Big/ B_n(\lambda_j)^{1/2}V_n(\lambda_j)^{1/2} \ge \varepsilon\right\}$$

$$\le \sum_{j=k}^{\bar n} P\left\{\left|\sum_{i=1}^{j} e_i\mu_i(\lambda_i +\cdot\lambda_j)^{-1}\right| \ge \tfrac{1}{2}\varepsilon B_n(\lambda_j)^{1/2}V_n(\lambda_j)^{1/2}\right\}$$

$$+ \sum_{j=k}^{\bar n} P\left\{ \sup_{\lambda_{j+1}\le h\le\lambda_j} \left|\sum_{i=1}^{j} e_i\mu_i\left((\lambda_i+h)^{-1}-(\lambda_i+\lambda_j)^{-1}\right)\right| \ge \tfrac{1}{2}\varepsilon B_n(\lambda_j)^{1/2}V_n(\lambda_j)^{1/2}\right\}.$$

By Chebyshev's inequality, the first term of the last expression does not exceed

$$\sum_{j=k}^{\bar n}\left(\tfrac{1}{2}\varepsilon B_n(\lambda_j)^{1/2}V_n(\lambda_j)^{1/2}\right)^{-4} E\left(\sum_{i=1}^{j} e_i\mu_i(\lambda_i+\lambda_j)^{-1}\right)^4. \tag{2.7}$$

The second term is also no greater than (2.7). To see this, observe that $(\lambda_i + h)^{-1} - (\lambda_i + \lambda_j)^{-1} = (\lambda_i + \lambda_j)^{-1}(\lambda_j - h)(\lambda_i + h)^{-1}$ and that for $\lambda_{j+1} \le h \le \lambda_j$ and $i \le j$, $(\lambda_j - h)(\lambda_i + h)^{-1}$ is nondecreasing in i and is no greater than 1. Now in Lemma 2.1 set $W_i = e_i\mu_i(\lambda_i + \lambda_j)^{-1}$, $a = 1$, and $\delta = \frac{1}{2}\varepsilon B_n(\lambda_j)^{1/2}V_n(\lambda_j)^{1/2}$ to yield the desired bound. Now since

$$E\left(\sum_{i=1}^{j} e_i\mu_i(\lambda_i+\lambda_j)^{-1}\right)^4 \le C\left(\sum_{i=1}^{j}\mu_i^2(\lambda_i+\lambda_j)^{-2}\right)^2$$

for some constant C, (2.7) does not exceed $16C\varepsilon^{-4}\sum_{j=k}^{\bar n}V_n(\lambda_j)^{-2}$. Finally, it is clear that for $\lambda_j \ne 0$, $V_n(\lambda_j) \ge \sum_{i=1}^{j}\lambda_i^2(\lambda_i + \lambda_j)^{-2} \ge \frac{1}{4}j$. Thus (2.5) is established for $l = 1$.

Turning to the case $l = 2$, since $h^2B_n(h)$ is nondecreasing in h, the left side of (2.5) does not exceed

$$\sum_{j=k}^{\bar n} P\left\{ \sup_{\lambda_{j+1}\le h\le\lambda_j} \left|\sum_{i=j+1}^{n} e_i\mu_i h(\lambda_i+h)^{-1}\right| \Big/ \lambda_{j+1}B_n(\lambda_{j+1})^{1/2}V_n(\lambda_j)^{1/2} \ge \varepsilon\right\}$$

$$\le \sum_{j=k}^{\bar n} P\left\{\left|\sum_{i=j+1}^{n} e_i\mu_i\lambda_{j+1}(\lambda_i+\lambda_{j+1})^{-1}\right| \Big/ \lambda_{j+1}B_n(\lambda_{j+1})^{1/2}V_n(\lambda_j)^{1/2} \ge \tfrac{1}{2}\varepsilon\right\}$$

$$+ \sum_{j=k}^{\bar n} P\left\{ \sup_{\lambda_{j+1}\le h\le\lambda_j} \left|\sum_{i=j+1}^{n} e_i\mu_i\left(h(\lambda_i+h)^{-1}-\lambda_{j+1}(\lambda_i+\lambda_{j+1})^{-1}\right)\right| \ge \tfrac{1}{2}\varepsilon\lambda_{j+1}B_n(\lambda_{j+1})^{1/2}V_n(\lambda_j)^{1/2}\right\}.$$

As before, using Chebyshev's inequality for the first term and Lemma 2.1 for the second term, we may obtain the desired bound. Note that when using Lemma 2.1, we observe that $h(\lambda_i + h)^{-1} - \lambda_{j+1}(\lambda_i + \lambda_{j+1})^{-1} = (\lambda_i + \lambda_{j+1})^{-1}\lambda_i(h - \lambda_{j+1})(\lambda_i + h)^{-1}$ and set $W_i = e_i\mu_i(\lambda_i + \lambda_{j+1})^{-1}$, $a = \lambda_{j+1}$ and $\delta = \frac{1}{2}\varepsilon\lambda_{j+1}B_n(\lambda_{j+1})^{1/2}V_n(\lambda_j)^{1/2}$. The details are omitted. This completes the proof of (2.5). Hence (2.1) is established. □

To prove (2.2), it suffices to show that

$$\text{(2.2}'\text{)}\qquad \sup_{h\geq 0}\left|\sum_{i=1}^{n}\left(\sigma^2 - e_i^2\right)\lambda_i(\lambda_i + h)^{-1}\right| \Big/ V_n(h)^{1/2}(nR_n(h))^{1/2} \to 0.$$

Now compare (2.2′) with (2.1′). By the correspondence of $\sigma^2 - e_i^2$ to e_i, λ_i to μ_i, and $V_n(h)^{1/2}$ to $B_n(h)^{1/2}$, it is clear that (2.2′) holds by similar arguments.

It remains to establish (2.3). Clearly we need only to prove that

$$\text{(2.8)}\qquad \sup_{h\geq 0}\left|\sum_{i=1}^{n} e_i\mu_i\lambda_i(\lambda_i + h)^{-2}\right| \Big/ B_n(h)^{1/2}(nR_n(h))^{1/2} \to 0$$

and that

$$\text{(2.9)}\qquad \sup_{h\geq 0}\left|\sum_{i=1}^{n}\left(\sigma^2 - e_i^2\right)\lambda_i^2(\lambda_i + h)^{-2}\right| \Big/ V_n(h)^{1/2}(nR_n(h))^{1/2} \to 0.$$

The proof of (2.8) will be similar to that of (2.1′). First it is enough to show that for any fixed natural number k, and for $l = 1, 2$,

$$\text{(2.10)}\qquad \sup_{h\geq\lambda_k}\left|\sum_{i\in I_l(k)} e_i\mu_i\lambda_i(\lambda_i + h)^{-2}\right| \Big/ B_n(h)^{1/2}Q_n^{1/2} \to 0,$$

and that for any $\varepsilon > 0$, there exist constants $c_1(\varepsilon)$ and $c_2(\varepsilon)$ such that for $l = 1, 2$,

$$\text{(2.11)}\qquad \begin{aligned} &P\left\{\sup_{j=k,\dots,\bar{n}}\ \sup_{\lambda_{j+1}\leq h\leq\lambda_j}\left|\sum_{i\in I_l(j)} e_i\mu_i\lambda_i(\lambda_i + h)^{-2}\right| \Big/ B_n(h)^{1/2}V_n(h)^{1/2} > \varepsilon\right\} \\ &\quad\leq c_l(\varepsilon)\sum_{j=k}^{\infty} j^{-2}. \end{aligned}$$

Proof of (2.10). For $l = 1$, since $\lambda_i/(\lambda_i + h) < 1$, the proof is exactly the same as in (2.4). For $l = 2$, the analogue of (2.6) is

$$\text{(2.12)}\qquad P\left\{\sup_{h\geq\lambda_k}\left|\sum_{i=k+1}^{n} e_i\mu_i h\lambda_i(\lambda_i + h)^{-2}\right| \Big/ hB_n(h)^{1/2}Q_n^{1/2} > \varepsilon\right\} \to 0.$$

Now the left side of (2.12) does not exceed

$$P\left\{\sup_{h\geq\lambda_k}\left|\sum_{\iota=k+1}^{n} e_\iota\mu_\iota h\lambda_\iota(\lambda_\iota+h)^{-2}\right|\Big/\lambda_k B_n(\lambda_k)^{1/2}Q_n^{1/2}\geq\varepsilon\right\}$$

$$\leq P\left\{\left|\sum_{\iota=k+1}^{n} e_\iota\mu_\iota\lambda_k\lambda_\iota(\lambda_\iota+\lambda_k)^{-2}\right|\Big/\lambda_k B_n(\lambda_k)^{1/2}Q_n^{1/2}>\tfrac{1}{2}\varepsilon\right\}$$

$$+P\left\{\sup_{\lambda_k\leq h}\left|\sum_{i=k+1}^{n} e_i\mu_i\lambda_i\left(h(\lambda_\iota+h)^{-2}-\lambda_k(\lambda_\iota+\lambda_k)^{-2}\right)\right|\right.$$

$$\left.>\tfrac{1}{2}\varepsilon\lambda_k B_n(\lambda_k)^{1/2}Q_n^{1/2}\right\}.$$

Now by Chebyshev's inequality and noting that $\lambda_\iota/(\lambda_\iota+\lambda_k)\leq 1$, the first term does not exceed $4\varepsilon^{-2}Q_n^{-1}\sigma^2\to 0$. The second term can also be shown to be no greater than $4\varepsilon^{-2}Q_n^{-1}\sigma^2$ due to Lemma 2.1. Here we observe that $h(\lambda_\iota+h)^{-2}-\lambda_k(\lambda_\iota+\lambda_k)^{-2}=(\lambda_\iota+\lambda_k)^{-2}(h-\lambda_k)(\lambda_i^2-h\lambda_k)(\lambda_\iota+h)^{-2}$ and that for $i\geq k+1$ and $h\geq\lambda_k$, $(h-\lambda_k)(h\lambda_k-\lambda_\iota^2)(\lambda_\iota+h)^{-2}$ is nondecreasing in i and is no greater than λ_k. Thus setting $W_i=-e_i\mu_\iota\lambda_\iota(\lambda_\iota+\lambda_k)^{-2}$ and $a=\lambda_k$ in Lemma 2.1 we obtain the upper bound $4\lambda_k^2\sigma^2\Sigma_{\iota=k+1}^{n}\mu_\iota^2\lambda_\iota^2(\lambda_\iota+\lambda_k)^{-4}/\varepsilon^2\lambda_k^2B_n(\lambda_k)Q_n$, which is no greater than $4\sigma^2\varepsilon^{-2}Q_n^{-1}$ as desired. This completes the proof of (2.10). □

Proof of (2.11). For $l=1$, the left side of (2.11) does not exceed

$$\sum_{j=k}^{\bar{n}}P\left\{\left|\sum_{\iota=1}^{j}e_\iota\mu_\iota\lambda_i(\lambda_i+\lambda_j)^{-2}\right|\Big/B_n(\lambda_j)^{1/2}V_n(\lambda_j)^{1/2}\geq\tfrac{1}{2}\varepsilon\right\}$$

$$+\sum_{j=k}^{\bar{n}}P\left\{\sup_{\lambda_{j+1}\leq h\leq\lambda_j}\left|\sum_{\iota=1}^{j}e_\iota\mu_\iota\lambda_\iota\left((\lambda_\iota+h)^{-2}-(\lambda_\iota+\lambda_j)^{-2}\right)\right|\Big/\right.$$

$$\left.B_n(\lambda_j)^{1/2}V_n(\lambda_j)^{1/2}\geq\tfrac{1}{2}\varepsilon\right\}.$$

By Chebyshev's inequality and Lemma 2.1 again, both terms in the above expression are bounded by some constant times (2.7). Here we observe that

$$(\lambda_\iota+h)^{-2}-(\lambda_\iota+\lambda_j)^{-2}=(\lambda_\iota+\lambda_j)^{-2}(\lambda_\iota-h)(2\lambda_i+\lambda_j+h)(\lambda_\iota+h)^{-2}$$

and that for $i\leq j$ and $h\leq\lambda_j$, $(\lambda_\iota-h)(2\lambda_\iota+\lambda_j+h)(\lambda_\iota+h)^{-2}$ is nonincreasing in i and is not greater than 3. Put $W_\iota=e_i\mu_\iota\lambda_\iota(\lambda_\iota+\lambda_j)^{-2}$ and $a=3$ to yield the desired bound.

Turning to the case $l = 2$, the left side of (2.11) does not exceed

$$\sum_{j=k}^{\bar{n}} P\left\{\left|\sum_{i=j+1}^{n} e_i\mu_i\lambda_{j+1}\lambda_i(\lambda_i+\lambda_{j+1})^{-2}\right| \Big/ \lambda_{j+1}B_n(\lambda_{j+1})^{1/2}V_n(\lambda_j)^{1/2} \geq \tfrac{1}{2}\varepsilon\right\}$$

$$+ \sum_{j=k}^{\bar{n}} P\left\{\sup_{\lambda_{j+1}\leq h\leq\lambda_j}\left|\sum_{i=j+1}^{n} e_i\mu_i\lambda_i\left(h(\lambda_i+h)^{-2} - \lambda_{j+1}(\lambda_i+\lambda_{j+1})^{-2}\right)\right| \Big/ \right.$$

$$\left. \lambda_{j+1}B_n(\lambda_{j+1})^{1/2}V_n(\lambda_j)^{1/2} \geq \tfrac{1}{2}\varepsilon\right\}.$$

These two terms are both no greater than (2.7). Here note that $W_i = -e_i\mu_i\lambda_i(\lambda_i + \lambda_{j+1})^{-2}$ and $a = \lambda_{j+1}$ when using Lemma 2.1. This completes the proof of (2.11). (2.8) is now established. □

Finally, comparing (2.9) with (2.8), we see that the former can be proved in a similar way. Hence (2.3) is established. The proof of Theorem 1 is now complete.

3. Stein estimates and GCV. Consider the following simplified version of Stein estimates and the associated unbiased risk estimate,

$$\tilde{\mu}_n(h) = \mathbf{y}_n - \sigma^2 \operatorname{tr} A_n(h)\|A_n(h)\mathbf{y}_n\|^{-2}A_n(h)\mathbf{y}_n$$

and

$$\mathrm{SURE}_n(h) = \sigma^2 - \sigma^4(\operatorname{tr} A_n(h))^2/n\|A_n(h)\mathbf{y}_n\|^2$$

where $A_n(h) = I - M_n(h)$. Clearly $\hat{h}_G$ minimizes $\mathrm{SURE}_n(h)$ over $h \geq 0$. Li (1985) has shown that $\mathrm{SURE}_n(\hat{h}_G)$ is a consistent estimate of the true loss $\tilde{L}_n(\hat{h}_G) = n^{-1}\|\mu_n - \tilde{\mu}_n(\hat{h}_G)\|^2$ essentially without any assumptions on the matrix $M_n(h)$ and μ_n. With (A.1) and other conditions to be given, we may strengthen this result.

PROPOSITION 3.1. *Under* (A.1), *for any* $\hat{h}$, *random or not, such that*

$$(3.1) \qquad (n^{-1}\operatorname{tr} M_n(\hat{h}))^2/n^{-1}\operatorname{tr} M_n^2(\hat{h}) \to 0$$

and

$$(3.2) \qquad n^{-1}\|A_n(\hat{h})\mathbf{y}_n\|^2 \to \sigma^2,$$

we have

$$(3.3) \qquad \left|\mathrm{SURE}_n(\hat{h}) - \tilde{L}_n(\hat{h}) - n^{-1}\|\mathbf{e}_n\|^2 + \sigma^2\right|/L_n(\hat{h}) \to 0$$

and

$$(3.4) \qquad n^{-1}\|\tilde{\mu}_n(\hat{h}) - \hat{\mu}_n(\hat{h})\|^2/L_n(\hat{h}) \to 0.$$

PROOF. Rewrite the left side of (3.3) as

$$2\left|\frac{\sigma^2\operatorname{tr}A_n(\hat{h})}{n\|A_n(\hat{h})\mathbf{y}_n\|^2}\langle\mathbf{e}_n, A_n(\hat{h})\mathbf{y}_n\rangle - \frac{\sigma^4(\operatorname{tr}A_n(\hat{h}))^2}{n\|A_n(\hat{h})\mathbf{y}_n\|^2} - n^{-1}\|\mathbf{e}_n\|^2 + \sigma^2\right| \Big/ L_n(\hat{h})$$

$$\le 2\sigma^2\operatorname{tr}A_n(\hat{h})\left|\langle\mathbf{e}_n, A_n(\hat{h})\boldsymbol{\mu}_n\rangle\right| \Big/ n\|A_n(\hat{h})\mathbf{y}_n\|^2 L_n(\hat{h})$$

$$+2\sigma^2\operatorname{tr}A_n(\hat{h})\left|\langle\mathbf{e}_n, M_n(\hat{h})\mathbf{e}_n\rangle - \sigma^2\operatorname{tr}M_n(\hat{h})\right| \Big/ n\|A_n(\hat{h})\mathbf{y}_n\|^2 L_n(\hat{h})$$

$$+2\left|\left(\frac{\sigma^2\operatorname{tr}A_n(\hat{h})}{\|A_n(\hat{h})\mathbf{y}_n\|^2} - 1\right)(\sigma^2 - n^{-1}\|\mathbf{e}_n\|^2)\right| \Big/ L_n(\hat{h}).$$

Now by (2.1)–(2.3) and (3.2), the first two terms of the last expression tend to 0. To show that the third also converges to 0, it is enough to prove

$$\left|\sigma^2 n^{-1}\operatorname{tr}A_n(\hat{h}) - n^{-1}\|A_n(\hat{h})\mathbf{y}_n\|^2\right|\left|\sigma^2 - n^{-1}\|\mathbf{e}_n\|^2\right|/L_n(\hat{h}) \to 0.$$

Since $\|A_n(\hat{h})\mathbf{y}_n\|^2 = \|\mathbf{e}_n\|^2 + 2\langle\mathbf{e}_n, A_n(\hat{h})\boldsymbol{\mu}_n\rangle - 2\langle\mathbf{e}_n, M_n(\hat{h})\mathbf{e}_n\rangle + L_n(\hat{h})$, the first absolute value factor in the last expression does not exceed

$$\left|\sigma^2 - n^{-1}\|\mathbf{e}_n\|^2\right| + L_n(\hat{h}) + 2n^{-1}\left|\langle\mathbf{e}_n, A_n(\hat{h})\boldsymbol{\mu}_n\rangle\right|$$

$$+2n^{-1}\left|\langle\mathbf{e}_n, M_n(\hat{h})\mathbf{e}_n\rangle - \sigma^2\operatorname{tr}M_n(\hat{h})\right| + n^{-1}\sigma^2\operatorname{tr}M_n(\hat{h}).$$

Thus by (2.1) and (2.2) again, it remains to show that

$$(3.5)\qquad (\sigma^2 - n^{-1}\|\mathbf{e}_n\|^2)^2/L_n(\hat{h}) \to 0$$

and

$$(3.6)\qquad (n^{-1}\operatorname{tr}M_n(\hat{h}))\left|\sigma^2 - n^{-1}\|\mathbf{e}_n\|^2\right|/L_n(\hat{h}) \to 0.$$

By the central limit theorem, (A.1), and (2.3), we get (3.5). Finally by (3.5) and (2.3), (3.6) holds because $(n^{-1}\operatorname{tr}M_n(\hat{h}))^2 \le R_n(\hat{h})$. Hence we have proved (3.3). The proof of (3.4) is omitted since it is similar to the proof of (6.7) of Li (1984) (see also Li and Hwang (1984) for the case where $\hat{h}$ is nonrandom). □

(3.2) is equivalent to the consistency of $\hat{\boldsymbol{\mu}}_n(\hat{h})$, i.e.,

$$(3.7)\qquad L_n(\hat{h}) \to 0.$$

This condition may imply (3.1) if we assume the following condition on the asymptotic distribution of the eigenvalues λ_i:

For any m such that $m/n \to 0$, we have

$$(\text{A.2})\qquad \left(\frac{1}{n}\sum_{i=m+1}^{n}\lambda_i\right)^2 \Big/ \frac{1}{n}\sum_{i=m+1}^{n}\lambda_i^2 \to 0.$$

LEMMA 3.1. *Under* (A.1) *and* (A.2), (3.7) *implies* (3.1).

PROOF. Define $\hat{m} = i$ if $\lambda_{i+1} \le \hat{h} \le \lambda_i$. Clearly we have

$$(3.8) \qquad \begin{aligned} 2^{-1}n^{-1}\left[\hat{m} + \sum_{i=\hat{m}+1}^{n} \lambda_i^l/\lambda_{\hat{m}}^l\right] &\le n^{-1}\sum_{i=1}^{n} \left[\lambda_i/(\lambda_i + \hat{h})\right]^l \\ &\le n^{-1}\left[\hat{m} + \sum_{i=\hat{m}+1}^{n} \lambda_i^l/\lambda_{\hat{m}}^l\right] \end{aligned}$$

for $l = 1, 2$. From this it follows that (3.1) is equivalent to

$$(3.9) \qquad \left[n^{-1}\left(\hat{m}\lambda_{\hat{m}} + \sum_{i=\hat{m}+1}^{n} \lambda_i\right)\right]^2 \Bigg/ n^{-1}\left(\hat{m}\lambda_{\hat{m}}^2 + \sum_{i=\hat{m}+1}^{n} \lambda_i^2\right) \to 0.$$

On the other hand due to (2.3), (3.7) implies that $R_n(\hat{h}) \to 0$, which in turn implies $n^{-1}\mathrm{tr}\, M_n^2(\hat{h}) \to 0$. Hence by (3.8), $\hat{m}/n \to 0$. Now it can be seen that (3.9) follows from (A.2). This completes the proof of Lemma 3.1. □

We are ready to prove the following main result of this section.

THEOREM 2. *Assume that* (A.1), (A.2), *and the following condition hold*:

$$(\mathrm{A}.3) \qquad \inf_{h \ge 0} L_n(h) \to 0.$$

Then $\hat{h}_G$ *is a.o. Moreover* $\tilde{L}_n(\hat{h}_G)/L_n(\hat{h}_G) \to 1$.

PROOF. Let h^* be the minimizer of the left side of (A.3). Then by Lemma 3.1 and Proposition 3.1, we have

$$(3.10) \qquad \mathrm{SURE}_n(h^*) - n^{-1}\|\mathbf{e}_n\|^2 + \sigma^2 = L_n(h^*)\left(1 + o_p(1)\right).$$

On the other hand, by Theorem 5.4 of Li (1985), (3.7) holds for $\hat{h} = \hat{h}_G$. Therefore, we also have

$$(3.11) \qquad \mathrm{SURE}_n(\hat{h}_G) - n^{-1}\|\mathbf{e}_n\|^2 + \sigma^2 = L_n(\hat{h}_G)\left(1 + o_p(1)\right).$$

Since $\mathrm{SURE}_n(\hat{h}_G) \le \mathrm{SURE}_n(h^*)$, Theorem 2 is now proved by comparing (3.10) with (3.11). □

We may apply Theorem 2 to the problem of spline smoothing. For instance, if x_i's are equispaced, then Craven and Wahba (1979) showed that $\lambda_i \approx ci^{-2k}$, for some constant c. Now

$$\frac{1}{n}\sum_{i=m}^{n} \lambda_i \approx c\frac{n-m}{n} n^{-2k}\int_{m/n}^{1} x^{-2k}\,dx \approx c(2k-1)^{-1}n^{-1}m^{-2k+1}$$

and

$$\frac{1}{n}\sum_{i=m}^{n} \lambda_i^2 \approx c\frac{n-m}{n} n^{-4k}\int_{0}^{1} x^{-4k}\,dx \approx c(4k-1)^{-1}n^{-1}m^{-4+1}.$$

Hence (A.2) holds. (A.3) is guaranteed by the existence of a consistent estimate of

f. (A.1) will be satisfied unless f is a polynomial of degree $k-1$ or less. Thus we have

COROLLARY. *For the problem of spline smoothing, if f is not a polynomial of degree $k-1$ or less, and x_i's are equispaced, then $\hat{h}_G$ is a.o.*

Finally we give an example to show that violating (A.2) may incur the inefficiency of $\hat{h}_G$.

EXAMPLE. Let

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{[n^{1/2}]} = n^{1/2},$$

$$\lambda_{[n^{1/2}]} = \cdots = \lambda_{n/2} = 2,$$

$$\lambda_{n/2+1} = \cdots = \lambda_n = 1,$$

$$\mu_1 = \cdots = \mu_{[n^{1/2}]} = n^{1/4},$$

and

$$\mu_{[n^{1/2}]+1} = \cdots = \mu_n = 0.$$

For any h such that $h \to \infty$ and $h \ll n^{1/2}$, we have $nR_n(h) \approx \sigma^2(n^{1/2} + 2.5nh^{-2}) + h^2$. Thus (A.1) and (A.3) hold. In fact, $[\inf_{h\geq 0} nR_n(h)]/n^{1/2}(\sigma^2 + 10^{1/2}\sigma) \to 1$, as $n \to \infty$ and the minimizer $h^* \approx (2.5n\sigma^2)^{1/4}$. On the other hand, (1.2) can be written as

$$C_1 D \sum_{\iota=1}^{[n^{1/2}]} y_i^2 + C_2 D \sum_{[n^{1/2}]+1}^{n/2} e_\iota^2 + C_3 D \sum_{n/2+1}^{n} e_\iota^2, \tag{3.12}$$

where

$$C_1 = n(h+n^{1/2})^{-2}, \qquad C_2 = n(h+2)^{-2}, \qquad C_3 = n(h+1)^{-2},$$

and

$$D = \left(n^{1/2}(h+n^{1/2})^{-1} + \left(\tfrac{1}{2}n - n^{1/2}\right)(h+2)^{-1} + \tfrac{1}{2}n(h+1)^{-1}\right)^{-2}.$$

Now using Taylor's expansion, for h such that $h \to \infty$ and $h \ll n^{1/2}$, we have

$$C_1 D = n^{-2}h^2 + o(n^{-2}h^2),$$

$$C_2 D = n^{-1}\left(1 - h^{-1} + \tfrac{7}{4}h^{-2} + 2n^{-1/2} + o(n^{-1/2} + h^{-2})\right),$$

and

$$C_3 D = n^{-1}\left(1 + h^{-1} - \tfrac{5}{4}h^{-2} + 2n^{-1/2} + o(h^{-2} + n^{-1/2})\right).$$

Substituting into (3.12), we obtain the leading terms

$$n^{-2}h^2 \sum_{i=1}^{n^{1/2}} y_i^2 + n^{-1}h^{-2}\left(\tfrac{7}{4}\sum_{n^{1/2}}^{n/2} e_i^2 - \tfrac{5}{4}\sum_{n/2}^{n} e_i^2\right) + n^{-1}h^{-1}\left(\sum_{n/2}^{n} e_i^2 - \sum_{n^{1/2}}^{n/2} e_i^2\right)$$

$$+\left(\frac{1}{n}\sum_{n^{1/2}}^{n} e_i^2\right)(1 + 2n^{-1/2}) \approx n^{-1}h^2 + \frac{1}{4}h^{-2}\sigma^2 + \left(\frac{1}{n}\sum_{n^{1/2}}^{n} e_i^2\right)(1 + 2n^{-1/2}).$$

Thus $\hat{h}_G \approx (4^{-1}n\sigma^2)^{1/4}$. Compared with h^*, we see that $\hat{h}_G$ is not a.o. Note that the condition (5.6) of Li (1985) is satisfied and hence $\hat{h}_G$ is consistent.

REFERENCES

Cox, D. D. (1983). Personal communication.

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

Erdal, A. (1983). Cross validation for ridge regression and principal component analysis. Thesis, Div. of Applied Mathematics, Brown Univ.

Golub, G., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

Li, K. C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation. *Ann. Statist.* **13** 1352–1377.

Li, K. C. (1984). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Unpublished.

Li, K. C. and Hwang, J. (1984). The data smoothing aspect of Stein estimates. *Ann. Statist.* **12** 887–897.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics* **15** 661–675.

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

Speckman, P. (1982). Efficient nonparametric regression with cross-validated smoothing splines. Unpublished.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

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.

Department of Mathematics
University of California
Los Angeles, California 90024