*The Annals of Statistics*
1985, Vol. 13, No. 4, 1352–1377

# FROM STEIN'S UNBIASED RISK ESTIMATES TO THE METHOD OF GENERALIZED CROSS VALIDATION

BY KER-CHAU LI[1]

*University of California, Los Angeles*

Dedicated to the memory of Jack Kiefer–advisor, teacher, and, above all, friend.

This paper concerns the method of generalized cross validation (GCV), a promising way of choosing between linear estimates. Based on Stein estimates and the associated unbiased risk estimates (Stein, 1981), a new approach to GCV is developed. Many consistency results are obtained for the cross-validated (Steinized) estimates in the contexts of nearest-neighbor nonparametric regression, model selection, ridge regression, and smoothing splines. Moreover, the associated Stein's unbiased risk estimate is shown to be uniformly consistent in assessing the true loss (not the risk). Consistency properties are examined as well when the sampling error is unknown. Finally, we propose a variant of GCV to handle the case that the dimension of the raw data is known to be greater than that of their expected values.

**1. Introduction.** We consider the problem of choosing a good estimator from those being tentatively proposed. In this selection process, it is desirable to let the data speak for themselves. The generalized cross-validation (GCV) method of Craven and Wahba (1979) is one of many promising data-driven techniques of selection. While the extension to the choice among nonlinear estimators has been underway (for example, Wahba, 1982), we shall nevertheless focus our study on linear ones.

Specifically, let $y_1, y_2, \dots, y_n$ be $n$ independent observations with unknown means $\mu_1, \mu_2, \dots, \mu_n$. Write

$$y_i = \mu_i + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1.1}$$

and assume that $\varepsilon_i$ has mean 0 and common variance $\sigma^2$. To estimate $\boldsymbol{\mu}_n = (\mu_1, \dots, \mu_n)'$, a class of linear estimators $\hat{\boldsymbol{\mu}}_n(h)$, indexed by $h$, is proposed. Let $H_n$ be the index set and $M_n(h)$ be the $n \times n$ matrix associated with $\hat{\boldsymbol{\mu}}_n(h)$ such that $\hat{\boldsymbol{\mu}}_n(h) = M_n(h)\mathbf{y}_n$ where $\mathbf{y}_n = (y_1, \dots, y_n)'$. GCV chooses $h$ by minimizing the quantity

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

Received September 1983; revised December 1984 and July 1985.

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

*AMS* 1980 *subject classifications*. Primary 62G99, 62J99; secondary 62J05, 62J07.

*Key words and phrases*. Cross validation; generalized cross validation; model selection; nearest-neighbor estimates; ridge regression; smoothing spline; Stein estimates; Stein's unbiased risk estimates.

Here $\|\cdot\|$ is the Euclidean norm of $R^n$ and tr denotes trace. The $h$ which is chosen according to GCV is written as $\hat{h}$.

EXAMPLE 1. Periodic curve and moving averages. Suppose $\mu_i = f(x_i)$ for an unknown continuous function on $[0, 1]$ with $f(0) = f(1)$ and $0 \leq x_1 < x_2 < \cdots < x_n < 1$. Due to the continuity of $f$, it is reasonable to estimate $\mu_i$ by $(2h+1)^{-1}\sum_{j=-h}^{h} y_{i+j}$, for some $h \leq (n-1)/2$. (Here we identify $y_{-j}$ with $y_{n-j}$.) It is clear that the rows of $M_n(h)$ must be permutations of the row vector $(2h+1)^{-1}(1, 1, \dots, 1, 0, 0, \dots, 0)$ having $2h+1$ nonzero entries. In this case (1.2) reduces to

$$\mathrm{GCV}_n(h) = (2h+1)^2(2h)^{-2}n^{-1}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2.$$

EXAMPLE 2. Model selection. Associated with each $y_i$ there are $p_n$ explanatory variables $x_{i1}, x_{i2}, \dots, x_{ip_n}$, arranged in the decreasing order of importance. To estimate $\boldsymbol{\mu}_n$ one may employ the first $h$ variables to form a linear model $y_i = \sum_{j=1}^{h} x_{ij}\beta_j + \varepsilon_i$ with $\beta_j$ being unknown parameters, and then use the least-squares estimator

$$\hat{\boldsymbol{\mu}}_n(h) = X_h(X_h'X_h)^{-1}X_h'\mathbf{y}_n, \tag{1.3}$$

where $X_h$ is the $n \times h$ design matrix. Now $M_n(h) = X_h(X_h'X_h)^{-1}X_h'$ is a projection matrix with rank $h$ and (1.2) becomes

$$\mathrm{GCV}_n(h) = n(n-h)^{-2}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2. \tag{1.4}$$

EXAMPLE 3. Ridge regression. Consider the regression model

$$y_i = \sum_{j=1}^{p_n} x_{ij}\beta_j + \varepsilon_i, \qquad i = 1, 2, \dots, n,$$

with the $n \times p_n$ design matrix $X = (x_{ij})$. If $I_{p_n}$ is the $p_n \times p_n$ identity matrix, the ridge-regression estimate of $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{p_n})'$ is $(X'X + hI_{p_n})^{-1}X'\mathbf{y}_n$ is estimated by

$$\hat{\boldsymbol{\mu}}_n(h) = X\left(X'X + hI_{p_n}\right)^{-1}X'\mathbf{y}_n. \tag{1.5}$$

Here the ridge parameter $h$ is a nonnegative number to be chosen. The trace of $M_n(h) = X(X'X + hI_{p_n})^{-1}X'$ may be obtained by the singular value decomposition (Golub and Reinsch, 1970). In particular, let $X = UDV$ where $U$ and $V$ are orthogonal matrices with ranks $n$ and $p_n$, respectively, and $D$ is an $n \times p_n$ matrix with the nonnegative and nonincreasing entries $d_{ii} = \lambda_{i,n}^{1/2}$, $i = 1, 2, \dots, \min\{n, p_n\}$, and with all other entries zero. A straightforward manipulation yields $\operatorname{tr} M_n(h) = \sum_{i=1}^{p_n}\lambda_{i,n}(h + \lambda_{i,n})^{-1}$ and

$$\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2 = \sum_{i=1}^{n} h^2(h + \lambda_{i,n})^{-2}\bar{y}_i^2 \quad \text{where } \bar{\mathbf{y}}_n = (\bar{y}_1, \dots, \bar{y}_n)' = U'\mathbf{y}_n.$$

(1.2) becomes

$$\mathrm{GCV}_n(h) = \left[ n^{-1} \sum_{i=1}^{n} (h + \lambda_{i,n})^{-2} \tilde{y}_i^2 \right] \Big/ \left[ n^{-1} \sum_{i=1}^{n} (h + \lambda_{i,n})^{-1} \right]^2,$$

where we write $\lambda_{i,n} = 0$ for $i = p_n + 1, \ldots, n$. We shall conveniently take $H_n = \{h\colon 0 \le h \le \infty\}$ without any difficulties in defining any quantities associated with the extreme cases $h = 0$ and $\infty$ by continuity.

EXAMPLE 4. Smoothing splines. Suppose $\mu_i = f(x_i)$ with $f \in W_2^k[0,1] = \{f\colon f$ has absolutely continuous derivatives, $f, f', \ldots, f^{(k-1)}$, and $\int_0^1 f^{(k)}(x)^2\,dx < \infty\}$, $x_i \in [0,1]$. The smoothing spline $\hat{f}_h$ is the solution of

$$\min_{f \in W_2^k[0,1]} n^{-1} \sum_{i=1}^{n} (y_i - f(x_i))^2 + h \int_0^1 f^{(k)}(x)^2\,dx. \tag{1.6}$$

Here the smoothing parameter $h$ is a nonnegative number to be chosen. $\hat{f}_h$ is well known to be linear in the $y_i$s and the matrix $M_n(h)$ such that $\hat{\boldsymbol{\mu}}_n = (\hat{f}_h(x_1), \ldots, \hat{f}_h(x_n))' = M_n(h)\mathbf{y}_n$ has been studied extensively (Reinsch 1967; Demmler and Reinsch 1975; Wahba 1975, 1978; Craven and Wahba 1979; Speckman 1981, 1982, 1985). To implement GCV one may either employ the fast algorithm of Utreras (1979, 1980) or carry out a singular value decomposition (Craven and Wahba, 1979). More recently Silverman (1984) has developed a new algorithm.

GCV was first proposed to choose the smoothing parameter for spline smoothing (Craven and Wahba, 1979); then applications were extended to the problems of selecting the ridge parameter, choosing a model, and many others (Golub, Heath, and Wahba, 1979) (but not including Example 1). Until now, the most useful and persuasive arguments for GCV are two theorems in Golub, Heath, and Wahba (1979): The first one compares the expected value of (1.2) with the mean-squared error for the linear estimator $\hat{\boldsymbol{\mu}}_n(h)$; the second one justifies GCV from a Bayesian viewpoint. In fact, certain types of asymptotic optimality have been obtained for GCV spline smoothing (Craven and Wahba, 1979; Speckman, 1982). These theoretical results, together with numerical evidence from simulation and real data, are very encouraging, but other issues remain. For instance, is GCV consistent in the sense that $n^{-1}\|\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(\hat{h})\|^2 \to 0$ in probability? Note that Craven and Wahba's Theorem does not lead to an answer because the asymptotic optimality result was only established for the $h$ selected by minimizing the expected value of (1.2), not for $\hat{h}$. The stronger result of Speckman dealt with $\hat{h}$, but he restricted $H_n$ to be a bounded interval that converges to 0 in some fashion. It is also unclear how to extend these results to other settings, although the formal application (1.2) is always possible [e.g., Rice (1983), in the context of kernel nonparametric regression]. In addition, maybe the most puzzling feature about GCV is that this selection needs no information about the sampling variance $\sigma^2$. Consequently for a given data set, GCV always selects the same $h$ (hence yielding the same degree of smoothness in smoothing spline), no matter whether we are told that the magnitude of noise $\sigma$ is 100 or is just 0.01.

Our main goal in this paper is to develop a new approach to GCV under the general setting of selecting a linear estimate. On the heuristic level, no restriction on the structure of the problem is necessary, although to obtain rigorous results case-by-case treatments may be unavoidable.

Section 2 gives the motivation for our approach. Briefly, instead of working directly on the linear estimates $\hat{\mu}_2(h)$, we consider the associated Stein estimates (Stein, 1981). This replacement will be justified in terms of efficiency and model robustness. After some simplification, we shall see that minimizing the corresponding unbiased risk estimates (for Stein estimates) yields the GCV. The surprising aspect of this is that Stein estimates and their unbiased risk estimates depend on $\sigma^2$ while GCV does not. But this helps us explain the puzzling feature about GCV mentioned before.

Section 3 studies the asymptotic behavior of a simplified version of Stein's unbiased risk estimate (SURE hereafter). It turns out that SURE does more than anticipated: It consistently estimates the *true squared error loss*, not the risk, for the corresponding Stein estimate. While it may be inconsistent for estimating the risk, it will be the true loss that really concerns us. Two important features about this consistency result are that it does not depend on how to embed a particular SURE into a sequence and that the consistency is uniform over $\mu_n$ in $R^n$.

In Section 4, we shall establish the consistency for the Stein estimate selected by GCV. This is carried out case-by-case for nearest-neighbor nonparametric regression, model selection, and ridge regression including spline smoothing. The key step that keeps us from obtaining a unified proof lies in verifying that SURE is consistent, uniformly for $h$ in $H_n$. But it will be clear that the main idea should easily carry over to other situations.

Stein estimates require knowledge of $\sigma^2$. If this is unknown, we may replace it by a suitable estimate. Or we may return to the original linear estimate. Both procedures will be consistent under appropriate conditions and this will be demonstrated in Section 5.

From Sections 2–5, we implicitly assume that $\mu_n$ could be any vector in $R^n$. Section 6 discusses a natural generalization of our approach to the case that $\mu_n$ is known to be in a proper linear subspace of $R^n$. This leads to a selection procedure different from GCV. All the technical proofs will be given in Section 7.

After this work was complete, the author learned that the consistency result for GCV in the context of ridge regression was independently obtained by Erdal (1983) using arguments different from ours. Assumptions imposed there were: (i) the number of parameters $p_n$ is fixed; (ii) the maximum and minimum eigenvalues of the information matrix must grow to infinity at the same rate $O(n)$. Under these conditions the problem becomes well posed and ridge regression seems unnecessary. Erdal also obtained results for principal component analysis under restrictions of a similar nature.

To close this section, we remark that many authors have realized one way or another that the shrinkage phenomena of Stein estimates are relevant in choosing estimates. However none of them directly employ the Stein estimates as we do here. In particular, in ridge regression our approach is completely different from others [i.e., Casella (1980) and the references given there] which are aimed at the minimax estimation.

**2. Heuristics.** We start with (1.1) and temporarily assume the normality of $\varepsilon_i$s. Consider first the case where $M_n(h)$ is symmetric. Define the Stein estimate associated with $\hat{\mu}_n(h)$,

$$\tilde{\mu}_n^0(h) = \mathbf{y}_n - \frac{\sigma^2}{\mathbf{y}_n' B_n(h)\mathbf{y}_n} A_n(h)\mathbf{y}_n, \tag{2.1}$$

where $A_n(h) = I_n - M_n(h)$ and

$$B_n(h) = (\operatorname{tr} A_n(h) \cdot I_n - 2A_n(h))^{-1} A_n(h)^2. \tag{2.2}$$

Here the largest characteristic root of $A_n(h)$ is assumed to be less than half of the trace of $A_n(h)$. Stein (1981) showed that $\tilde{\mu}_n^0(h)$ dominates $\mathbf{y}_n$ under the usual squared error loss. The relationship between $\tilde{\mu}_n^0(h)$ and $\hat{\mu}_n(h)$ was studied from an asymptotic viewpoint by Li and Hwang (1984). A result there will be useful for our development.

THEOREM 2.1. *For any sequence $\{h_n\}$ such that $\hat{\mu}_n(h_n)$ is consistent in the mean square sense,*

$$En^{-1}\|\mu_n - \hat{\mu}_n(h_n)\|^2 \to 0 \quad \text{as } n \to \infty, \tag{2.3}$$

*the associated Stein estimator $\tilde{\mu}_n^0(h_n)$ is also consistent,*

$$n^{-1}\|\tilde{\mu}_n^0(h_n) - \mu_n\|^2 \to 0 \quad \text{in probability.} \tag{2.4}$$

Moreover, they proved that the convergence rate of (2.4) is no slower than that of (2.3) except for the pathological case that the latter is faster than $n^{-1}$. Under the additional condition that

$$n^{-1}\operatorname{tr} M_n^2(h_n)/\left(n^{-1}\operatorname{tr} M_n(h_n)\right)^2 \to \infty, \tag{2.5}$$

$\tilde{\mu}_n^0(h_n)$ and $\hat{\mu}(h_n)$ will be asymptotically indistinguishable in the sense that $\|\tilde{\mu}_n^0(h_n) - \hat{\mu}_n(h_n)\|^2/E\|\hat{\mu}_n(h_n) - \mu_n\|^2 \to 0$ in probability. Condition (2.5) is frequently satisfied by good estimates (see Golub, Heath, and Wahba, or Li and Hwang).

These results justify the replacement of the original linear class $\{\hat{\mu}_n(h): h \in H_n\}$ by the Stein class $\{\tilde{\mu}_n^0(h): h \in H_n\}$. Asymptotically, good estimates in the new class remain as good as before. But for any finite sample size, Stein estimates enjoy an additional property: They have bounded risk. Another aspect of Stein estimates can be viewed as model robustness and is well illustrated by Example 2. If $\hat{\mu}_n(h)$ is good (the case that model $h$ is appropriate), then the shrinkage factor $\sigma^2/\mathbf{y}_n' B_n(h)\mathbf{y}_n$ should be close to 1 and $\tilde{\mu}_n^0(h)$ would be about the same as $\hat{\mu}_n(h)$. Otherwise, $\tilde{\mu}_n^0(h)$ shrinks $\hat{\mu}_n(h)$ toward the raw data to guard against model violation. See Huber (1975) for a careful distinction between model robustness and distributional robustness.

Now define

$$\mathrm{SURE}_n^0(h) = \sigma^2 - \frac{\sigma^4 \|A_n(h)\mathbf{y}_n\|^2}{n(\mathbf{y}_n' B_n(h)\mathbf{y}_n)^2}.$$

Stein showed that $\mathrm{SURE}_n^0(h)$ is an unbiased estimate for the risk of $\tilde{\mu}_n^0(h)$; $E\ \mathrm{SURE}_n^0(h) = En^{-1}\|\mu_n - \tilde{\mu}_n^0(h)\|^2$ for any $\mu_n \in R^n$. To select a good $h$, it is natural to minimize $\mathrm{SURE}_n^0(h)$ over $h \in H_n$, and is equivalent to minimizing

$$n(\mathbf{y}_n' B_n(h)\mathbf{y}_n)^2 / \|A_n(h)\mathbf{y}_n\|^2. \tag{2.6}$$

Suppose $n$ is large enough so that the largest eigenvalue of $A_n(h)$ is negligible compared to the trace. Then we may approximate $B_n(h)$ by $(\operatorname{tr} A_n(h))^{-1} \cdot A_n^2(h)$. By substituting this quantity into (2.6), $\mathrm{GCV}_n(h)$ of (1.2) is obtained!

For a general $M_n(h)$, Li and Hwang (1984) replaced (2.2) by

$$B_n(h) = (\operatorname{tr} A_n(h) - 2\lambda(A_n(h)))^{-1} A_n(h)' A_n(h),$$

where $\lambda(A_n(h))$ denotes the maximum eigenvalue of $\frac{1}{2}(A_n(h)' + A_n(h))$. They showed that all the desired properties we discussed above are preserved. The corresponding Stein's unbiased risk estimates have a complicated form but they can be simplified. This again leads to GCV.

REMARK 1. There is another simple way to derive GCV by means of Mallows' $C_L$ statistics (Mallows, 1973) and the notion of nil-trace estimate. Briefly, consider $\bar{\mu}_n(h) = -\alpha \mathbf{y}_n + (1+\alpha)\hat{\mu}_n(h)$ with $\alpha = \operatorname{tr} M_n(h)/(n - \operatorname{tr} M_n(h))$. The matrix associated with $\bar{\mu}_n(h)$, $-\alpha I_n + (1+\alpha)M_n(h)$, has trace 0. Now using $C_L$ procedure to select an estimate from the class $\{\bar{\mu}_n(h)\colon h \in H_n\}$ amounts to choosing $h$ by minimizing $n^{-1}\|\mathbf{y}_n - \bar{\mu}_n(h)\|^2 = \mathrm{GCV}_n(h)$! A counterpart of Theorem 2.1 for $\bar{\mu}_n(h)$ was demonstrated in Li (1983), justifying the replacement of $\mu_n(h)$ by $\bar{\mu}_n(h)$ for large sample sizes. The connection between cross validation (Stone, 1974; Geisser, 1975) and $C_L$ can also be drawn by a similar argument.

## 3. Estimating the true loss, not the risk!

From now on, only the following simplified version of Stein estimates and SURE will be considered:

$$\tilde{\mu}_n(h) = \mathbf{y}_n - \frac{\sigma^2 \operatorname{tr} A_n(h)}{\|A_n(h)\mathbf{y}_n\|^2} A_n(h)\mathbf{y}_n, \tag{3.1}$$

$$\mathrm{SURE}_n(h) = \sigma^2 - \frac{\sigma^4 (\operatorname{tr} A_n(h))^2}{n\|A_n(h)\mathbf{y}_n\|^2}. \tag{3.2}$$

It is clear that minimizing (3.2) is exactly the same as the procedure of GCV. In this section we shall show that $\mathrm{SURE}_n(h)$ is a consistent estimate of the true loss $n^{-1}\|\tilde{\mu}_n(h) - \mu_n\|^2$ for each $h \in H_n$. The consistency will be uniform over $\mu_n \in R^n$ but not over $h \in H_n$. To obtain uniformity over both, we need more information about the cardinality of $H_n$ or some topological properties about $M_n(h)$ as a function of $h$. Case-by-case treatment is easier to pursue and will appear in Section 4.

For brevity we shall omit the index $h$ and assume that $M_n$ is not an identity matrix in this section.

THEOREM 3.1. *Assume that*

(A.1) *The fourth moments of* $\varepsilon_i$s *are bounded by a constant* $m$;

(A.2) *There exists a constant* $K$ *such that for any* $a \geq 0$, *we have*

$$\sup_{x \in R} P\{x - a \leq \varepsilon_i \leq x + a\} \leq Ka \quad \textit{for any } i.$$

*Then for any* $\delta > 0$ *we have*

$$\sup_{\mu_n \in R^n} P\left\{\left|\mathrm{SURE}_n - n^{-1}\|\tilde{\mu}_n - \mu_n\|^2\right| \geq \delta\right\} \to 0 \quad \textit{as } n \to \infty. \tag{3.3}$$

Looking at the forms of (3.1) and (3.2), it is clear that if $\|A_n y_n\|^2$ takes too small values, then $\tilde{\mu}_n$ and $\mathrm{SURE}_n$ will not be good estimates. (A.2) is simply made to monitor the chance that this will happen. It can be easily satisfied, for instance, by assuming that $\varepsilon_i$s have a common bounded density. On the other hand, it seems possible to avoid this assumption by modifying (3.1) and (3.2) a little bit; for instance, by adding a positive constant to the denominators there. Note that no assumptions about the matrix $M_n$ are required here. Roughly speaking, the boundedness of the risks of Stein estimates makes this theorem plausible. The same result does not seem likely to hold for linear estimates together with their unbiased risk estimates.

The most interesting phenomenon implied by this theorem is that SURE may not be a consistent estimate of the risk but it always estimates the true loss consistently. This is illustrated by the following example.

EXAMPLE 5. Take $A_n = \mathrm{diag}(1, n^{-1/2}, n^{-1/2}, \dots, n^{-1/2})$, and $\mu_n = (0, 0, \dots, 0)'$. Then

$$\mathrm{SURE}_n = \sigma^2 - \frac{\sigma^4\left(1 + (n-1)n^{-1/2}\right)^2}{n\left(\varepsilon_1^2 + n^{-1}\sum_{i=2}^n \varepsilon_i^2\right)},$$

which tends to $\sigma^2\varepsilon_1^2(\varepsilon_1^2 + \sigma^2)^{-1}$ (a random variable!)) as $n \to \infty$. Hence $\mathrm{SURE}_n$ cannot be a consistent estimate of the risk $En^{-1}\|\tilde{\mu}_n - \mu_n\|^2$ since the risk is a nonrandom number. On the other hand, Theorem 3.1 shows that $\mathrm{SURE}_n$ estimates the true loss $n^{-1}\|\tilde{\mu}_n - \mu_n\|^2$ consistently.

**4. Consistency results for Stein estimate selected by GCV.** In this section we shall show that for many cases (including Examples 1–4), the Stein estimate selected by minimizing $\mathrm{SURE}_n(h)$ over $h \in H_n$ is consistent:

$$n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2 \to 0 \quad \text{in probability}. \tag{4.1}$$

The crucial step will be to establish the uniform consistency of SURE over both

$\mu_n \in R^n$ and $h \in H_n$: For any $\delta > 0$,

$$\sup_{\mu_n \in R^n} P\left\{ \sup_{h \in H_n} |\mathrm{SURE}_n(h) - n^{-1}\|\tilde{\mu}_n(h) - \mu_n\|^2| > \delta \right\} \to 0. \tag{4.2}$$

THEOREM 4.1. *Assume that there exists a deterministically chosen sequence* $h_n \in H_n$ *such that* (2.3) *holds. Then* (4.2) *implies* (4.1).

PROOF. First, by an argument similar to the proof of Theorem 2.1 [given in Li and Hwang (1984)], we may easily show that (2.3) implies $n^{-1}\|\tilde{\mu}_n(h_n) - \mu_n\|^2 \to 0$ in probability. Then from (4.2), we have $n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2 = \mathrm{SURE}_n(\hat{h}) + o_p(1) \leq \mathrm{SURE}_n(h_n) + o_p(1) = n^{-1}\|\tilde{\mu}_n(h_n) - \mu_n\|^2 + o_p(1) = o_p(1)$, yielding (4.1). □

Condition (2.3) simply says that the given class of linear estimates contains at least one good estimate and should be satisfied in most cases. (4.2) is an extension of (3.3), to be established case by case in the following subsections.

REMARK 2. (4.2) implies that $\mathrm{SURE}_n(\hat{h})$ is a uniformly consistent estimate of $n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2$, in the sense that

$$\sup_{\mu_n \in R^n} P\{|\mathrm{SURE}_n(\hat{h}) - n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2| > \delta\} \to 0,$$

for any $\delta > 0$. This leads to a valid conservative confidence set for $\mu_n$ whose size, as measured by $n^{-1}$ times the squared radius, will tend to 0 whenever (2.3) holds. For details, see Li (1983).

REMARK 3. The uniform consistency of $\mathrm{SURE}_n(\hat{h})$ implies that for any $\delta > 0$,

$$\inf_{\mu_n \in R^n} P\left\{ \inf_{h \in H_n} \mathrm{GCV}_n(h) \geq (1 - \delta)\sigma^2 \right\} \to 1.$$

To see this we simply observe that since $n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2$ is nonnegative, the consistency of $\mathrm{SURE}_n(\hat{h})$ implies $P\{\mathrm{SURE}_n(\hat{h}) \leq -\delta\} \to 0$, which in turn yields the desired result.

4.1. *Bounded* $\#H_n$. The following result follows immediately from Theorem 3.1.

THEOREM 4.2. *Under the assumptions of Theorem* 3.1, *if* $\sup\{\#H_n\colon n = 1, 2, \dots\} < \infty$. *Then* (4.2) *holds.*

4.2. *Finite* $\#H_n$. Consider the case that $\#H_n$ is finite but may be unbounded. Instead of (A.1), we assume the following stronger moment condition:

(A.1′) $\epsilon_i$s have mean 0, common second, fourth, and sixth moments, and their eighth moments are bounded by a constant $m$.

The following theorem will be useful in verifying (4.2). Let $\lambda'(A_n(h))$ denote the maximum singular value of $A_n(h)$.

THEOREM 4.3. *Under* (A.1′), *for any* $\delta > 0$ *there exist positive numbers* $C_1$ *and* $C_2$ (*depending on m only*) *such that for any* $\mu_n \in R^n$,

$$P\Big\{\sup_{h\in H_n} |\mathrm{SURE}_n(h) - n^{-1}\|\tilde{\mu}_n(h) - \mu_n\|^2| \geq 2\delta\Big\}$$

$$\leq P\{|n^{-1}\|\varepsilon_n\|^2 - \sigma^2| \geq \delta\} + C_1 n^{-2}\# H_n \tag{4.3}$$

$$+ C_2 \sum_{h\in H_n} [\lambda'(A_n(h))]^4 (\operatorname{tr} A_n'(h) A_n(h))^{-2}.$$

*4.2a. Nearest neighbor estimates in nonparametric regression.* Let $p$ be a natural number and $\chi$ be the compact closure of an open set in $R^p$. Suppose $y_1, y_2, \dots, y_n$ are observed at levels $x_1, x_2, \dots, x_n \in \chi$ with $x_i \neq x_j$ for $i \neq j$ such that the expected value $\mu_i$ of $y_i$ is equal to $f(x_i)$ for an unknown continuous function of $f$ on $\chi$. Let $x_{i(j)}$ denote the $j$th nearest neighbor of $x_i$ in the sense that $\|x_i - x_{i(j)}\|$ is the $j$th smallest number among the $n$ values $\|x_i - x_{i'}\|$, $i' = 1, 2, \dots, n$. Ties may be broken in any systematic manner. Take $H_n = \{1, 2, \dots, n\}$. For any $h \in H_n$, consider $\hat{\mu}_n(h)$, the $h$ nearest-neighbor estimate of $\mu_n$, with the $i$th coordinate given by $\sum_{j=1}^{h} w_{n,h}(j) y_{i(j)}$. Here $w_{n,h}(\cdot)$ is a nonnegative weight function such that

$$\sum_{i=1}^{h} w_{n,h}(i) = 1. \tag{4.4}$$

Each row of $M_n(h)$ is a permutation of $(w_{n,h}(1), \dots, w_{n,h}(h), 0, \dots, 0)$ and the diagonal elements are all equal to $w_{n,h}(1)$. To ensure (2.3), we assume that

$$\text{there exists a sequence } \{h_n\} \text{ such that } h_n/n \to 0 \text{ and } w_{n,h_n}(1) \to 0 \text{ as } n \to \infty. \tag{4.5}$$

It can be easily verified that (4.5) implies the consistency of $\hat{\mu}_n(h_n)$ [see, e.g., Li (1984a)]. Stone (1977) gives more general consistency results for nearest-neighbor nonparametric regression.

In addition to (4.4) and (4.5), we need two mild restrictions:

$$\text{there exists a positive number } \delta' \text{ such that } w_{n,h}(1) \leq 1 - \delta' \text{ for any } n, h \geq 2. \tag{4.6}$$

$$\text{for any } n, h, \text{and } i, w_{n,h}(i) \geq w_{n,h}(i+1). \tag{4.7}$$

LEMMA 4.1. *Under* (4.6) *and* (4.7), *there exists a constant* $\Lambda$ (*depending on the dimension p only*) *such that* $\lambda'(M_n(h)) \leq \Lambda$ *for any n and h.*

We may take, for instance, $\Lambda = \sqrt{2}$ for $p = 1$ and $\Lambda = \sqrt{6}$ for $p = 2$. From Lemma 4.1 it follows that $\lambda'(A_n(h)) \leq (1 + \Lambda)$. From this and the observation that $\operatorname{tr} A_n'(h) A_n(h) \geq n(1 - w_{n,h}(1))^2 \geq n(1 - \delta')^2$, it follows that the last term in (4.3) does not exceed $C_2(1 + \Lambda)^4 n^{-1}(1 - \delta')^{-4}$, tending to 0 as $n \to \infty$. (4.2) is now established. Hence by Theorem 4.1, we obtain the following.

THEOREM 4.4. *Under* (4.1′) *and* (4.4)–(4.7), *the Steinized nearest-neighbor estimate* $\tilde{\mu}_n(\hat{h})$ *with* $\hat{h}$ *chosen by GCV is consistent. Moreover, the associated*

*Stein unbiased risk estimate* $SURE_n(\hat{h})$ *is a uniformly consistent estimate of the true loss* $n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2$.

4.2b. *Model selection.* Consider Example 2 without the restriction that models to be selected are nested. In general, let $H_n$ denote a class of models. Associated with each $h$ in $H_n$ is a design matrix $X_h$ with $d(h)$ columns corresponding to $d(h)$ explanatory variables. Assume that $X_h'X_h$ is nonsingular. Consider the least-squares estimate $\hat{\mu}_n(h)$ of (1.3) and its Steinized version $\tilde{\mu}_n(h)$. With $d(h) = n$, define $\tilde{\mu}_n(h) = \mathbf{y}_n$ and $\mathrm{SURE}_n(h) = \sigma^2$. Here in advocating $\tilde{\mu}_n(h)$, we implicitly assume that none of the models with ranks less than $n$ are completely appropriate. Otherwise, we shall proceed differently; see Section 6 for details. Also we do not require that $p_n$ be finite since infinitely many parameter models can be useful sometimes (e.g., Shibata, 1981; Li, 1984b).

Since $A_n(h)$ is a projection with rank $n - d(h)$, the last term in (4.3) equals $C_2\sum_{h \in H_n}(n - d(h))^{-2}$, which may not tend to 0 asymptotically if the number of parameters $d(h)$ is too close to $n$ for some model $h$. But when there are not too many such models in $H_n$, this difficulty can be circumvented, using Theorem 4.2. This leads to the following theorem.

THEOREM 4.5. *Assume that* (A.1′) *and* (A.2) *hold,* $\#H_n/n^2 \to 0$ *as* $n \to \infty$, *and that*

(4.8) *for any positive number* $\varepsilon$, *there exists a natural number* $k$ *such that for any* $n$, *we can find a subset* $H_n' \subset H_n$ *with cardinality no greater than* $k$ *so that* $\sum_{h \notin H_n'}(n - d(h))^{-2} \le \varepsilon$.

*Then* $SURE_n(\hat{h})$ *with* $\hat{h}$ *chosen by GCV is uniformly consistent. Furthermore* $\tilde{\mu}_n(\hat{h})$ *is consistent whenever given* $\mu_n$, *there exists a sequence of models* $\{h_n \in H_n\}$, *such that the least-squares estimate* $\hat{\mu}_n(h_n)$ *is consistent.*

EXAMPLE 2 (cont.). In this case, $\#H_n = p_n \le n$. Put $H_n' = \{n, n-1, \dots, n-k-1\} \cap H_n$. Then $\sum_{h \notin H_n'}(n - d(h))^{-2} \le \sum_{i=k}^{n} i^{-2}$. Since $\sum_{i=1}^{\infty} i^{-2}$ converges, it is easy to see that (4.8) can be satisfied for a suitably chosen $k$. Hence Theorem 4.5 applies here.

4.3. *Continuous* $H_n$. Two cases for $H_n = \{h\colon h \ge 0\}$ will be considered: ridge regression and smoothing splines. In fact, the results for smoothing splines follow immediately from those for ridge regression.

4.3a. *Ridge regression.* Consider the Steinized ridge-regression estimate $\tilde{\mu}_n(h)$ associated with $\hat{\mu}_n(h)$ of (1.5). Here $\tilde{\mu}_n(0)$ is defined by $\lim_{h \to 0}\tilde{\mu}_n(h)$ ($\neq \mathbf{y}_n$ unless all $\lambda_{i,n}$, $i = 1, \dots, n$ are equal) and similarly for $\mathrm{SURE}_n(0)$, $\tilde{\mu}_n(\infty)$, and $\mathrm{SURE}_n(\infty)$. Again in advocating $\tilde{\mu}_n(h)$, we implicitly assume that our regression model is imperfect if its rank is less than $n$. The true model may be $\mu_i = \sum_{j=1}^{p_n} x_{ij}\beta_j + \delta_i$ with $\delta_i$s being nuisance parameters. This is the approximate linear model of Sacks and Ylvisaker (1978) although we do not specify a bound for $\delta_i$s.

We shall assume that

$$\text{(A.1}''\text{)} \qquad \epsilon_i s \text{ are i.i.d. } N(0, \sigma^2).$$

Under this assumption, the transformed data $\bar{y}_i$, $i = 1, \dots, n$ (defined in Example 3), are again independent normal random variables. It is not difficult to see that for $\bar{y}_i$s, our problem takes a simpler form:

$$\text{(4.9)} \quad M_n(h) \text{ is a diagonal matrix with } \lambda_i(\lambda_i + h)^{-1} \text{ as the } i\text{th diagonal element.}$$

LEMMA 4.2. *For $M_n(h)$ defined in* (4.9), $h \in H_n = [0, \infty]$, (A.1) *and* (A.2) *imply* (4.2).

THEOREM 4.6. *Under* (A.1″), *for the ridge-regression problem with ridge estimate* (1.5), *$SURE_n(\hat{h})$ with $\hat{h}$ chosen by GCV is uniformly consistent. In addition, if given $\{\mu_n\}$ there exists a sequence of positive numbers $\{h_n\}$ such that $\hat{\mu}_n(h_n)$ is consistent, then $\tilde{\mu}_n(\hat{h})$ is consistent.*

4.3b. *Smoothing splines.* Consider Example 4. It is well known that $\hat{f}_h$ is a natural polynomial spline of degree $2k - 1$ with knots at $x_i$s. Specifically, let $S_n^k = \{f\colon\ f \in C^{2k-2}[0,1],\ f$ is a polynomial of degree $2k - 1$ on $(x_i, x_{i+1})$, $i = 1, \dots, n - 1$, and $f^{(k)} \equiv 0$ on $[0, x_1]$ and $[x_n, 1]\}$. Consider the basis for $S_n^k$ introduced by Demmler and Reinsch (1975) (see also, Speckman, 1981a, b, 1982) consisting of eigenfunctions $\{\phi_{jn}\}_{j=1}^n$ along with eigenvalues $\{\rho_{jn}\}_{j=1}^n$ satisfying

$$\text{(4.10)} \qquad \frac{1}{n}\sum_{i=1}^{n} \phi_{jn}(x_i)\phi_{j'n}(x_i) = \delta_{jj'}, \qquad \int_0^1 \phi_{jn}^{(k)}(x)\phi_{j'n}^{(k)}(x)\,dx = \rho_{jn}\delta_{jj'},$$

for $1 \le j,\ j' \le n$, with

$$0 = \rho_{1n} = \cdots = \rho_{kn} < \rho_{k+1,n} \le \cdots \le \rho_{nn}.$$

Here $\delta_{jj'}$ is the Kronecker delta. Using this basis, (1.6) is equivalent to

$$\text{(4.11)} \qquad \min_{\mathbf{c} \in R^n} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{n} c_j n^{-1/2}\phi_j(x_i) \right)^2 + h \sum_{j=k+1}^{n} c_j^2 \rho_{jn}.$$

Here $\mathbf{c} = (c_1, c_2, \dots, c_n)'$. Let $U_n$ denote the $n \times n$ matrix with the $ij$th element $n^{-1/2}\phi_j(x_i)$. From (4.10) it follows that $U_n'U_n = I_n$. Put $\bar{\mathbf{y}}_n = U_n'\mathbf{y}_n$. Then (4.11) reduces to

$$\min_{\mathbf{c} \in R^n} \|\bar{\mathbf{y}}_n - \mathbf{c}\|^2 + h \sum_{j=k+1}^{n} c_j^2 \rho_{jn}.$$

The solution $c^*$ of this minimization problem can be obtained easily by standard calculus. It turns out that

$$c^* = \left(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k, (1 + h\rho_{k+1,n})^{-1}\bar{y}_{k+1}, \dots, (1 + h\rho_{n,n})^{-1}\bar{y}_n\right)'.$$

Put $\lambda_1 = \lambda_2 = \cdots = \lambda_k = \infty$ and $\lambda_i = \rho_{i,n}^{-1}$ for $i = k + 1, \dots, n$. We see that $c^* = M_n(h)\bar{\mathbf{y}}_n$, with $M_n(h)$ satisfying (4.9). Hence in terms of $\bar{\mathbf{y}}_n$, our problem is exactly the same as that of the ridge regression. Therefore applying Lemma 4.2, we obtain the following

THEOREM 4.7. *Under* (A.1″), *for the smoothing splines, the* $SURE_n(\hat{h})$ *with* $\hat{h}$ *chosen by GCV is uniformly consistent in estimating the true loss. In addition, if given the true* $\mu_n$, *there exists a sequence of nonnegative numbers* $\{h_n\}$ *such that the corresponding smoothing-spline solution of* (1.6) *is consistent, then* $\tilde{\mu}_n(\hat{h})$ *is consistent.*

Speckman (1982) derived an interesting variant of smoothing splines. It is conceivable that similar results may hold for his procedure.

## 5. Unknown variance of sampling error.

Two procedures will be discussed in Sections 5.1 and 5.2 to cover the case that $\sigma^2$ is unknown.

5.1. *Estimating* $\sigma^2$. As mentioned before, GCV does not require $\sigma^2$, so we may still use it to select $\hat{h}$. After $\hat{h}$ is chosen, we may estimate $\mu_n$ by the Stein estimate $\tilde{\mu}_n(\hat{h})$ with the unknown $\sigma^2$ substituted by a good estimate $\hat{\sigma}_n^2$. We denote such an estimate by $\tilde{\mu}_n(\hat{h}, \hat{\sigma}_n)$. To assess the performance of $\tilde{\mu}_n(\hat{h}, \hat{\sigma}_n)$ we use $\mathrm{SURE}_n(\hat{h}, \hat{\sigma}_n)$ defined to be the $\mathrm{SURE}_n(\hat{h})$ with $\hat{\sigma}_n$ substituted for $\sigma$.

THEOREM 5.1. *Assume that* $\hat{\sigma}_n^2$ *is a consistent estimate of* $\sigma^2$. *Then* $\tilde{\mu}_n(\hat{h}, \hat{\sigma}_n)$ *and* $SURE_n(\hat{h}, \hat{\sigma}_n)$ *are consistent whenever given* $\sigma^2$, $\tilde{\mu}_n(\hat{h})$ *and* $SURE_n(\hat{h})$ *are consistent, respectively. Moreover, if the distribution of* $\hat{\sigma}_n^2$ *does not depend on* $\mu_n$, *then* $SURE_n(\hat{h}, \hat{\sigma}_n)$ *is uniformly consistent, whenever given* $\sigma^2$, $SURE_n(\hat{h})$ *is uniformly consistent.*

The natural case in which the distribution of $\hat{\sigma}_n^2$ does not depend on $\mu_n$ is when there are replicated observations. Another possibility comes up in ridge regression or model selection, where one may assume a true model and use the residual sum of squares for the least-squares estimates to construct $\hat{\sigma}_n^2$ (in this case, a modification of GCV is needed; see Section 6). Sometimes $\hat{\sigma}_n$ can depend on $\mu_n$. For instance, in Example 1, we may take $\hat{\sigma}_n^2 = n^{-1}\Sigma_{i=1}^{n/2}(y_{2i-1} - y_{2i})^2$ for $n$ even. Supposing that as $n$ increases, the $x_i$ values get dense in [0, 1], it is easy to see that $\hat{\sigma}_n^2 \to \sigma^2$. This method of estimating $\sigma^2$ extends naturally to higher-dimensional cases. Rice (1984) has considered such variance estimates in a study of utilizing Mallows' $C_L$ procedure to select the bandwidth of a kernel nonparametric regression in $R^1$. He even suggested the use of higher-order differences in place of the first-order difference $y_{2i-1} - y_{2i}$ to reduce the bias.

5.2. *Returning to the original estimates.* After $\hat{h}$ is selected by GCV, the common practice has been to return to the original linear estimate, $\hat{\mu}_n(\hat{h})$. The results of Section 4 will be used to establish the consistency of $\hat{\mu}_n(\hat{h})$.

First we have seen in many cases that $\tilde{\mu}_n(\hat{h})$ is consistent. Since

$$n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2 = n^{-1}\|\varepsilon_n - \sigma^2 \mathrm{tr}\, A_n(\hat{h}) \cdot \|A_n(\hat{h})\mathbf{y}_n\|^{-2} \cdot A_n(\hat{h})\mathbf{y}_n\|^2$$

and $n^{-1}\|\varepsilon_n\|^2 \to \sigma^2$, it follows that

$$n\|A_n(\hat{h})\mathbf{y}_n\|^2/(\operatorname{tr} A_n(\hat{h}))^2 \to \sigma^2. \tag{5.1}$$

On the other hand,

$$n^{-1}\|\tilde{\mu}_n(\hat{h}) - \hat{\mu}_n(\hat{h})\|^2 = \left(1 - \sigma^2 \operatorname{tr} A_n(\hat{h}) \cdot \|A_n(\hat{h})\mathbf{y}_n\|^{-2}\right)^2 \cdot n^{-1}\|A_n(\hat{h})\mathbf{y}_n\|^2.$$

Hence $n^{-1}\|\hat{\mu}_n(\hat{h}) - \mu_n\|^2$ will tend to 0 if

$$n^{-1}\operatorname{tr} A_n(\hat{h}) \to 1 \quad \text{in probability.} \tag{5.2}$$

In fact, it is easy to obtain the following stronger result.

LEMMA 5.1. *Assume that $\tilde{\mu}_n(\hat{h})$ is consistent. Then $\hat{\mu}_n(\hat{h})$ is consistent if and only if* (5.2) *holds.*

It remains to verify (5.2). A technical step to be established case by case is to derive

LEMMA 5.2. *Under the assumptions of Theorem* 4.4, 4.5, *or* 4.6

$$\lim_{n\to\infty} P\left\{ \frac{\|A_n(\hat{h})\mathbf{y}_n\|^2}{\|A_n(\hat{h})\mu_n\|^2 + \sigma^2 \operatorname{tr} A_n^2(\hat{h})} \le 1 - \delta \right\} = 0 \quad \textit{for any } \delta > 0. \tag{5.3}$$

*Note that the denominator in* (5.3) *is just the expected value of the numerator when $\hat{h}$ is considered as deterministic. From* (5.1) *and* (5.3), *it follows that* $(\operatorname{tr} A_n(\hat{h}))^2 \ge n \operatorname{tr} A_n^2(\hat{h}) \cdot (1 - o_p(1))$, *and then*

$$(\operatorname{tr} A_n(\hat{h}))^2/n \operatorname{tr} A_n^2(\hat{h}) \to 1 \quad \textit{in probability,} \tag{5.4}$$

*because of the inequality* $(\operatorname{tr} A_n(\hat{h}))^2 \le n \operatorname{tr} A_n^2(\hat{h})$. *From* (5.4) *we may obtain* (5.2) *in many cases.*

5.2*a*. *Model selection.* This is the simplest case. Because $A_n(\hat{h})$ is a projection, $A_n^2(\hat{h}) = A_n(\hat{h})$ and (5.4) is identical to (5.2).

THEOREM 5.2. *For the model-selection problem, under the same assumptions as those in Theorem* 4.5, *$\hat{\mu}_n(\hat{h})$ is consistent whenever given $\mu_n$, there exists a sequence of models $\{h_n \in H_n\}$ such that the least-squares estimate $\hat{\mu}_n(h_n)$ is consistent.*

REMARK 4. Hocking's criterion $S_p$ (Hocking, 1976; Thompson, 1978) selects $h$ by minimizing $(n-1)(n-h)^{-1}(n-h-1)^{-1}\|\mathbf{y}_n - \tilde{\mu}_n(h)\|^2$. In view of (1.4), there is little difference between GCV and $S_p$. Recently Breiman and Freedman (1983) established an asymptotic optimality for $S_p$ in the setting of Example 2 with $p_n = \infty$ and $H_n = \{1, 2, \ldots, n/2\}$, under the assumptions that *all explanatory variables and random errors are jointly normal and that there are infinitely many nonzero $\beta_j$s.* Shibata (1981) considered a different criterion: select $h$ by minimizing $n^{-1}(n + 2h)\|\mathbf{y}_n - \hat{\mu}_n(h)\|^2$. Because for $h$, small compared with $n$,

$(1 - h/n)^{-2} \approx 1 + 2hn^{-1}$, this criterion was claimed to be asymptotically equivalent to the GCV. However, the two criteria could be quite different without the crucial assumption $h = o(n)$! Shibata obtained an asymptotic optimality for his criterion but the underlying assumption about $H_n$ (i.e., $\max_{h \in H_n} d(h) = o(n)$) makes it not completely data-driven.

5.2b. *Nearest neighbor regression.* Here $n^{-1}\operatorname{tr} A_n^2(\hat{h}) = (1 - w_{n,\hat{h}}(1))^2 + \sum_{i=2}^{\hat{h}} w_{n,\hat{h}}^2(i)$ and $\operatorname{tr} A_n(\hat{h}) = n(1 - w_{n,\hat{h}}(1))$, (5.4) implies $\sum_{i=2}^{\hat{h}} w_{n,\hat{h}}^2(i)/(1 - w_{n,\hat{h}}(1))^2 \to 0$, which in turn implies $(\hat{h} - 1)^{-1} \to 0$ by the Cauchy–Schwartz inequality and (4.4). Hence we have $\hat{h} \to \infty$ as $n \to \infty$. This together with the following additional regularity condition on the weight function implies (5.2):

(5.5) for any sequence $\{h_n\}$ such that $h_n \to \infty$, we have $w_{n,h_n}(1) \to 0$.

THEOREM 5.3. *For the nearest-neighbor nonparametric regression problem, under* (A.1′), (4.4)–(4.7), *and* (5.5), $\hat{\mu}_n(\hat{h})$ *is consistent.*

5.2c. *Ridge regression.* Assume the following condition on the eigenvalues $\lambda_{i,n}$:

(5.6) there exist constants $p$ and $q$, $0 < p < q < 1$, such that
$$\limsup \lambda_{[qn],n}/\lambda_{[pn],n} < 1,$$

where $[x]$ denotes the greatest integer less than or equal to $x$.

From (5.4) and (5.6) we can derive (5.2) as follows. Let $\lambda$ be the random variable taking value $\lambda_{i,n}$ with probability $n^{-1}$ for each $i$. Now (5.4) means that

$$\left(E\hat{h}(\hat{h} + \lambda)^{-1}\right)^2 / E\hat{h}^2(\hat{h} + \lambda)^{-2} \to 1,$$

where $E$ denotes the expectation with respect to $\lambda$ only ($\hat{h}$ is fixed). This implies that $\hat{h}(\hat{h} + \lambda)^{-1}/E\hat{h}(\hat{h} + \lambda)^{-1} \to 1$ in probability, which in turn implies that both $\hat{h}(\hat{h} + \lambda_{[pn],n})$ and $\hat{h}(\hat{h} + \lambda_{[qn],n})$ tend to $E\hat{h}(\hat{h} + \lambda)^{-1}$. Now because of (5.6), we see that $E\hat{h}(\hat{h} + \lambda)^{-1}$ must tend to 1, proving (5.2).

THEOREM 5.4. *For the ridge-regression problem, under* (A.1″) *and* (5.6), $\hat{\mu}_n(h_n)$ *is consistent whenever given* $\mu_n$ *there exists a sequence* $h_n$ *such that* $\hat{\mu}_n(h_n)$ *is consistent.*

Condition (5.6) is more or less necessary for $\hat{\mu}_n(\hat{h})$ to be consistent. The following example illustrates this point.

EXAMPLE 6. Consider the canonical case with $X_n = \operatorname{diag}(2, 1, \dots, 1)$. Here $\lambda_{1,n} = 4$, $\lambda_{2,n} = \cdots = \lambda_{n,n} = 1$, and GCV chooses $h$ by minimizing

$$n\left[(h + 4)^{-2} y_1^2 + (h + 1)^{-2} \sum_{i=2}^{n} y_i^2\right] \Big/ \left[(h + 4)^{-1} + (n - 1)(h + 1)^{-1}\right]^2$$

over $h \geq 0$. By the straightforward inspection of the derivative, we see that

$$\begin{aligned} \hat{h} &= \infty \quad \text{if } r \geq 1, \\ &= (4r - 1)(1 - r)^{-1} \quad \text{if } \tfrac{1}{4} < r < 1, \\ &= 0 \quad \text{if } 0 \leq r \leq \tfrac{1}{4}, \end{aligned}$$

where $r = (n - 1)^{-1}\Sigma_{i=2}^{n} y_i^2/y_1^2$. For the case that $\mu_n = 0$ and $\varepsilon_i$s are normal, this leads to $\lim P\{\hat{h} = 0\} = P\{4\sigma^2 \leq \varepsilon_1^2\} \approx 0.05$. Since $n^{-1}\operatorname{tr} A_n(0) = 0$, (5.2) does not hold. However, by Theorem 4.6, $\tilde{\mu}_n(\hat{h})$ and $\text{SURE}_n(\hat{h})$ are consistent. Using Lemma 5.1 we see that $\hat{\mu}_n(\hat{h})$ is inconsistent. Note that $\tilde{\mu}_n(0) = \lim_{h \to 0}\tilde{\mu}_n(h) \neq \hat{\mu}_n(0)$.

This example and the condition (5.6) indicate that GCV does not perform well if the problem is not ill-posed. This observation was implicit in Craven and Wahba. However, it is important to note that the inconsistency occurs only because of the insistence on returning to the original linear estimates. The method of Section 5.1 does not have this problem.

5.2d. *Spline smoothing.* The result of ridge regression applies directly to spline smoothing. (5.6) holds easily for the case that $x_i$s are equispaced, in view of the result on eigenvalues from Craven and Wahba. One would even conjecture that it holds if $x_i$s get dense in $[0, 1]$.

THEOREM 5.5. *For the spline-smoothing problem, under* (A.1″), $\hat{\mu}_n(\hat{h})$ *is consistent if* $x_i$s *are equispaced.*

**6. A variant of GCV.** GCV does not depend on $\sigma^2$. But consider the ridge-regression problem. Suppose the number of parameters $p_n$ is less than $n$ and the given regression model is correct. $\mu_n$ is known to be in a proper subspace of $R^n$. Then for estimating $\mu_n$, it is sufficient to consider only the projection of $\mathbf{y}_n$ on this subspace, say $\mathbf{y}_n^*$, if $\epsilon_i$s are normal with $\sigma^2$ given. Any estimate that depends on $\|\mathbf{y}_n - \mathbf{y}_n^*\|^2$ will be inadmissable! This is the case for GCV as can be seen by writing (1.2) as

$$n^{-1}\left(\|\mathbf{y}_n^* - \hat{\mu}_n(h)\|^2 + \|\mathbf{y}_n - \mathbf{y}_n^*\|^2\right)/\left(1 - n^{-1}\operatorname{tr} M_n(h)\right)^2,$$

so that GCV seems inefficient here!

In the following we shall see how our approach of using Stein estimates does not depend on $\sigma^2$ known or unknown. Only the following assumption is crucial:

$$\mu_n \in S_n \subsetneqq R^n, \tag{6.1}$$

where $S_n$ is a subspace of $R^n$ with dimension $s_n < n$.

Under (6.1) it is clear that one should not use $\tilde{\mu}_n(h)$ since it may take value outside $S_n$. Instead, we may replace the raw data $\mathbf{y}_n$ by its projection on $S_n$, say, $\mathbf{y}_n^* = P_n\mathbf{y}_n$, with $P_n$ the projection matrix from $R^n$ to $S_n$. Change the simplified

Stein estimate (3.1) to

$$\tilde{\mu}_n^*(h) = \mathbf{y}_n^* - \frac{\sigma^2 \operatorname{tr}(P_n - M_n(h))}{\|\mathbf{y}_n^* - \hat{\mu}_n(h)\|^2}(\mathbf{y}_n^* - \hat{\mu}_n(h))$$
$$= \mathbf{y}_n^* - \frac{\sigma^2(s_n - \operatorname{tr} M_n(h))}{\|\mathbf{y}_n^* - \hat{\mu}_n(h)\|^2}(\mathbf{y}_n^* - \hat{\mu}_n(h)).$$

Similarly, $\mathrm{SURE}_n(h)$ of (3.2) becomes

$$\mathrm{SURE}_n^*(h) = \sigma^2 - \frac{\sigma^4(s_n - \operatorname{tr} M_n(h))^2}{s_n\|\mathbf{y}_n^* - \hat{\mu}_n(h)\|^2},$$

which estimates the loss $s_n^{-1}\|\mu_n - \tilde{\mu}_n^*(h)\|^2$. Now we see that to choose $h$, one should minimize

$$\mathrm{GCV}_n^*(h) = \frac{s_n^{-1}\|\mathbf{y}_n^* - \hat{\mu}_n(h)\|^2}{\left(1 - s_n^{-1}\operatorname{tr} M_n(h)\right)^2}.$$

REMARK 5. $\mathrm{GCV}_n^*(h)$ can also be derived from the $C_L$-nil-trace estimate argument. But it is unclear whether or not the invariance argument in Golub, Heath, and Wahba is still applicable here.

## 7. Proofs.

Proofs of the results in Sections 3–5 will be sketched in the following subsections, 7.1 to 7.3, respectively. Details may be found in Li (1983). We shall use the following notations:

$$A^{(2)} = A'A \quad \text{for any } n \times n \text{ matrix } A;$$
$$Q_n(h) = \|A_n(h)\mu_n\|^2 + \sigma^2 \operatorname{tr} A_n^{(2)}(h) \quad \left(= E\|A_n(h)\mathbf{y}_n\|^2\right).$$

7.1. *Proofs for Section 3.* Toward proving Theorem 3.1, compute

$$\mathrm{SURE}_n - n^{-1}\|\hat{\mu}_n - \mu_n\|^2 = \sigma^2 - n^{-1}\|\varepsilon_n\|^2 + 2\sigma^2 n^{-1}\operatorname{tr} A_n \cdot\|A_n\mathbf{y}_n\|^{-2}\left(\langle\varepsilon_n, A_n\mu_n\rangle + \langle\varepsilon_n, A_n\varepsilon_n\rangle - \sigma^2\operatorname{tr} A_n\right).$$

It is enough to show that for any $\delta_1, \delta_2 > 0$, there exists an integer $N$ (independent of $\mu_n$) such that for $n \geq N$,

$$P\left\{n^{-1}|\operatorname{tr} A_n| \cdot \|A_n\mathbf{y}_n\|^{-2}|\langle\varepsilon_n, A_n\mu_n\rangle| \geq \delta_1\right\} \leq \delta_2, \tag{7.1.1}$$

$$P\left\{n^{-1}|\operatorname{tr} A_n| \cdot \|A_n\mathbf{y}_n\|^{-2}|\langle\varepsilon_n, A_n\varepsilon_n\rangle - \sigma^2\operatorname{tr} A_n| \geq \delta_1\right\} \leq \delta_2. \tag{7.1.2}$$

The following lemma will be used.

LEMMA 7.1. *Assume* (A.2) *holds. Then for any sequence of nonnegative numbers* $\{a_n\}$ *converging to* 0, *any sequence of real numbers* $\{b_n\}$ *and any sequence of vectors* $\{\mathbf{c}_n \in R^n\}$ *with* $\|\mathbf{c}_n\| = 1$, *we have*

$$\lim_{n\to\infty} P\{|\mathbf{c}_n'\varepsilon_n + b_n| \leq a_n\} = 0.$$

The proof of this lemma will be given later. We proceed with the proof of Theorem 3.1. Since both $\tilde{\mu}_n$ and $\mathrm{SURE}_n$ are invariant under the scale change of $A_n$, we may, without loss of generality, assume that the maximum singular value of $A_n$, $\lambda'(A_n)$, equals 1. (7.1.1) and (7.1.2) will hold if there exists a positive number $a_n$, such that

$$P\{\|A_n\mathbf{y}_n\|^2 \le a_nQ_n\} \le \delta_2/2, \tag{7.1.3}$$

$$P\{n^{-1}|\mathrm{tr}\, A_n| \cdot |\langle \varepsilon_n, A_n\varepsilon_n\rangle - \sigma^2 \mathrm{tr}\, A_n| \ge \delta_1 a_n Q_n\} \le \delta_2/2, \tag{7.1.4}$$

$$P\{n^{-1}|\mathrm{tr}\, A_n| \cdot |\langle \varepsilon_n, A_n\mu_n\rangle| \ge \delta_1 a_n Q_n\} \le \delta_2/2, \tag{7.1.5}$$

where $Q_n$ is the $Q_n(h)$ defined in the beginning of this section with $h$ omitted.

Now by the Chebyshev inequality and the observation that $\mathrm{Var}\langle \varepsilon_n, A_n\varepsilon_n\rangle \le m\, \mathrm{tr}\, A_n^{(2)}$, we can show that (7.1.4) and (7.1.5) hold for

$$a_n \ge cn^{-1/2}, \tag{7.1.6}$$

with some positive number $c$ (for instance, $c = \max\{(2m\sigma^{-2}\delta_1^{-2}\delta_2^{-1})^{1/2}, (2\delta_1^{-2}\delta_2^{-1})^{1/2}\}$).

To obtain (7.1.3), we set $a_n \le 1/2$ and use the Chebyshev inequality again. The left side of (7.1.3) turns out to be no greater than $(8m\sigma^{-2} + 32)Q_n^{-1}$. If $Q_n$ tends to infinity, then (7.1.3) follows immediately. Thus it suffices to take care of the case of bounded $Q_n$.

Recall that we have assumed that $\lambda'(A_n) = 1$. It follows that $A_n^{(2)} \ge \mathbf{c}_n\mathbf{c}_n'$ in the nonnegative definite sense, where $\mathbf{c}_n$ is an eigenvector for $A_n^{(2)}$ with eigenvalue 1 and $\|\mathbf{c}_n\| = 1$. Therefore the left side of (7.1.3) does not exceed

$$P\{|\mathbf{c}_n'\varepsilon_n + \mathbf{c}_n'\mu_n| \le a_n^{1/2}Q_n^{1/2}\},$$

which can be made arbitrarily small if

$$a_n \to 0, \tag{7.1.7}$$

because of the boundedness of $Q_n$ and Lemma 7.1. Finally it is clear that there exists $\{a_n\}$ satisfying (7.1.6) and (7.1.7), completing the proof of Theorem 3.1.

PROOF OF LEMMA 7.1. Without loss of generality, assume that the first coordinate $c_{1n}$ of $\mathbf{c}_n = (c_{1n}, \dots, c_{nn})'$ is positive and is no less than, $|c_{2n}|, \dots, |c_{nn}|$. Rewrite the event $\{|\mathbf{c}_n'\varepsilon_n + b_n| \le a_n\}$ as $\{|\varepsilon_1 + c_{1n}^{-1}(\Sigma_{i=2}^n c_{in}\varepsilon_i + b_n)| \le a_n c_{1n}^{-1}\}$ and consider the conditional probability that this event will hapen given $\varepsilon_2, \dots, \varepsilon_n$. By (A.2) this conditional probability does not exceed $Ka_nc_{1n}^{-1}$, tending to 0 provided that $c_{1n}$ is bounded by away from 0. It remains to consider the case that $c_{1n} \to 0$. By checking the Linderberg–Feller condition, $\mathbf{c}_n'\varepsilon_n$ is seen to be asymptotically normal with mean 0 and variance 1. Lemma 8.1 follows obviously. □

### 7.2. *Proofs for Section 4.*

*7.2a. Proof of Theorem 4.3.* Without loss of generality assume $\lambda'(A_n(h)) = 1$ for each $h \in H_n$. Similar to the proof of Theorem 3.1, it suffices to show that the sum of the last two terms in (4.3) is no less than

$$\begin{aligned}&\sum_{h\in H_n} P\{\|A_n(h)\mathbf{y}_n\|^2 \le 2^{-1}Q_n(h)\}\\ &\quad + \sum_{h\in H_n} P\{2\sigma^2 n^{-1}|\operatorname{tr} A_n(h)|\cdot|\langle\varepsilon_n, A_n(h)\varepsilon_n\rangle - \sigma^2\operatorname{tr} A_n(h)| \ge 4^{-1}\delta Q_n(h)\}\\ &\quad + \sum_{h\in H_n} P\{2\sigma^2 n^{-1}|\operatorname{tr} A_n(h)|\cdot|\langle\varepsilon_n, A_n(h)\mu_n\rangle| \ge 4^{-1}\delta Q_n(h)\}.\end{aligned}$$

Now using the Chebyshev inequality, the above expression is bounded by

$$\begin{aligned}&\sum_{h\in H_n} 16c\left[\sigma^4\left(\operatorname{tr} A_n^{(2)}(h)\right)^2 + \|A_n(h)\mu_n\|^4\right]Q_n(h)^{-4}\\ &\quad + \sum_{h\in H_n} 16^3 c\sigma^8\delta^{-4}n^{-4}\left(\operatorname{tr} A_n(h)\right)^4\left(\operatorname{tr} A_n^{(2)}(h)\right)^2 Q_n(h)^{-4}\\ &\quad + \sum_{h\in H_n} 16^3 c\sigma^{12}\delta^{-4}n^{-4}\left(\operatorname{tr} A_n(h)\right)^4\|A_n(h)\mu_n\|^4 Q_n(h)^{-4},\end{aligned} \tag{7.2.1}$$

for some constant $c$. Here we have used the following inequalities:

$$E\left(\langle\varepsilon_n, A_n(h)\varepsilon_n\rangle - \sigma^2\operatorname{tr} A_n(h)\right)^4 \le c\operatorname{tr} A_n^{(2)}(h);$$

$$E\langle\varepsilon_n, A_n(h)\mu_n\rangle^4 \le c\|A_n(h)\mu_n\|^4;$$

$$E\left(\|A_n(h)y_n\|^2 - Q_n(h)\right)^4 \le c\left[\sigma^4\left(\operatorname{tr} A_n^{(2)}(h)\right)^2 + \|A_n(h)\mu_n\|^4\right].$$

All of these can be verified using Theorem 2 of Whittle (1960).

Finally, the first term in (7.2.1) does not exceed $16c\sigma^{-4}\sum_{h\in H_n}(\operatorname{tr} A_n^{(2)}(h))^{-2}$; the second term does not exceed $16^3\delta^{-4}cn^{-2}\#H_n$; the third term does not exceed $16^3\sigma^8\delta^{-4}n^{-2}\#H_n$. The proof of Theorem 4.3 is now complete by taking $c_1 = 16^3\delta^{-4}(c + \sigma^8)$ and $c_2 = 16c\sigma^{-4}$. □

7.2b. *Proof of Lemma 4.1.* Observe that

$$\begin{aligned}\|M_n(h)\mathbf{y}_n\|^2 &= \sum_{i=1}^{n}\left(\sum_{j=1}^{h} w_{n,h}(j)y_{i(j)}\right)^2 \le \sum_{i=1}^{n}\sum_{j=1}^{h} w_{n,h}(j)y_{i(j)}^2\\ &= \sum_{j=1}^{h} w_{n,h}(j)\sum_{i=1}^{n} y_{i(j)}^2 = \sum_{j=1}^{h}\left(w_{n,h}(j) - w_{n,h}(j+1)\right)\left(\sum_{k=1}^{j}\sum_{i=1}^{n} y_{i(k)}^2\right),\end{aligned}$$

where $w_{n,h}(h+1)$ is set to 0. For $1 \le l \le n$, $2 \le j \le n$, let $n(l, j)$ denote the cardinal number of the set $\bigcup_{k=2}^{j}\{i\colon i(k) = l\}$. It is clear that

$$\sum_{k=1}^{j}\sum_{i=1}^{n} y_{i(k)}^2 = \sum_{i=1}^{n} y_i^2 + \sum_{k=2}^{j}\sum_{i=1}^{n} y_{i(k)}^2 = \sum_{i=1}^{n} y_i^2 + \sum_{l=1}^{n} n(l, j)y_l^2.$$

Now Lemma 2.2 of Li (1984a) showed that there exists a universal constant $\lambda_5$ (depending only on the dimension $p$) such that $n(l, j) \le \lambda_5(j - 1)$ for any

$n, l, j$. Therefore we see that

$$\|M_n(h)\mathbf{y}_n\|^2 \le \lambda_5 \sum_{j=1}^{h} \left(w_{n,h}(j) - w_{n,h}(j+1)\right) \cdot \left(\sum_{l=1}^{n} y_l^2\right) = \lambda_5 \sum_{l=1}^{n} y_l^2.$$

This means $\lambda'(M_n^{(2)}(h)) \le \lambda_5$. Taking $\Lambda = \lambda_5^{1/2}$, the proof of Lemma 4.1 is complete. □

7.2c. *Proof of Lemma 4.2.* As before, it suffices to show the following counterparts for (7.1.3) to (7.1.5): There exists $a_n \to 0$ such that

$$P\left\{\inf_{h \ge 0} h^2 \sum_{i=1}^{n} (\mu_i + \varepsilon_i)^2 (h + \lambda_i)^{-2} / Q_n(h) \le a_n\right\} \le \delta_2/2; \tag{7.2.2}$$

$$P\left\{\sup_{h \ge 0} \frac{h^2}{nQ_n(h)} \sum_{i=1}^{n} (h + \lambda_i)^{-1} \cdot \left|\sum_{i=1}^{n} \frac{\varepsilon_i^2 - \sigma^2}{h + \lambda_i}\right| \ge a_n \delta_1\right\} \le \delta_2/2; \tag{7.2.3}$$

$$P\left\{\sup_{h \ge 0} \frac{h^2}{nQ_n(h)} \sum_{i=1}^{n} (h + \lambda_i)^{-1} \left|\sum_{i=1}^{n} \frac{\mu_i \varepsilon_i}{h + \lambda_i}\right| \ge a_n \delta_1\right\} \le \delta_2/2. \tag{7.2.4}$$

Here $Q_n(h) = h^2 \sum_{i=1}^{n} (h + \lambda_i)^{-2} (\mu_i^2 + \sigma^2)$.

PROOF OF (7.2.2). Define $G_n(h) = h^{-2}(h + \lambda_n)^2 Q_n(h)$. Clearly, $G_n(h)$ is nondecreasing in $h$. Let $L$ be a large number to be chosen later. Let $l$ and $k$ be the largest integers such that $2^l L < G_n(0)$ and $2^k L \le G_n(\infty)$, respectively. Define $h_i = G_n^{-1}(2^i L)$ for $l < i \le k$, $h_l = 0$, and $h_{k+1} = \infty$. Note that $l$, $k$, and $h_i$ are all depending on $n$.

First, consider the case the $l \ge 1$. The left-hand side of (7.2.2) does not exceed

$$\begin{aligned} &\sum_{j=l}^{k} P\left\{\inf_{h_j \le h \le h_{j+1}} G_n(h)^{-1} \sum_{i=1}^{n} (\mu_i + \varepsilon_i)^2 (h + \lambda_n)^2 (h + \lambda_i)^{-2} \le a_n\right\} \\ &\quad \le \sum_{j=l}^{k} P\left\{G_n(h_j)^{-1} \sum_{i=1}^{n} (\mu_i + \varepsilon_i)^2 (h_j + \lambda_n)^2 (h_j + \lambda_i)^{-2} \le 2a_n\right\}. \end{aligned} \tag{7.2.5}$$

Here we have used the fact $G_n(h_j)/G_n(h_{j+1}) \ge 2^{-1}$. Now by the Chebyshev inequality, the last expression does not exceed

$$\begin{aligned} &(1 - 2a_n)^{-2} \sum_{j=l}^{k} \operatorname{Var}\left\{G_n(h_j)^{-1} \sum_{i=1}^{n} (\mu_i + \varepsilon_i)^2 (h_j + \lambda_n)^2 (h_j + \lambda_i)^{-2}\right\} \\ &\quad \le (l - 2a_n)^{-2} C \sum_{j=l}^{k} G_n(h_j)^{-1} \\ &\quad = (1 - 2a_n)^{-2} C L^{-1} \sum_{j=l}^{k} 2^{-j} \le (1 - 2a_n)^{-2} C L^{-1}. \end{aligned}$$

Note that we have used the inequality that $\operatorname{Var}(\mu_i + \varepsilon_i)^2 \le C(\mu_i^2 + \sigma^2)$ for some

$C > 0$ in deriving the first inequality above. Now setting

$$L = 8C\delta_2^{-1}, \tag{7.2.6}$$

we see that (7.2.2) holds for large $n$.

Turning to the case $l \le 0$, we may evaluate the left-hand side of (7.2.2) by splitting $h \ge 0$ into two parts: $0 \le h \le h_1$ and $h_1 \le h$. The second part can be evaluated by (7.2.5) with $l = 1$, leading to the bound $(1 - 2a_n)^{-2}CL^{-1}$. The first part is no greater than

$$P\left\{ \inf_{0 \le h \le h_1} L^{-1} \sum_{i=1}^{n} (\mu_i + \varepsilon_i)^2 (h + \lambda_n)^2 (h + \lambda_i)^{-2} \le a_n \right\}$$

$$\le P\left\{ (\mu_n + \varepsilon_n)^2 L^{-1} \le a_n \right\} \le KL^{1/2} a_n^{1/2},$$

where the last inequality is due to (A.2). Putting these two parts together and using (7.2.6), we see that $KL^{1/2}a_n^{1/2} + (1 - 2a_n)^{-2}CL^{-1} \le \delta_2/2$ for large $n$, proving (7.2.2). □

Before turning to the proofs of (7.2.3) and (7.2.4), we first state a useful lemma.

LEMMA 7.2. *Assume that $W_i$, $i = 1, \dots, n$ are independent random variables with mean* 0 *and finite fourth moments. Then for any $\delta > 0$,*

$$P\left\{ \sup_{0 \le c_1 \le c_2 \le \cdots \le c_n \le 1} \left| \sum_{i=1}^{n} c_i W_i \right| \ge \delta \right\} \le \delta^{-4} E\left( \sum_{i=1}^{n} W_i \right)^4. \tag{7.2.7}$$

This lemma follows immediately from Kolmogorov's inequality [see, e.g., Chung (1974)], after observing that

$$\sup_{0 \le c_1 \le c_2 \le \cdots \le c_n \le 1} \left| \sum_{i=1}^{n} c_i W_i \right| = \sup_{1 \le j \le n} \left| \sum_{i=j}^{n} W_i \right|.$$

A slightly different version of (7.2.7) was first used by Speckman (1981a, 1982).

To proceed, we need only to prove (7.2.4) because by taking $\mu_i = 1$ and treating $\varepsilon_i$ as $\varepsilon_i^2 - \sigma^2$, we see that (7.2.3) is essentially a special case of (7.2.4).

PROOF OF (7.2.4). Write $\lambda_0 = \infty$ and $\lambda_{n+1} = 0$. Let $I_1(j) = \{1, 2, \dots, j\}$ and $I_0(j) = \{j + 1, \dots, n\}$. Using the inequality $\sum_{i=1}^{n}(h + \lambda_i)^{-1} \le \sqrt{n}(\sum_{i=1}^{n}(h + \lambda_i)^{-2})^{1/2}$, we see that (7.2.4) will follow from

$$\sum_{j=l}^{n+l-1} P\left\{ \sup_{\lambda_{j+1} \le h \le \lambda_j} \left| \sum_{i \in I_l(j)} \frac{\mu_i \varepsilon_i}{h + \lambda_i} \right| (nQ_n(h))^{-1/2} h \ge \tfrac{1}{2} a_n \delta_1 \sigma \right\} \le \frac{\delta_2}{4} \tag{7.2.8}$$

for $l = 0$ and 1.

We now prove (7.2.8). Consider $l = 1$ first. Since $h^{-2}Q_n(h)$ is nonincreasing in $h$, the left-hand side of (7.2.8) does not exceed

$$\sum_{j=1}^{n} P\left\{n^{-1/2}\left|\sum_{i=1}^{j}(\lambda_j+\lambda_i)^{-1}\mu_i\varepsilon_i\right|Q_n(\lambda_j)^{-1/2}\lambda_j \geq \tfrac{1}{4}a_n\delta_1\sigma\right\}$$

$$\text{(7.2.9)} \qquad + \sum_{j=1}^{n} P\left\{\sup_{\lambda_{j+1}\leq h\leq\lambda_j} n^{-1/2}\left|\sum_{i=1}^{j}\left[(h+\lambda_i)^{-1}-(\lambda_j+\lambda_i)^{-1}\right]\mu_i\varepsilon_i\right| \cdot Q_n(\lambda_j)^{-1/2}\lambda_j \geq \tfrac{1}{4}a_n\delta_1\sigma\right\}.$$

By the Chebyshev inequality, the first term does not exceed

$$\text{(7.2.10)} \qquad \left(\tfrac{1}{4}a_n\delta_1\sigma\right)^{-4}mn^{-2}\sum_{j=1}^{n}\left(\sum_{i=1}^{j}(\lambda_j+\lambda_i)^{-2}\mu_i^2\right)^2 Q_n(\lambda_j)^{-2}\lambda_j^4,$$

which clearly is no greater than $(1/4a_n\delta_1\sigma)^{-4}mn^{-1}$. Thus the first term in (7.2.9) will tend to 0 if

$$\text{(7.2.11)} \qquad a_n^4 n \to \infty.$$

The second term in (7.2.9) can be evaluated by using Lemma 7.2. Write $(h+\lambda_i)^{-1} - (\lambda_j+\lambda_i)^{-1} = (\lambda_j - h)(h+\lambda_i)^{-1}(\lambda_j+\lambda_i)^{-1}$ and observe that $(\lambda_j - h)(h+\lambda_i)^{-1}$ is nondecreasing in $i$ and is no greater than 1 for $\lambda_{j+1} \leq h \leq \lambda_j$. We may put $c_i = (\lambda_j - h)(h+\lambda_i)^{-1}$, $W_i = (\lambda_j+\lambda_i)^{-1}\mu_i\varepsilon_i$, and $\delta = 1/4a_n\delta_1\sigma n^{1/2}Q_n(\lambda_j)^{1/2}\lambda_j^{-1}$ in (7.2.7), yielding (7.2.10) as the desired upper bound for the second term in (7.2.9). Therefore we have seen that with $l = 1$, (7.2.8) holds for large $n$ if $a_n$ is chosen to satisfy (7.2.11).

The case $l = 0$ can be treated in a similar manner. Because $Q_n(h)$ is nondecreasing in $h$, the counterpart of (7.2.9) becomes

$$\sum_{j=0}^{n-1} P\left\{n^{-1/2}\left|\sum_{i=j+1}^{n}\lambda_{j+1}(\lambda_{j+1}+\lambda_i)^{-1}\mu_i\varepsilon_i\right|Q_n(\lambda_{j+1})^{-1/2} \geq \tfrac{1}{4}a_n\delta_1\sigma\right\}$$

$$\text{(7.2.12)} \qquad + \sum_{j=0}^{n-1} P\left\{\sup_{\lambda_{j+1}\leq h\leq\lambda_j}\left(nQ_n(\lambda_{j+1})\right)^{-1/2}\left|\sum_{i=j+1}^{n}\mu_i\varepsilon_i\left(\frac{h}{h+\lambda_i}-\frac{\lambda_{j+1}}{\lambda_{j+1}+\lambda_i}\right)\right| \geq \tfrac{1}{4}a_n\delta_1\sigma\right\}.$$

Both terms are again no greater than $(1/4a_n\delta_1\sigma)^{-4}mn^{-1}$ by the Chebyshev inequality and Lemma 7.2. Here we should put $c_i = (h - \lambda_{j+1})(h+\lambda_i)^{-1}$, $W_i = \lambda_i(\lambda_{j+1}+\lambda_i)^{-1}\mu_i\varepsilon_i$, and $\delta = 1/4a_n\delta_1\sigma n^{1/2}Q_n(\lambda_{j+1})^{1/2}$ when applying (7.2.7). In conclusion, under (7.2.11), (7.2.8) holds for large $n$ where $l = 0$. This completes the proof of (7.2.4). □

7.3. *Proofs for Section 5.*

7.3a. *Proof of Theorem 5.1.* The consistency of $\mathrm{SURE}_n(\hat{h}, \hat{\sigma}_n)$ will follow from

$$|\mathrm{SURE}_n(\hat{h}, \hat{\sigma}_n) - \mathrm{SURE}_n(\hat{h})| \to 0, \tag{7.3.1}$$

$$n^{-1}\|\tilde{\mu}_n(\hat{h}, \hat{\sigma}_n) - \tilde{\mu}_n(\hat{h})\|^2 \to 0. \tag{7.3.2}$$

To obtain (7.3.1), observe that the absolute value term in (7.3.1) is no greater than

$$|\hat{\sigma}_n^2 - \sigma^2| + |\sigma^{-4}\hat{\sigma}_n^4 - 1|\left(\sigma^2 - \mathrm{SURE}_n(\hat{h})\right). \tag{7.3.3}$$

The consistency of $\mathrm{SURE}_n(\hat{h})$ implies that

$$\sigma^2 - \mathrm{SURE}_n(\hat{h}) = \sigma^2 - n^{-1}\|\mu_n - \tilde{\mu}_n(\hat{h})\|^2 + o_p(1) \le \sigma^2 + o_p(1).$$

We see that (7.3.3) vanishes asymptotically, implying (7.3.1). To obtain (7.3.2) first observe that $\tilde{\mu}_n(\hat{h}, \hat{\sigma}_n) - \tilde{\mu}_n(\hat{h}) = (1 - \sigma^{-2}\hat{\sigma}_n^2)(\mathbf{y}_n - \tilde{\mu}_n(h))$. Therefore the right-hand side of (7.3.2) does not exceed

$$2\left(1 - \sigma^{-2}\hat{\sigma}_n^2\right)^2\left(n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2 + n^{-1}\|\varepsilon_n\|^2\right). \tag{7.3.4}$$

Since $n^{-1}\|\varepsilon_n\|^2 \to \sigma^2$ and $n^{-1}\|\tilde{\mu}_n(\hat{h}) - \mu_n\|^2 = \mathrm{SURE}_n(\hat{h}) + o_p(1) \le \sigma^2 + o_p(1)$, we see that (7.3.4) converges to 0, proving (7.3.2).

The consistency of $\tilde{\mu}_n(\hat{h}, \hat{\sigma}_n)$ also follows from (7.3.2). The proof of Theorem 5.1 is now complete.

7.3b. *Proof of Lemma 5.2.* (I) *Nearest neighbor regression.* It suffices to show that for any $\delta > 0$,

$$P\left\{\inf_{h \in H_n} \|A_n(h)\mathbf{y}_n\|^2 Q_n(h)^{-1} \le 1 - \delta\right\} \to 0. \tag{7.3.5}$$

But this can be proved easily by the Chebyshev inequality as in Section 7.2a. We omit the details.

(II) *Model selection.* (7.3.5) holds if replacing "$h \in H_n$" by "$h \notin H_n'$." Hence it suffices to show that for any $h_n \in H_n'$ such that $n - d(h_n)$ is bounded, say by $M$,

$$P\{\hat{h} = h_n\} \to 0. \tag{7.3.6}$$

Now by Theorem 4.5, $\mathrm{SURE}_n(\hat{h})$ and $\tilde{\mu}_n(\hat{h})$ are consistent, implying that $\mathrm{SURE}_n(\hat{h}) \to 0$. This means that for any $\delta_1 > 0$,

$$\begin{aligned} 1 &= \lim_{n\to\infty} P\left\{\|A_n(\hat{h})\mathbf{y}_n\|^2 \le \left(\sigma^2 + \delta_1\right)n^{-1}\left(n - d(\hat{h})\right)^2\right\} \\ &\le \lim_{n\to\infty} P\{\hat{h} \ne h_n\} + \lim_{n\to\infty} P\left\{\|A_n(h_n)y_n\|^2 \le \left(\sigma^2 + \delta_1\right)n^{-1}M^2\right\}. \end{aligned}$$

The second term tends to 0 by Lemma 7.1 because $\|A_n(h_n)\mathbf{y}_n\|^2 \ge (\mathbf{c}_n'\mathbf{y}_n)^2$ for some $\mathbf{c}_n \in R^n$ with $\|\mathbf{c}_n\| = 1$. (7.3.6) is obtained, completing the proof.

(III) *Ridge regression.* It suffices to consider the canonical case (4.9). Define

$$D_n(h) = \sum_{i=1}^{n} (h + \lambda_i)^{-2} y_i^2 \cdot h^2 Q_n(h)^{-1}.$$

We shall show that given $\varepsilon$, $\delta > 0$, the following holds for large $n$:

$$P\{D_n(\hat{h}) \le 1 - \delta\} \le \varepsilon. \tag{7.3.7}$$

Let positive numbers, $\delta_1$, $\delta_2$, and $c$ be small and $F$ be large. They will be chosen appropriately later on. Define

$$h_n^* = \text{the largest } h \text{ such that } n\delta_1(\sigma^2 + \delta)^{-1} \ge \left(\sum_{i=1}^{n} (h + \lambda_n)/(h + \lambda_i)\right)^2,$$

$$h_n^0 = \text{the largest } h \text{ such that } (n - 1) \ge c\left(\sum_{i=1}^{n-1} (h + \lambda_n)/(h + \lambda_i)\right)^2,$$

$$h_n' = \text{the smallest } h \text{ such that } G_n(h) \ge F,$$

where $G_n(h)$ is defined in the beginning of the proof of (7.2.2). (7.3.7) will hold if we can show that $\varepsilon$ is no less than

$$\begin{aligned} &P\{y_n^2 \le \delta_1\} + P\{y_n^2 > \delta_1, \hat{h} \in [0, h_n^*]\} + P\{y_n^2 > \delta_1, \hat{h} \in [h_n^*, h_n^0]\} \\ &+ P\{y_n^2 > \delta_1, \hat{h} \in [h_n^0, h_n']\} + P\Big\{\inf_{h \in [h_n', \infty]} D_n(h) \le 1 - \delta\Big\}. \end{aligned} \tag{7.3.8}$$

Each term in (7.3.8) will be made no greater than $\varepsilon/5$ in the following.

It is easy to treat the first term. For the second term, we observe that it does not exceed

$$P\left\{ny_n^2(\sigma^2 + \delta)^{-1} \ge \left(\sum_{i=1}^{n} (\hat{h} + \lambda_n)/(\hat{h} + \lambda_i)\right)^2\right\} \tag{7.3.9}$$

because of the definition of $h_n^*$. (7.3.9) is obviously no greater than

$$P\left\{n\sum_{i=1}^{n} (\hat{h} + \lambda_i)^{-2} y_i^2 \Big/ \left(\sum_{i=1}^{n} (\hat{h} + \lambda_i)^{-1}\right)^2 \ge \sigma^2 + \delta\right\} = P\{\mathrm{GCV}_n(\hat{h}) \ge \sigma^2 + \delta\}.$$

On the other hand, by Theorem 4.6, $\mathrm{SURE}_n(\hat{h}) \to 0$, implying that

$$\mathrm{GCV}_n(\hat{h}) \to \sigma^2. \tag{7.3.10}$$

Therefore we see that the second term in (7.3.8) converges to 0.

To bound the third term in (7.3.8), we first observe that it does not exceed

$$o(1) + P\Big\{y_n^2 > \delta_1, \inf_{h \in [h_n^*, h_n^0]} |\mathrm{GCV}_n(h) - \sigma^2| \le \delta_2\Big\} \tag{7.3.11}$$

because of (7.3.10). On the other hand, if $r_i$ denotes $(h + \lambda_n)(h + \lambda_i)^{-1}$, then the definition of $h_n^*$ implies that for any $h \ge h_n^*$, $\sum_{i=1}^n r_i$ converges to $\infty$, which in turn shows that $\Sigma_{i=1}^{n-1} r_i / \Sigma_{i=1}^{n} r_i$. From this we see that (7.3.11) does not exceed

$$o(1) + P\left\{ y_n^2 > \delta_1, \inf_{h \in [h_n^*, h_n^0]} \left| (n-1)\left( \sum_{i=1}^{n-1} r_i \right)^{-2} \cdot \left( \sum_{i=1}^{n-1} r_i^2 y_i^2 + y_n^2 \right) - \sigma^2 \right| \leq \delta_2 \right\}, \tag{7.3.12}$$

which, by the definition of $h_n^0$, is no greater than

$$o(1) + P\left\{ \inf_{h \in [h_n^*, h_n^0]} (n-1)\left( \sum_{i=1}^{n-1} (h + \lambda_i)^{-1} \right)^{-2} \times \sum_{i=1}^{n-1} (h+\lambda_i)^{-2} y_i^2 \leq \sigma^2 - c\delta_1 + \delta_2 \right\}$$

$$\leq o(1) + P\left\{ \inf_{h \in [0, \infty]} \mathrm{GCV}_{n-1}(h) \leq \sigma^2 - c\delta_1 + \delta_2 \right\}.$$

By Remark 3 of Section 4, the last term converges to 0, providing that

$$c > \delta_2 \delta_1^{-1}. \tag{7.3.13}$$

Turning to the fourth term in (7.3.8), we shall assume $h_n' > h_n^0$; otherwise this term vanishes. As before, we may bound it by expression (7.3.11) with $[h_n^*, h_n^0]$ replaced by $[h_n^0, h_n']$, which in turn is no greater than

$$o(1) + P\left\{ cn(n-1)^{-1} \sum_{i=1}^{n} (h_n' + \lambda_n)^2 (h_n' + \lambda_i)^{-2} y_i^2 > \sigma^2 - \delta_2 \right\}$$

by the definition of $h_n^0$. Now by the Chebyshev inequality, the above expression does not exceed

$$o(1) + (\sigma^2 - \delta_2)^{-1} cn(n-1)^{-1} \sum_{i=1}^{n} (h_n' + \lambda_n)^2 (h_n' + \lambda_i)^{-2} (\mu_i^2 + \sigma^2)$$

$$= o(1) + (\sigma^2 - \delta_2)^{-1} cn(n-1)^{-1} F,$$

where the quality is due to the definition of $h_n'$ (note that $h_n' \neq 0$), Hence we see that the fourth term in (7.3.8) does not exceed $\varepsilon/5$ if

$$c < F^{-1}(\sigma^2 - \delta_2)\varepsilon/5. \tag{7.3.14}$$

It remains to bound the fifth term. The argument will be similar to the proof of (7.2.2). Let $L$ be a large number to be chosen later. Let $l$ and $k$ be large integers such that $(1+\delta)^l L \leq G_n(h_n')$ and $(1+\delta)^k L \leq G_n(\infty)$. Define $h_i = G_n^{-1}((1+\delta)^i L)$ for $l < i \leq k$, $h_l = h_n'$, and $h_{k+1} = \infty$. Note that $l$, $k$, and $h_i$ are depending on $n$. Now set

$$(1+\delta)L < F. \tag{7.3.15}$$

By the definition of $h_n'$, we have $l \geq 1$. Similar to (7.2.5), the fifth term of (7.3.8) does not exceed

$$\sum_{j=l}^{k} P\{D_n(h_j) \le 1 - \delta^2\}.$$

As before, using the Chebyshev inequality, the last expression does not exceed

$$\delta^{-4}C\sum_{j=l}^{k} G_n(h_j)^{-1} = \delta^{-4}CL^{-1}\sum_{j=l}^{k}(1+\delta)^{-j} \le C\delta^{-5}L^{-1}.$$

Therefore we see that the fifth term of (7.3.8) does not exceed $\varepsilon/5$, if

$$L \ge 5C\delta^{-5}\varepsilon^{-1}. \tag{7.3.16}$$

Finally, we choose $F$ large enough so that (7.3.15) and (7.3.16) hold for some $L$. Then choose $\delta_2$ small enough so that (7.3.13) and (7.3.14) hold for some $c$. Each term in (7.3.8) is now no greater than $\varepsilon/5$, completing the proof. □

**Acknowledgment.** I am grateful to Professor D. Ylvisaker for useful discussions and editorial assistance.

## REFERENCES

BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation. *J. Amer. Statist. Assoc.* **78** 131–136.

CASELLA, G. (1980). Minimax ridge regression estimation. *Ann. Statist.* **8** 1036–1056.

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

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.

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

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

GEISSER, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320–328.

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

GOLUB, G. and REINSCH, C. (1970). Singular value decomposition and least squares solutions. *Numer. Math.* **14** 403–420.

HOCKING, R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32** 1–49.

HUBER, P. J. (1975). Robustness and designs. *A Survey of Statistical Design and Linear Models*. 287–303, North Holland, Amsterdam.

LI, K. C. (1983). From Stein's unbiased risk estimates to the method of generalized cross-validation. Technical Report, Department of Statistics, Purdue University.

LI, K. C. (1984a). Consistency for cross-validated nearest neighbor estimates in nonparametric regression. *Ann. Statist.* **12** 230–240.

LI, K. C. (1984b). Regression models with infinitely many parameters: consistency of bounded linear functionals. *Ann. Statist.* **12** 601–611.

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 *Cp*. *Technometrics* **15** 661–675.

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

RICE, J. (1984). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.*, **12**, 1215–1230.

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

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

SPECKMAN, P. (1981). The asymptotic integrated error for smoothing noisy data by splines. *Numer. Math.*. To appear.

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

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

SILVERMAN, B. W. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Amer. Statist. Assoc.* **79** 584–589.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1155.

STONE, C. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statist. Soc. Ser. B* **36** 111–147.

THOMPSON, M. (1978). Selection of variables in multiple regression. *Internat. Statist. Review* **46** 1–49 and 129–146.

UTRERAS, F. (1979). Cross validation techniques for smoothing in one or two dimensions. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.). Lecture Notes in Mathematics No. 757. Springer, New York.

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

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

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

WAHBA, G. (1982). Constrained regularization for ill-posed linear operator equations, with application in meteorology and medicine. In *Statistical Decision Theory and Related Topics III* **2** (S. Gupta and J. Berger, eds.) 383–418.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, LOS ANGELES
LOS ANGELES, CALIFORNIA, 90024