Asymptotic Optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set*

by

Ker-Chau Li
Department of Mathematics, U.C.L.A.
and
Department of Statistics, Purdue University

Technical Report #84-17

Department of Statistics
Purdue University

May 1984

*This work is sponsored by National Science Foundation under Grant No. MCS-8200631 at Purdue University.

AMS 1980 Subject Classification: Primary 62G99, 62J99
Secondary 62J05, 62J07

Key Words and Phrases: model-selection; nearest neighbor estimates; nil-trace linear estimates; non-parametric regression; Stein estimates; Stein's unbiased risk estimates.

Asymptotic Optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set

KER-CHAU LI

Abstract. Data-driven techniques of selecting a good estimate from a proposed class of linear estimates $\hat{\mu}(h)$, $h \in H_n$, are studied. Here $\hat{\mu}(h)$ takes the form $M(h)y$ with y being the vector of n independent observations whose mean value $\mu = (\mu_1,\ldots,\mu_n)'$ is to be estimated. Many selection procedures have been proposed including Mallows' C_p, C_L, cross-validation and generalized cross-validation. Let $\hat{h}$ denote the h selected by any of these procedures. Under certain reasonable conditions, it is shown that

$$n^{-1}||\mu - \hat{\mu}(\hat{h})||^2/\inf\{n^{-1}||\mu - \hat{\mu}(h)||^2 : h \in H_n\} \rightarrow 1$$

in probability. The applications in nearest neighbor nonparametric regression and model-selection are discussed.

1. INTRODUCTION

Let $y_n = (y_1,y_2,\ldots,y_n)'$ be a vector of n independent observations with unknown means μ_1, μ_2, ..., μ_n. Write

$$y_i = \mu_i + e_i, \quad i = 1, 2, \ldots, n, \tag{1.1}$$

and assume that the random errors e_i are identically distributed with mean 0 and variance σ^2. Suppose that to estimate $\underset{\sim}{\mu}_n = (\mu_1, \mu_2, \ldots, \mu_n)'$, a class of linear estimators $\underset{\sim}{\hat{\mu}}_n(h) = M_n(h)\underset{\sim}{y}_n$, indexed by $h \in H_n$, is proposed. Here $M_n(h)$ is an $n \times n$ matrix and H_n is just an index set. After observing y_i's, our concern is to select an $\hat{h}$ from H_n so that the average squared error $L_n(\hat{h}) = n^{-1}||\underset{\sim}{\mu}_n - \underset{\sim}{\hat{\mu}}_n(\hat{h})||^2$ may be as small as possible ($||\cdot||$ denotes the Euclidean norm). The following are three well-known procedures of selection.

(i) Mallows' C_L (Mallows 1973): select $\hat{h}$, denoted by $\hat{h}_M$, that achieves

$$\min_{h \in H_n} n^{-1}||\underset{\sim}{y}_n - \underset{\sim}{\hat{\mu}}_n(h)||^2 + 2\sigma^2 n^{-1} \operatorname{tr} M_n(h).$$

(ii) Generalized cross-validation (Craven and Wahba 1979): select $\hat{h}$, denoted by $\hat{h}_G$, that achieves

$$\min_{h \in H_n} \frac{n^{-1}||\underset{\sim}{y}_n - \underset{\sim}{\hat{\mu}}(h)||^2}{(1 - n^{-1} \operatorname{tr} M_n(h))^2} \, .$$

(iii) Cross-validation (Allen 1974, Stone 1974, Geisser 1975, Wahba and Wold 1975): select $\hat{h}$, denoted by $\hat{h}_C$, that minimized the sum of squared prediction errors for y_i with y_i itself being excluded from the data set. A rigorous definition of this procedure requires the specification of estimators (or predictors) to be used when sample size is $n - 1$. Suppose given $y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$, we want to predict y_i by $\sum_{j=1}^n \bar{m}_{ij}(h)y_j$ with $\bar{m}_{ii}(h)$ being zero. Then $\hat{h}_c$ achieves

$$\min_{h \in H_n} ||\underset{\sim}{y}_n - \bar{M}_n(h)\underset{\sim}{y}_n||^2,$$

where $\bar{M}_n(h)$ is an $n \times n$ matrix with $\bar{m}_{ij}(h)$ as the ijth entry.

In defining C_L, we assume that σ^2 is known. If σ^2 is unknown, the stability of its estimate may play an influential role in the selection. GCV and CV do not depend on σ^2; this seems to be a great advantage. The goal of this paper is to demonstrate that under reasonable conditions, these procedures are asymptotically optimal (A.O. hereafter) in the sense that

$$\frac{L_n(\hat{h})}{\inf_{h \in H_n} L_n(h)} \to 1, \tag{1.2}$$

in probability. Thus using these procedures, statisticians may do as well as if they knew the true $\underset{\sim}{\mu}_n$. We shall consider only the case that the cardinality of H_n is finite. The continuous case, particularly the ridge regression setting will be treated in a forthcoming paper.

Consider the expected average squared error $R_n(h) = EL_n(h)$, $h \in H_n$. Among the conditions to be imposed in order to derive the desired results, the following one is most crucial

$$\sum_{h \in H_n} (nR_n(h))^{-m} \to \infty, \text{ for some natural number } m. \qquad (A.1)$$

Sometimes (A.1) may be implied by the weaker condition:

$$\inf_{h \in H_n} nR_n(h) \to \infty, \qquad (1.3)$$

which means that the optimal convergent rate, when $\underset{\sim}{\mu}_n$ is known, is slower that n^{-1}. It seems that without (1.3) we cannot find any selection procedure $\hat{h}$ to satisfy (1.2); otherwise the resulted estimate of $\underset{\sim}{\mu}_n$ would have unattainably small error.

As prime examples to demonstrate the application of our general results we shall treat the following two special settings in details:

Example 1. Model selection: Suppose associated with y_i there are p_n explanatory variables x_{i1}, x_{i2}, x_{ip_n}, arranged in the decreasing order of importance. Take $H_n = \{1,2,\ldots,p_n\}$. To estimate $\underset{\sim}{\mu}_n$, one may employ the first h variables to form a linear model $y_i = \sum_{j=1}^{h} x_{ij}\beta_j + e_i$ with unknown parameters β_j, $j = 1, \ldots, h$ and then use the least squares estimator $\underset{\sim}{\hat{\mu}}(h) = X_h(X'_hX_h)^{-1}X'_h\underset{\sim}{y}_n$, where X_h

is the $h \times h$ design matrix (x_{ij}). Assume that the information matrix is non-singular. Now, $M_n(h) = X_h(X_h'X_h)^{-1}X_h$ is a projection matrix of rank h. C_L procedure reduces to the more famous C_p criterion which select $\hat{h}$ that minimizes

$$\min_{h \in H_n} ||\underset{\sim}{y}_n - \underset{\sim}{\hat{\mu}}(h)||^2 + 2\sigma^2 h.$$

Later in Section 2 we shall show that (1.3) implies (A.1) for $m = 2$. $\hat{h}_M$ and $\hat{h}_G$ are both A.O. In view of the relationship between $\hat{h}_C$ and $\hat{h}_G$, there seems no guarantee that $\hat{h}_C$ will be A.O. unless the diagonal elements of $M_n(h)$ are nearly equal to each other, in which case $\hat{h}_C$ is almost the same as $\hat{h}_G$. The assumption (1.3) seems quite reasonable if p_n, the number of explanatory variables, grows as the sample size increases. For instance, in the problem of selecting the suitable degree of a polynomial to fit a response curve, (1.3) will hold when the true regression function is not a polynomial (Shibata 1981).

Shibata (1981) demonstrated the A.O. property of a related selection procedure, Final Prediction Error (FPE) criterion, which selects $\hat{h}$ by minimizing $n^{-1}||\underset{\sim}{y}_n - \underset{\sim}{\hat{\mu}}(h)||^2(n + 2h)$, under the normality assumption of e_1, (1.3), and the rank of the largest model considered $p_n = o(n)$. The last condition makes his selection procedure not completely data-driven because it is hard to judge when p_n will be small enough compared with n. It was also claimed that C_p and FPE are asymptotically equivalent. But apparently without the assumption that

$p_n = o(n)$ this can be false; for example if $p_n = n$, then FPE always selects $\hat{h} = p_n$. Our results (both C_p and GCV) do not require this assumption.

Breiman and Freedman (1983) studied another procedure, S_p, firstly proposed by Hocking (1976), then explored by Thompson (1978), which selects $\hat{h}$ by minimizing $(n - 1)||\underset{\sim}{y} - \underset{\sim}{\hat{\mu}}(h)||^2/(n - h)(n - h - 1)$. Obviously S_p is almost identical to GCV for large n. Under a regression model with infinitely many non-zero parameters, Breiman and Freedman showed that S_p is A.O., under the conditions that all explanatory variables and random errors are jointly normal and that $H_n = \{1,\ldots,n/2\}$. The first condition excludes many interesting applications; for instance, fitting a response curve by polynomials.

Example 2. Nearest neighbor nonparametric regression (Stone 1977): Let p be a natural number and X be the compact closure of an open connected set in R^p. Suppose $y_1, y_2, \ldots, y_n$ are observed at distinct levels $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$, which become dense in X uniformly as $n \rightarrow \infty$. Assume that $\mu_i = f(\underset{\sim}{x}_i)$ for an unknown continuous function f on X. Let $\underset{\sim}{x}_{i(j)}$ denote the jth nearest neighbor of $\underset{\sim}{x}_i$ in the sense that $||\underset{\sim}{x}_i - \underset{\sim}{x}_{i(j)}||$ is the jth smallest number among the n values $||\underset{\sim}{x}_i - \underset{\sim}{x}_{i'}||$, $i' = 1, 2, \ldots, n$. Ties may be broken in any systematic manner. Take $H_n = \{1,2,\ldots,n\}$. For any $h \in H_n$, let $\underset{\sim}{\hat{\mu}}_n(h)$ be the h-nearest neighbor estimate of $\underset{\sim}{\mu}_n$, with the ith coordinate given by

$\sum_{j=1}^{h} w_{n,h}(j)y_{i(j)}$ for some nonnegative weight function $w_{n,h}(\cdot)$ such that

$$\sum_{i=1}^{h} w_{n,h}(i) = 1, \tag{1.4}$$

and

$$w_{n,h}(i) \geq w_{n,h}(i+1) \quad \text{for any} \quad i, \quad 1 \leq i \leq h-1. \tag{1.5}$$

Under these and some other conditions, we shall demonstrate that $\hat{h}_M$, $\hat{h}_G$ and $\hat{h}_C$ are A.O., if

$$\lim_{n\to\infty} \left[\inf_{h \in H_n} nR_n(h) \right] / n^{1/m} \to \infty, \tag{1.6}$$

for some natural number m, which is a sufficient condition for (A.1). In view of the results for the optimal convergent rates in nonparametric regression (e.g., Stone 1982), (1.6) is quite natural although it is slightly stronger than (1.3).

Section 2 establishes the asymptotic efficiency of $\hat{h}_M$. Then these results will be used to treat $\hat{h}_G$ in Section 3. The main idea involved here is the notion of nil-trace linear estimates, firstly introduced by Li (1983) to bring a connection between GCV and C_L. This approach to GCV though valid asymptotically may look shaky when the sample size is not large. A better approach suggested by Li (1983) is by means of

Stein's estimates and Stein's unbiased risk estimates, Stein (1981). In Section 4 we shall show that the G-cross-validated Stein estimate is A.O. Finally, Section 5 is devoted to the study of Cross-validation with special attention to the nearest neighbor nonparametric regression whose consistency was established by Li (1984). All the proofs will be given in Section 6.

To simplify the notation without ambiguity, we shall frequently omit the subscript n.

2. C_p AND C_L.

Let $\underset{\sim}{e}_n = (e_1,e_2,\ldots,e_n)'$ and $A_n(h) = I - M_n(h)$ where I is the $n \times n$ identity matrix. The motivation of C_L comes from the simple identity that

$$n^{-1}||\underset{\sim}{y}_n - \underset{\sim}{\hat{\mu}}(h)||^2 + 2\sigma^2 n^{-1} \text{ tr } M_n(h)$$

$$= n^{-1}||\underset{\sim}{e}_n||^2 + L_n(h) + 2n^{-1} \langle \underset{\sim}{e}_n, A_n(h)\underset{\sim}{\mu}_n\rangle + 2n^{-1}(\sigma^2 \text{ tr } M_n(h) - \langle \underset{\sim}{e}_n, M_n(h)\underset{\sim}{e}_n\rangle).$$

Since $n^{-1}||\underset{\sim}{e}_n||^2$ is independent of h, $\hat{h}_M$ also minimizes

$$L_n(h) + 2n^{-1}\langle \underset{\sim}{e}_n, A_n(h)\underset{\sim}{\mu}_n\rangle + 2n^{-1}(\sigma^2 \text{ tr } M_n(h) - \langle \underset{\sim}{e}_n, M_n(h)\underset{\sim}{e}_n\rangle)$$

over $h \in H_n$. If we can show that $n^{-1}\langle \underset{\sim}{e}_n, A_n(h)\underset{\sim}{\mu}_n\rangle$ and

$n^{-1}(\sigma^2 \text{ tr } M_n(h) - \langle e_n, M_n(h) e_n \rangle)$ are negligible compared with $L_n(h)$ uniformly for any $h \in H_n$, then (1.2) is established for $\hat{h} = \hat{h}_M$. In other words it remains to show that in probability,

$$\sup_{h \in H_n} n^{-1} \langle e_n, A_n(h) \mu_n \rangle / R_n(h) \to 0 \tag{2.1}$$

$$\sup_{h \in H_n} n^{-1} |\sigma^2 \text{ tr } M_n(h) - \langle e_n, M_n(h) e_n \rangle| / R_n(h) \to 0 \tag{2.2}$$

and

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

Some assumptions will be made to establish (2.1) ~ (2.3). Specifically, we have

Theorem 2.1. Assume that (A.1) and the following conditions hold:

$$Ee_1^{4m} < \infty, \tag{A.2}$$

$$\overline{\lim_{n \to \infty}} \sup_{h \in H_n} \lambda(M_n(h)) < \infty, \tag{A.3}$$

where $\lambda(M_n(h))$ denotes the maximum singular value of $M_n(h)$. Then $\hat{h}_M$ is asymptotically optimal.

(A.2) might be weakened at the expense of introducing more complicated proofs. (A.3) is usually satisfied unless the class of estimators $\hat{\underset{\sim}{\mu}}(h), h \in H_n$, is poorly motivated. In fact, if $\lambda(M_n(h)) > 1$ then $\hat{\underset{\sim}{\mu}}(h)$ is inadmissible and other better linear estimators can be easily constructed to replace $\hat{\underset{\sim}{\mu}}(h)$. We now turn to the application of this theorem.

Consider the model selection of Example 1. (A.3) holds obviously because $M_n(h)$ is a projection matrix. To see that (1.3) implies (A.1) with $m = 2$, observe that

$$nR_n(h) = ||A_n(h)\underset{\sim}{\mu}||^2 + h\sigma^2 \geq h\sigma^2.$$

Hence for any fixed natural number k

$$\sum_{h \in H_n} (nR_n(h))^{-2} \leq \sum_{h=1}^{k} (nR_n(h))^{-2} + \sigma^{-4} \sum_{h=k+1}^{p_n} h^{-2}$$

$$\leq k \left[\inf_{h \in H_n} nR_n(h) \right]^{-2} + \sigma^{-4} \sum_{h=k+1}^{\infty} h^{-2}.$$

Now by (1.3), we can chose $k \to \infty$ slowly enough so that $k(\inf nR_n(h))^{-2} \to 0$. This proves (A.1). We summarize the result by the following.

<u>Corollary 2.1</u>: For the model-selection setting of Example, C_p is

A.O., if (1.3) and (A.2) with $m = 2$ are satisfied.

Note that if σ^2 is replaced by a consistent estimate then the above corollary also holds.

Next consider the nearest neighbor nonparametric regression problem of Example 2. Under (1.4) and (1.5), (A.3) follows from Lemma 4.1 of Li (1983). Now we have the following desired result.

Corollary 2.2: For the nearest neighbor nonparametric regression problem of Example 2, $\hat{h}_M$ is A.O., if (1.6) and (A.2) hold.

3. GENERALIZED CROSS-VALIDATION

A simple way to derive GCV is by means of nil-trace linear estimates (N.T.L.E.) $\bar{\underset{\sim}{\mu}}(h) = \bar{M}_n(h)\underset{\sim}{y}_n$ where

$\bar{M}_n(h) = -\alpha_n(h)I + (1 + \alpha_n(h))M_n(h)$ with $\alpha_n(h) = \mathrm{tr}\ M_n(h)/\mathrm{tr}\ A_n(h)$. Clearly the trace of $\bar{M}_n(h)$ is zero and when applying C_L to the new class of estimates $\{\bar{\underset{\sim}{\mu}}(h)\colon h \in H_n\}$ we end up with GCV. The asymptotic justification of replacing $\hat{\underset{\sim}{\mu}}(h)$ by $\bar{\underset{\sim}{\mu}}(h)$ was given in Li (1983), Theorem 2.1. The following version of this theorem is more appropriate for further development.

Theorem 3.1. For any sequence of random variables $\hat{h}_n$, taking values in H_n, such that

$$\lim_{n \to \infty} p\{|n^{-1} \operatorname{tr} A_n(\hat{h}_n)| > \delta\} = 1, \quad \text{for some} \quad \delta > 0 \tag{3.1}$$

and

$$\lim_{n \to \infty} (n^{-1} \operatorname{tr} M_n(\hat{h}_n))^2 / n^{-1} \operatorname{tr} M_n(\hat{h}_n) M'_n(\hat{h}_n) = 0. \tag{3.2}$$

We have

$$n^{-1} || \bar{\underset{\sim}{\mu}}(\hat{h}_n) - \hat{\underset{\sim}{\mu}}(\hat{h}_n) ||^2 / R_n(\hat{h}_n) \to 0. \tag{3.3}$$

This theorem shows that under (3.1) and (3.2), N.T.L.E., $\bar{\underset{\sim}{\mu}}(\hat{h}_n)$, is a good approximation to the original estimate $\hat{\underset{\sim}{\mu}}(\hat{h}_n)$. Put $M_n(\hat{h}_n) = (a_{ij})_{i,j=1,\ldots,n}$. Clearly

$$(n^{-1} \operatorname{tr} M_n(\hat{h}))^2 = \left[\frac{1}{n} \sum_{i=1}^{n} a_{ii}\right]^2 \le \frac{1}{n} \sum_{i=1}^{n} a_{ii}^2 \le \frac{1}{n} \sum_{i,j} a_{ij}^2$$

$$= n^{-1} \operatorname{tr} M_n(\hat{h}) M_n(\hat{h}).$$

(3.2) amounts to saying that the weight a_{ii} of y_i for estimating its own mean μ_i should be small. This seems to be a prerequisite for consistency. For most settings (3.1) is weaker than (3.2). Therefore it is natural to anticipate that the A.O. property of $\hat{h}_G$ may follow from the result of Section 2 when applied to the class of N.T.L.E. Specifically we have

Theorem 3.2. Assume that (A.1) ~ (A.3) and the following conditions hold:

$$\inf_{h \in H_n} L_n(h) \to 0; \tag{A.4}$$

for any sequence $\{h_n \in H_n\}$ such that (A.5)

$n^{-1} \text{ tr } M_n(h_n)M'_n(h_n) \to 0$

we have $(n^{-1} \text{ tr } M_n(h_n))^2/n^{-1} \text{ tr } M_n (h_n)M'_n(h_n) \to 0;$

$$\sup_{h \in H_n} n^{-1} \text{ tr } M_n(h) \leq \gamma_1 \quad \text{for some} \quad 1 > \gamma_1 > 0; \tag{A.6}$$

$$\sup_{h \in H_n} (n^{-1} \text{ tr } M_n(h))^2/n^{-1} \text{ tr } M_n(h)M'_n(h) \leq \gamma_2 \tag{A.7}$$

for some $1 > \gamma_2 > 0$.

Then $\hat{h}_G$ is A.O.

(A.4) claims only the existence of a consistent selection procedure when $\underset{\sim}{\mu}_n$ is known. Since the risk of $\underset{\sim}{\hat{\mu}}(h)$, $R_n(h)$, is no less than its variance $\sigma^2 \text{ tr } M_n(h)M'_n(h)$, (A.5) implies that the good candidates (in terms of having small risks) in H_n, put a small weight on y_i for estimating its own expectation μ_i.

Now consider the model selection problem. Since $\text{tr } M_n(h) = \text{tr } M_n(h)M'_n(h) = h$, (A.5) is obviously satisfied. However

(A.6) and (A.7) require that the largest model has rank $p_n \leq n\gamma$ for some $0 < \gamma < 1$. But this constraint can be easily removed by the following arguments.

First let h^* be the minimizer of $\inf_{h \in H_n} L_n(h)$. (2.3), which follows from (A.1) - (A.3), implies that $R_n(h^*) \to 0$ because of (A.4). From this it follows that $h^* n^{-1} \to 0$. Therefore denoting $H'_n = H_n \prod \{h: h \leq n\gamma\}$, we see that the minimum loss does not increase for the restricted class H'_n: i.e., $\inf_{h \in H_n} L_n(h) = \inf_{h \in H'_n} L_n(h)$ except for a small probability that tends to 0 as $n \to \infty$. On the other hand, Li (1983) proved that $\underline{\hat{\mu}}(\hat{h}_G)$ is consistent providing that the following addition condition on the random errors hold:

there exists a constant k' so that for any $a \geq 0$, (3.4)

$$\sup_{x \in R} P\{x - a \leq e_1 \leq x + a\} \leq k'a.$$

(A.8) is satisfied if e_1 has a bounded density. Using this result and the arguments for h^* before, we see that $\hat{h}_G n^{-1} \to 0$. Thus asymptotically $\hat{h}'_G$, the model selected by GCV when the class of models considered is restricted to H'_n, will be the same as $\hat{h}_G$, the model selected from the entire class H_n. Therefore we see that it is not necessary to have the condition $P_n \leq n\gamma$. The following theorem conveys the result we have established.

<u>Corollary 3.1</u>: For the model selection problem of Example 1, $\hat{h}_G$

is A.O. if (1.3), (A.2) with $m = 2$, (A.4) and (3.4) hold.

Next, turning to the nearest neighbor nonparametric problem, observe that n^{-1} tr $M_n(h) = w_{n,h}(1)$ and n^{-1} tr $M_n(h)M'_n(h) = \sum_{i=1}^{h} w_{n,h}(i)^2 \geq w_{n,h}(1)^2 + h^{-1}(1 - w_{n,h}(1))^2$. Thus it is clear that the following condition implies (A.5):

there exist fixed positive numbers λ_1 and λ_2 such that (3.5)
$w_{n,h}(1) \leq \lambda_1 h^{-(1/2+\lambda_2)}$ for any n, h.

This condition was used in Li (1983) and can be easily satisfied by most commonly used weights; for example, uniform weight, $w_{n,h}(i) = h^{-1}$. In addition, (A.6) is also a reasonable restriction on the weight functions providing that $H_n = \{2,\ldots,n\}$ (note that GCV is undefined for $h = 1$ because $||\underset{\sim}{y}_n - \hat{\mu}_n(1)||^2 = 0$ and $1 - n^{-1}$ tr $M_n(1) = 0$.) It reduces to the following condition:

$$\sup_{h=2,\ldots,n} w_{n,h}(1) \leq \gamma_1 \text{ for some } \gamma_1, \; 0 < \gamma_1 < 1. \qquad (3.6)$$

Finally it is obvious that (3.5) and (3.6) imply (A.7). Therefore we obtain the following desired result.

Corollary 3.2: Suppose that the weight functions satisfy the regularity conditions of (1.4), (1.5), (3.5) and (3.6). Then $\hat{h}_G$ is A.O., if (A.2), (A.4) and (1.6) hold and $H_n = \{2,\ldots,n\}$.

4. STEIN ESTIMATES

The replacement of $\hat{\boldsymbol{\mu}}_n(h)$ by $\bar{\boldsymbol{\mu}}_n(h)$ does not seem appropriate if n^{-1} tr $M_n(h)$ is not negligible. This is a weak point for considering $\bar{\boldsymbol{\mu}}_n(h)$. A better viewpoint is by means of Stein estimates and the associated unbiased risk estimates, defined by

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

and

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

The original version of these quantities given in Stein (1981) was a little complicated and only for a symmetric $A_n(h)$. Stein estimates possess the nice property that they dominate the raw data $\boldsymbol{y}_n$ as estimates of $\boldsymbol{\mu}_n$ under the normality of the error distribution and some mild assumption about the largest characteristic root of $A_n(h)$. Li and Hwang (1984) studied the asymptotic behavior of $\tilde{\boldsymbol{\mu}}_n(h)$ for the nonparametric regression problem. Basically $\tilde{\boldsymbol{\mu}}_n(h)$ and $\hat{\boldsymbol{\mu}}_n(h)$ will be very close to each other providing that $\hat{\boldsymbol{\mu}}_n(h)$ is very close to the true value $\boldsymbol{\mu}_n$. Hence using Stein estimates we do not lose any efficiency if it is the case that the corresponding linear estimate performs well; if not, by the property of bounded risks, we still have some guarantee that estimation error may not be as big as the linear ones which usually have unbounded risks. This justifies the replacement of $\tilde{\mu}(h)$ by $\tilde{\mu}_n(h)$. Now it is easy to see the interesting result that the natural way of selecting $\tilde{\mu}_n(h)$, minimizing $\mathrm{SURE}_n(h)$, is exactly the same as GCV. Li (1984) argued that $\mathrm{SURE}_n(h)$, initially proposed as an estimate of the risk of Stein estimate $\tilde{\mu}_n(h)$, indeed does more than anticipated: it is always a consistent estimate of the true loss $n^{-1}||\mu_n - \tilde{\mu}_n(h)||^2$ although sometimes the true loss does not converge (hence for this case $\mathrm{SURE}_n(h)$ cannot be a consistent estimate of the risk $En^{-1}||\mu_n(h) - \tilde{\mu}_n(h)||^2$). In addition, the consistency is uniform for $\mu_n \in R^n$. The consistency of the G-cross-validated Stein estimate $\tilde{\mu}_n(\hat{h}_G)$ was also established there. The following theorem strengthens this result by proving the asymptotic efficiency of $\tilde{\mu}_n(\hat{h}_G)$.

<u>Theorem 4.1</u>. Under the assumptions of Theorem 3.1, we have

$$n^{-1}||\tilde{\mu}_n(\hat{h}_G) - \mu_n||^2 / \inf_{h \in H_n} L_n(h) \to 1$$

in probability.

As in Section 3, Theorem 4.1 applies to the model selection and nearest neighbor nonparametric regression.

5. CROSS-VALIDATION.

Let $\underline{\mu}^c(h)$ denote the delete-one estimate of $\underline{\mu}_n$, $M_n(h)\underline{y}_n$. Intuitively $\underline{\mu}^c(h)$ will be only slightly different from $\underline{\hat{\mu}}(h)$ when the sample size is large. Since diagonal elements of $M_n(h)$ are all zero now, we see that C.V. is just the C_L applied to $\{\underline{\mu}^c(h), h \in H_n\}$. Thus one may use the results of Section 2 to establish the A.O. property of $\hat{h}_C$. However a rigorous proof requires suitable conditions on $M_n(h)$ to ensure that the difference between $\underline{\hat{\mu}}(h)$ and $\underline{\mu}^c(h)$, $n^{-1}||\underline{\hat{\mu}}(h) - \underline{\mu}^c(h)||^2$, is negligible in comparison with the loss $n^{-1}||\underline{\mu}_n - \underline{\hat{\mu}}(h)||^2$. For the case of nearest neighbor nonparametric regression we have the following theorem.

<u>Theorem 5.1</u>. Under the assumptions of Corollary 3.2, $\hat{h}_C$ is A.O.

For the model selection problem, we have not obtained any useful general results yet. But observe that

$$n^{-1}||\underline{y}_n - \underline{\mu}^c(h)||^2 = n^{-1} \sum_{i=1}^{n} (y_i - \hat{\mu}_{ni}(h))^2(1 - a_{ni}(h))^{-2}$$

where $\hat{\mu}_{ni}(h)$ is the ith coordinate of $\underline{\hat{\mu}}_n(h)$ and $a_{ni}(h)$ is the ith diagonal element of $M_n(h)$. Compared with G.C.V., it seems that if $a_{ni}(h)$, $i = 1, \ldots, n$, are very close to each other, then C.V. and G.C.V. may be almost equivalent. Hence the results of G.C.V. may be used to justify C.V.

6. PROOFS

Proof of Theorem 2.1. We shall prove (2.1) first. Given any $\delta > 0$, by Chebychev inequality we have

$$P\{\sup_{h \in H_n} n^{-1}|\langle \underset{\sim}{e}_n, A_n(h)\underset{\sim}{\mu}_n\rangle|/R_n(h) > \delta\}$$

$$\leq \sum_{h \in H_n} \frac{n^{-2m}E\langle e_n, A_n(h)\underset{\sim}{\mu}_n\rangle^{2m}}{\delta^{2m}R_n(h)^{2m}}$$

;

$$\leq C \cdot \delta^{-2m} \sum_{h \in H_n} n^{-2m} \cdot ||A_n(h)\underset{\sim}{\mu}_n||^{2m} \cdot R_n(h)^{-2m},$$

for some constant $C > 0$. Now since $n^{-1}||A_n(h)\underset{\sim}{\mu}_n||^2 \leq R_n(h)$, the last expression does not exceed $C\delta^{-2m} \sum_{h \in H_n} (nR_n(h))^{-m}$, which tends to 0 by (A.1). Thus (2.1) is proved. (2.2) can be established in a similar manner by noting that

$$E(\sigma^2 \text{ tr } M_n(h) - \langle \underset{\sim}{e}_n, M_n(h)\underset{\sim}{e}_n\rangle)^{2m} \leq C'(\text{tr } M_n(h)M'_n(h))^m$$

for some $C' > 0$ and that $\sigma^2 n^{-1} \text{ tr } M_n(h)M'_n(h) \leq R_n(h)$. Finally it is clear that (2.3) will follow from the following two statements:

$$\sup_{h \in H_n} n^{-1}|\langle A_n(h)\underset{\sim}{\mu}_n, M_n(h)\underset{\sim}{e}_n\rangle|/R_n(h) \to 0 \tag{6.1}$$

and

$$\sup_{h \in H_n} n^{-1}|\ ||M_n(h)\underset{\sim}{e}_n||^2 - \sigma^2 \text{ tr } M_n(h)M'_n(h)|/R_n(h) \to 0. \tag{6.2}$$

Since $\langle A_n(h)\underset{\sim}{\mu}_n, M_n(h)\underset{\sim}{e}_n\rangle = \langle M'_n(h)A_n(h)\underset{\sim}{\mu}_n, \underset{\sim}{e}_n\rangle$ and $||M'_n(h)A_n(h)\underset{\sim}{\mu}_n||^2 \le \lambda(M_n(h))^2||A_n(h)\underset{\sim}{\mu}_n||^2$, the proof of (6.1) will be the same as that of (2.1) in view of (A.3). Similarly, write $||M_n(h)\underset{\sim}{e}_n||^2 = \langle M'_n(h)M_n(h)\underset{\sim}{e}_n, \underset{\sim}{e}_n\rangle$ and observe that $\text{tr}(M'_n(h)M_n(h))^2 \le \lambda(M_n(h))^2 \text{ tr } M'_n(h)M_n(h)$. We see that (6.2) can be proved exactly as (2.2). This completes the proof of Theorem 2.1.

□

Proof of Theorem 3.1. Similar to the proof of Theorem 2.1, Li (1983).

□

Proof of Theorem 3.2. Put $\bar{L}_n(h) = n^{-1}||\bar{\underset{\sim}{\mu}}(h) - \underset{\sim}{\mu}||^2$ and $\bar{R}_n(h) = E\bar{L}_n(h)$. A simple computation leads to

$$\text{tr } \bar{M}_n(h)\overline{M_n(h)}' = \frac{\text{tr } M_n(h)M'_n(h) - n^{-1}(\text{tr } M_n(h))^2}{(n^{-1} \text{ tr } A_n(h))^2},$$

and

$$nR_n(h) = \frac{||A_n(h)\underline{\mu}||^2 + \mathrm{tr}\ M_n(h)M'_n(h) - n^{-1}(\mathrm{tr}\ M_n(h))^2}{(n^{-1}\ \mathrm{tr}\ A_n(h))^2} . \qquad (6.3)$$

Now by (A.6) and (A.7) it is easy to check that (A.1) and (A.3) hold when $R_n(h)$ and $M_n(h)$ are replaced by $\bar{R}_n(h)$ and $\bar{M}_n(h)$ respectively. Hence from Theorem 2.1 we see that

$$\bar{L}_n(\hat{h}_G)/\inf_{h \in H_n} \bar{L}_n(h) \to 1. \qquad (6.4)$$

In fact, the following analogue of (2.3) also holds:

$$\sup_{h \in H_n} |\bar{L}_n(h)/\bar{R}_n(h) - 1| \to 0. \qquad (2.3')$$

Let h_* be the minimizer of $L_n(h)$ over $h \in H_n$. From (6.4) it is clear that (1.2) will hold for $\hat{h} = \hat{h}_G$ if we can verify that

$$\bar{L}_n(h_*)/L_n(h_*) \to 1 \qquad (6.5)$$

and

$$\bar{L}_n(\hat{h}_G)/L_n(\hat{h}_G) \to 1. \qquad (6.6)$$

Theorem 3.1 can be used to prove (6.5) and (6.6). To see this, first observe that (3.1) always holds because of (A.6). Next, from (2.3) and (A.4), it follows that n^{-1} tr $M_n(h_*)M_n(h_*)' \to 0$. Hence by (A.5), (3.2) holds for $\hat{h}_n = h_*$. Theorem 3.1 applies for $\hat{h}_n = h_*$. (6.5) follows as a simple consequence of (3.3) and (2.3). Similarly, (6.6) will hold if we can show that n^{-1} tr $M_n(\hat{h}_G)M_n(\hat{h}_G)' \to 0$. Now by (6.5) and (A.4) we see that $\bar{L}_n(h_*) \to 0$, which implies $\bar{L}_n(\hat{h}_G) \to 0$ because of (6.4). Finally from (2.3') we conclude that $\bar{R}_n(\hat{h}_G) \to 0$ which in turn implies n^{-1} tr $\bar{M}_n(\hat{h}_G)\overline{M'}_n(\hat{h}_G) \to 0$ as desired. This completes the proof of Theorem 3.2.

□

<u>Proof of Theorem 4.1.</u> By Theorem 3.2 it suffices to show that

$$n^{-1}||\bar{\underset{\sim}{\mu}}_n(\hat{h}_G) - \hat{\underset{\sim}{\mu}}_n(\hat{h}_G)||^2/L_n(\hat{h}_G) \to 0. \qquad (6.7)$$

First observe that

$$n^{-1}||\bar{\underset{\sim}{\mu}}_n(\hat{h}_G) - \hat{\underset{\sim}{\mu}}_n(\hat{h}_G)||^2 = \left[\frac{n^{-1}\sigma^2 \text{ tr } A_n(\hat{h}_G)}{n^{-1}||A_n(\hat{h}_G)\underset{\sim}{y}||^2} - 1\right]^2 \cdot n^{-1}||A_n(\hat{h}_G)\underset{\sim}{y}||^2$$

$$= [(n^{-1}\sigma^2 - n^{-1}||\underset{\sim}{e}_n||^2) - L_n(\hat{h}_G) - 2n^{-1}\langle\underset{\sim}{e}_n, \underset{\sim}{\mu} - \hat{\underset{\sim}{\mu}}(\hat{h}_G)\rangle$$

$$- n^{-1}\sigma^2 \text{ tr } M_n(\hat{h}_G)]^2/n^{-1}||A_n(\hat{h}_G)\underset{\sim}{y}||^2$$

and that

$$n^{-1}||A_n(\hat{h}_G)\underset{\sim}{y}||^2 = n^{-1}||\underset{\sim}{e}_n + (\underset{\sim}{\mu}_n - \hat{\underset{\sim}{\mu}}(\hat{h}_G))||^2 \to \sigma^2$$

because of the consistency of $\hat{\underset{\sim}{\mu}}(\hat{h}_G)$. Therefore to prove (6.7) it is enough to verify

$$(n^{-1}\sigma^2 - n^{-1}||\underset{\sim}{e}_n||^2)^2/L_n(\hat{h}_G) \to 0 \tag{6.8}$$

$$(n^{-1}<\underset{\sim}{e}_n, \underset{\sim}{\mu} - \hat{\underset{\sim}{\mu}}(\hat{h}_G)>)^2/L_n(\hat{h}_G) \to 0 \tag{6.9}$$

and

$$(n^{-1} \text{ tr } M_n(\hat{h}_G))^2/L_n(\hat{h}_G) \to 0. \tag{6.10}$$

Now (1.3), which is weaker than (A.1), and (2.3) imply that $nL_n(\hat{h}_G) \to \infty$. Thus (6.8) follows from the central limit theorem. Next as was proved in the proof of Theorem 3.2, (3.2) holds for $\hat{h}_n = \hat{h}_G$. This together with (2.3) implies (6.10). Finally to prove (6.9), it suffices to show

$$(n^{-1}<\underset{\sim}{e}_n, A_n(\hat{h}_G)\underset{\sim}{\mu}_n>)^2/R_n(\hat{h}_G) \to 0 \tag{6.11}$$

and

$$(n^{-1}<\underset{\sim}{e}_n, M_n(\hat{h}_G)\underset{\sim}{e}_n>)^2/R_n(\hat{h}_G) \to 0. \tag{6.12}$$

It is not difficult to see that (6.11) follows from (2.1) and that (6.12) follows from (2.2), (6.10) and (2.3). This completes the proof of Theorem 4.1.

Proof of Theorem 5.1. Let $\bar{L}_n(h) = n^{-1}||\underset{\sim}{\mu}_n - \underset{\sim}{\mu}^c(h)||^2$ and $\bar{R}_n(h) = E\bar{L}_n(h)$. The following useful statement will be proved late:

$$\text{if} \quad h_n \to \infty \quad \text{then} \quad \bar{R}_n(h_n)/R_n(h_n) \to 1. \tag{6.13}$$

First Theorem 2.1 can be used to show that

$$\bar{L}_n(\bar{h}_C)/\inf_{h \in H_n} \bar{L}_n(h) \to 1, \tag{6.14}$$

and

$$\sup_{h \in H_n} |\bar{L}_n(h)/\bar{R}_n(h) - 1| \to 0, \tag{6.15}$$

providing that (1.6) and (A.3) also hold when $R_n(h)$ and $M_n(h)$ are replaced by $\bar{R}_n(h)$ and $\bar{M}_n(h)$ respectively. Denote the minimizers of $\inf R_n(h)$ and $\inf \bar{R}_n(h)$ by h_* and $\bar{h}_*$ respectively. Now from (A.4) and (2.3), we see that $h_* \to \infty$ and $R_n(h_*) \to 0$. Thus from (6.13) it follows that $R_n(\bar{h}_*) \to 0$, which clearly implies $\bar{h}_* \to 0$. By (6.13) again we see that $\bar{R}_n(\bar{h}_*)/R_n(h_*) \to 1$, Therefore (1.6) also holds when $R_n(h)$ is replaced by $\bar{R}_n(h)$. On the other hand the arguments in the proof of Lemma 4.1 of Li (1983) can be used to show that (A.3) also holds with $M_n(h)$ replaced by $\bar{M}_n(h)$. Now by (6.14), (6.15) and (2.3), it is enough to show that $\bar{R}_n(\hat{h}_C)/R_n(\hat{h}_C) \to 1$. This again will follow from (6.13) providing that $\hat{h}_C \to \infty$. Finally, from (6.14) and (6.15) it follows that $\bar{R}_n(\hat{h}_C) \to 0$, which clearly implies $\hat{h}_C \to \infty$ as desired. This completes the proof of Theorem 5.1.

<u>Proof of (6.13)</u>. Let $f_\infty = \sup\limits_{\underset{\sim}{x} \in X} f(\underset{\sim}{x})$. By the definition of

$\bar{M}_n(h)$,

$$n^{-1}||(M_n(h) - \bar{M}_n(h)\underset{\sim}{\mu}_n||^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=2}^{h}(w_{n,h}(j) - w_{n,h}(j-1))f(\underset{\sim}{x}_{i(j)})\right.$$

$$\left. + w_{n,h}(1)f(\underset{\sim}{x}_{i(j)}) - w_{n,h}(h)f(\underset{\sim}{x}_{i(h+1)})\right]^2$$

$$\le \left[\sum_{j=2}^{h}(w_{n,h}(j-1) - w_{n,h}(j)) + w_{n,h}(1) + w_{n,h}(h)\right]^2 f_\infty^2$$

$$\le 4w_{n,h}(1)^2 f_\infty^2,$$

where the first inequality follows from (1.5). Now compare

$\bar{R}_n(h) = n^{-1}||\underset{\sim}{\mu}_n - \bar{M}_n(h)\underset{\sim}{\mu}||^2 + \sigma^2 \sum\limits_{i=1}^{n} w_{n,h}(i)^2$ with

$R_n(h) = n^{-1}||\underset{\sim}{\mu}_n - M_n(h)\underset{\sim}{\mu}||^2 + \sigma^2 \sum\limits_{i=1}^{h} w_{n,h}(i)^2$. We see that

$\bar{R}_n(h)/R_n(h) \to 1$ if $w_{n,h}(1)^2/\sum\limits_{i=1}^{h} w_{n,h}(i)^2 \to 0$, or, due to (3.5) and the fact that $\sum_{i=1}^{h} w_{n,h}(i)^2 \geq h^{-1}$, if $h \to \infty$. This completes the proof of (6.13).

REFERENCES

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. Technometrics, **16**, 125-127.

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

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.

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

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

Li, K. C. (1983). From Stein's unbiased risk estimates to the method of generalized cross-validation. Technical Report #83-34, Department of Statistics, Purdue University.

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

Li, K. C., and Hwang, J. (1984). The data smoothing aspect of Stein estimates. Ann. Statist., to appear.

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

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

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

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

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

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. *International Statist. Review.* **46**, 1-49 and 129-146.

Asymptotic Optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set*

Short title: C_p, C_L, Cross-validation and G. C. V.

by

Ker-Chau Li

Department of Mathematics, U.C.L.A.

and

Department of Statistics, Purdue University

* This work is sponsored by National Science Foundation under Grant No. MCS-8200631.

AMS 1980 Subject Classification: Primary 62G99, 62J99

Secondary 62J05, 62J07

Key Words and Phrases: model-selection; nearest neighbor estimates; nil-trace linear estimates; non-parametric regression; Stein estimates; Stein's unbiased risk estimates.