The Annals of Statistics
1987, Vol. 15, No. 3, 958–975

ASYMPTOTIC OPTIMALITY FOR C_p, C_L, CROSS-VALIDATION AND GENERALIZED CROSS-VALIDATION: DISCRETE INDEX SET[1]

BY KER-CHAU LI

University of California, Los Angeles

C_p, C_L, cross-validation and generalized cross-validation are useful data-driven techniques for selecting a good estimate from a proposed class of linear estimates. The asymptotic behaviors of these procedures are studied. Some easily interpretable conditions are derived to demonstrate the asymptotic optimality. It is argued that cross-validation and generalized cross-validation can be viewed as some special ways of applying C_L. Applications in nearest-neighbor nonparametric regression and in model selection are discussed in detail.

1. Introduction. Let $\mathbf{y}_n = (y_1, y_2, \ldots, y_n)'$ be a vector of n independent observations with unknown mean $\mu_n = (\mu_1, \mu_2, \ldots, \mu_n)'$. Write

$$y_i = \mu_i + e_i, \qquad i = 1, 2, \ldots, n,$$

and assume that the random errors e_i are identically distributed with mean 0 and variance σ^2. Suppose that to estimate μ_n, a class of linear estimators $\hat{\mu}_n(h) = M_n(h)\mathbf{y}_n$, 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 the y_i's, our concern is to select an $\hat{h}$ from H_n so that the average squared error $L_n(\hat{h}) = n^{-1}\|\mu_n - \hat{\mu}_n(\hat{h})\|^2$ may be as small as possible ($\|\cdot\|$ denotes the Euclidean norm).

EXAMPLE 1. Model selection: Suppose associated with y_i there are p_n explanatory variables $x_{i1}, x_{i2}, \ldots, x_{ip_n}$, arranged in decreasing order of importance. To estimate μ_n, one may employ the first h variables to propose 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 $\hat{\mu}(h) = X_h(X_h'X_h)^{-1}X_h'\mathbf{y}_n$ to estimate μ_n. Here X_h denotes the $n \times h$ design matrix (x_{ij}) and the information matrix $X_h'X_h$ is assumed to be nonsingular. Clearly, $M_n(h) = X_h(X_h'X_h)^{-1}X_h$ is a projection matrix of rank h. We may take $H_n = \{1, 2, \ldots, p_n\}$. Our goal is to determine an appropriate model for the purpose of estimating μ_n.

EXAMPLE 2. Nearest-neighbor nonparametric regression: 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 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, which become dense in

Received May 1985; revised September 1986.

[1]This work was 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. Model-selection, nearest-neighbor estimates, nil-trace linear estimates, nonparametric regression, Stein estimates, Stein's unbiased risk estimates.

X as $n \to \infty$. Assume that $\mu_i = f(\mathbf{x}_i)$ for an unknown continuous function f on X. An h-nearest-neighbor estimate of f at $\mathbf{x}_i$ depends only on the h observations whose $\mathbf{x}$ values are closest to $\mathbf{x}_i$. Let $\mathbf{x}_{i(j)}$ denote the jth nearest neighbor of $\mathbf{x}_i$ in the sense that $\|\mathbf{x}_i - \mathbf{x}_{i(j)}\|$ is the jth smallest number among the n values $\|\mathbf{x}_i - \mathbf{x}_{i'}\|$, $i' = 1, 2, \ldots, n$. Ties may be broken in any systematic manner. For a given weight function $w_{n,h}(\cdot)$, the h-nearest-neighbor estimate of μ_i is $\hat{\mu}_i(h) = \sum_{j=1}^{h} w_{n,h}(j) y_{i(j)}$. Thus, $\hat{\boldsymbol{\mu}}_n(h) = (\hat{\mu}_1(h), \ldots, \hat{\mu}_n(h))'$ takes the form $M_n(h)\mathbf{y}_n$ with each row of $M_n(h)$ being some permutation of the vector of the weights $(w_{n,h}(1), \ldots, w_{n,h}(h), 0, \ldots, 0)$. Conditions on the weight function will be given later. Stone (1977) gave extensive studies on the asymptotic behavior of this procedure for a deterministically chosen sequence of h. For our purpose, we may take $H_n = \{1, 2, \ldots, n\}$. It is then desired to use the same data to decide the number of neighbors that should enter into our estimate.

For these two examples, the index set H_n is discrete and has finite cardinality. Examples with continuous H_n, including ridge regression, smoothing splines and kernel nonparametric regression, will not be treated here. But some basic results or principles derived in this paper are expected to carry over [see Li (1986) for results on ridge regression and smoothing splines].

The following three well-known procedures of selecting h will be studied in this paper.

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

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

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

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

(iii) (Delete-one) cross-validation [Allen (1974), Stone (1974), Geisser (1975) and Wahba and Wold (1975)]: Select $\hat{h}$, denoted by $\hat{h}_C$, that minimizes the sum of squared prediction errors for y_i with y_i itself being excluded from the data set. A rigorous definition of this (delete-one) procedure requires the specification of estimators (or predictors) to be used when the sample size is $n - 1$. But formally, given $y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$, we may write the predictor of y_i (or the estimator of μ_i) as $\hat{y}_{-i} = \sum_{j=1}^{n} \tilde{m}_{ij}(h) y_j$ with $\tilde{m}_{ii}(h)$ being zero. Then $\hat{h}_c$ achieves

$$\min_{h \in H_n} \left\|\mathbf{y}_n - \tilde{M}_n(h)\mathbf{y}_n\right\|^2, \tag{1.3}$$

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

EXAMPLE 1. Model selection (continued): Since $M_n(h)$ is a projection with rank h, the C_L procedure reduces to the more famous C_p criterion which selects

$\hat{h}$ that minimizes

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

On the other hand, GCV selects $\hat{h}$ by minimizing

$$\min_{h\in H_n} n(n-h)^{-2}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2.$$

GCV happens to take a form almost identical to another procedure S_p proposed in Hocking (1976), and Thompson (1978), which selects $\hat{h}$ by minimizing $(n-1)(n-h)^{-1}(n-h-1)^{-1}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2$. S_p was motivated by treating the response variable y and the explanatory variable $\mathbf{x}$ jointly as a multivariate normal random variable in a prediction problem. This aspect of S_p is further explored in Breiman and Freedman (1983). For (delete-one) cross-validation, y_i is predicted by $\hat{y}_{-i} = \mathbf{x}_i'(X_{h,-i}'X_{h,-i})^{-1}X_{h,-i}'\mathbf{y}_{n,-i}$, where $X_{h,-i}$ is the $(n-1)\times h$ submatrix of X_h with the ith row deleted and $\mathbf{y}_{n,-i}$ is the subvector of $\mathbf{y}_n$ with the ith coordinate deleted. Let $D_n(h)$ be an $n\times n$ diagonal matrix with the ith diagonal element equal to $(1-m_i(h))^{-1}$, where $m_i(h)$ is the ith diagonal matrix of $M_n(h)$. Then the delete-one estimate of $\boldsymbol{\mu}_n$ takes the form $\tilde{M}_n(h)\mathbf{y}_n$ with

$$\tilde{M}_n(h) = D_n(h)(M_n(h) - I) + I, \tag{1.4}$$

where I denotes the $n\times n$ identity matrix. To see this, first observe that a simple application of the Gauss–Markov theorem implies that the least-squares estimate $\hat{\mu}_i(h)$ of μ_i (given the model h) based on the whole sample $y_1,\ldots,y_n$ is the best linear combination of y_i and the delete-one estimate $\hat{y}_{-i}$. From this observation, it follows that

$$\hat{\mu}_i(h) = m_i(h)y_i + \sum_{j\neq i} m_{ij}(h)y_j = m_i(h)y_i + (1-m_i(h))\hat{y}_{-i}$$

[the weight for $\hat{y}_{-i}$ has to be $1-m_i(h)$, otherwise $\hat{\mu}_i(h)$ will be biased]. Therefore $\hat{y}_{-i} = (1-m_i(h))^{-1}\hat{\mu}_i(h) - (1-m_i(h))^{-1}m_i(h)y_i$, yielding (1.4). The cross-validation criterion (1.3) amounts to the following:

$$\min_{h\in H_n} n^{-1}\sum_{i=1}^{n}(1-m_i(h))^{-2}(y_i - \hat{\mu}_i(h))^2. \tag{1.5}$$

Note that the ith residual $y_i - \hat{\mu}_i(h)$ has variance $(1-m_i(h))\sigma^2$.

EXAMPLE 2. Nearest-neighbor nonparametric regression (continued): Since each $\mathbf{x}_i$ has itself as the first nearest neighbor, it follows that all the diagonal elements of $M_n(h)$ are equal to $w_{n,h}(1)$. C_L criterion (1.1) amounts to the following:

$$\min_{h\in H_n} n^{-1}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2 + 2\sigma^2 w_{n,h}(1).$$

The GCV criterion (1.2) becomes

$$\min_{h\in H_n} \frac{n^{-1}\|\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2}{(1-w_{n,h}(1))^2}.$$

For (delete-one) cross-validation, the delete-one estimate $\hat{y}_{-i}$ equals $\sum_{j=1}^{h} w_{n,h}(j) y_{i(j+1)}$ (for another possible definition see Remark 5.1 in Section 5.2). From this, we see that each row of $\tilde{M}_n(h)$ is again some permutation of the weights $w_{n,h}(j)$; but the diagonal elements of $\tilde{M}_n(h)$ are identically zero.

The primary goal of this paper is to demonstrate that under reasonable conditions, these procedures are asymptotically optimal in the sense that

$$\frac{L_n(\hat{h})}{\inf_{h \in H_n} L_n(h)} \to 1, \quad \text{in probability.} \tag{1.6}$$

Thus using these procedures, statisticians may do as well as if they knew the true μ_n [but are restricted to the use of the linear estimators $\hat{\mu}_n(h)$]. For brevity, we shall omit the phrase "in probability."

Let $R_n(h) = EL_n(h)$, the expected average-squared error. Denote the maximum singular value of $M_n(h)$ by $\lambda(M_n(h))$. Section 2 proves the asymptotic optimality of C_L criterion under the conditions that

$$\limsup_{n \to \infty} \sup_{h \in H_n} \lambda(M_n(h)) < \infty, \tag{A.1}$$

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

$$\sum_{h \in H_n} (nR_n(h))^{-m} \to 0, \tag{A.3}$$

for some natural number m. Condition (A.1) is quite natural. In fact, if $\lambda(M_n(h)) > 1$, then $\hat{\mu}(h)$ is inadmissible and is dominated by some other linear estimators [Cohen (1966)]. To explain (A.3), we first observe that in nonparametric regression or in parametric regression with the number of unknown parameters growing up to infinity as the sample size increases, the optimal risk $R_n(h)$ is typically of order $n^{-1+\delta}$, for some $\delta > 0$. Thus, if the cardinality of H_n is of polynomial order, $n^{\delta'}$, for some δ', then we can always find an m so that (A.3) is satisfied. (A.2) is just a moment condition on the error distribution which may not be satisfied if distribution-robustness is of concern (but that is a separate issue and will not be discussed here). For the model selection problem of Example 1, we need only to take $m = 2$ in (A.2) and replace (A.3) by the weaker condition (see Corollary 2.1)

$$\inf_{h \in H_n} nR_n(h) \to \infty. \tag{A.3$'$}$$

The assumption (A.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, (A.3$'$) will hold when the true regression function is not a polynomial [Shibata (1981)].

The asymptotic optimality of GCV is established in Section 3. This is based on a connection between C_L and GCV via the notion of nil-trace linear estimator [Li (1985), page 1357, Remark 1].

Li (1985) provides another way to look at GCV using Stein estimates and Stein's unbiased risk estimates [Stein (1981)]. Section 4 will strengthen the results in Li (1985) by showing that the Stein estimate selected by GCV is asymptotically optimal.

Section 5 discusses the asymptotic optimality of cross-validation. A general viewpoint is that cross-validation is closely related to C_L. In fact, since trace $\tilde{M}_n(h) = 0$ we see that cross-validation is just the C_L procedure applied to the class of delete-one estimates $\{\tilde{M}_n(h)\mathbf{y}_n\colon\ h \in H_n\}$. Thus, we have strong reasons to believe that the asymptotic optimality for cross-validation should follow from that for C_L, providing that the delete-one estimate is very close to the initial estimate. Examples 1 and 2 are treated in detail.

Technical proofs will be given in Section 6.

2. C_p and C_L. Let $\mathbf{e}_n = (e_1, e_2, \ldots, e_n)'$ and $A_n(h) = I - M_n(h)$, where I is the $n \times n$ identity matrix. First, observe the identity

$$\begin{aligned} &n^{-1}\|\mathbf{y}_n - \hat{\mu}_n(h)\|^2 + 2\sigma^2 n^{-1}\operatorname{tr} M_n(h) \\ &\quad = n^{-1}\|\mathbf{e}_n\|^2 + L_n(h) + 2n^{-1}\langle \mathbf{e}_n, A_n(h)\mu_n\rangle \\ &\qquad + 2n^{-1}\big(\sigma^2 \operatorname{tr} M_n(h) - \langle \mathbf{e}_n, M_n(h)\mathbf{e}_n\rangle\big). \end{aligned} \tag{2.1}$$

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

$$L_n(h) + 2n^{-1}\langle \mathbf{e}_n, A_n(h)\mu_n\rangle + 2n^{-1}\big(\sigma^2 \operatorname{tr} M_n(h) - \langle \mathbf{e}_n, M_n(h)\mathbf{e}_n\rangle\big)$$

over $h \in H_n$. If we can show that $n^{-1}\langle \mathbf{e}_n, A_n(h)\mu_n\rangle$ and $n^{-1}(\sigma^2 \operatorname{tr} M_n(h) - \langle \mathbf{e}_n, M_n(h)\mathbf{e}_n\rangle)$ are negligible [compared with $L_n(h)$] uniformly for any $h \in H_n$, then the asymptotic optimality property (1.6) is established for $\hat{h} = \hat{h}_M$. More precisely, it remains to show that in probability,

$$\sup_{h \in H_n} n^{-1}\langle \mathbf{e}_n, A_n(h)\mu_n\rangle / R_n(h) \to 0, \tag{2.2}$$

$$\sup_{h \in H_n} n^{-1}\big|\sigma^2 \operatorname{tr} M_n(h) - \langle \mathbf{e}_n, M_n(h)\mathbf{e}_n\rangle\big| / R_n(h) \to 0 \tag{2.3}$$

and

$$\sup_{h \in H_n} |L_n(h)/R_n(h) - 1| \to 0. \tag{2.4}$$

These statements will be verified in Section 6. Now we have

THEOREM 2.1. *Assume* (A.1)–(A.3) *hold. Then* C_L *is asymptotically optimal; i.e.,* (1.6) *holds for* $\hat{h}_M$, *defined by* (1.1).

The following two subsections are devoted to the application of this theorem in Examples 1 and 2.

2.1. *Model selection—Example* 1 (*continued*). Consider the model selection of Example 1. (A.1) holds obviously because $M_n(h)$ is a projection matrix. To see

that (A.3′) implies (A.3) with $m = 2$, observe that

$$nR_n(h) = \|A_n(h)\mu\|^2 + h\sigma^2 \geq h\sigma^2. \tag{2.5}$$

Hence for any fixed natural number k

$$\begin{aligned}\sum_{h\in H_n} (nR_n(h))^{-2} &\leq \sum_{h=1}^{k} (nR_n(h))^{-2} + \sigma^{-4} \sum_{h=k+1}^{p_n} h^{-2} \\ &\leq k\Big(\inf_{h\in H_n} nR_n(h)\Big)^{-2} + \sigma^{-4} \sum_{h=k+1}^{\infty} h^{-2}.\end{aligned} \tag{2.6}$$

Now by (A.3′), we can choose $k \to \infty$ slowly enough so that $k(\inf nR_n(h))^{-2} \to 0$. On the other hand, since $\Sigma_{h=1}^{\infty} h^{-2}$ is convergent, the last term in the above display converges to zero as k tends to ∞. This proves (A.3). We summarize the result by the following.

COROLLARY 2.1. *For the model-selection setting of Example* 1, C_p *is asymptotic optimal if* (A.3′) *and* (A.2) *with* $m = 2$ *are satisfied.*

When σ^2 is unknown, we should replace it by an estimate in (1.1).

COROLLARY 2.2. *If* σ^2 *is replaced by a consistent estimate* $\hat{\sigma}^2$, *then* C_p *is still asymptotically optimal under* (A.3′) *and* (A.2) *with* $m = 2$.

PROOF. The identity (2.1) still holds with σ^2 replaced by $\hat{\sigma}^2$. Therefore, in addition to (2.2) and (2.4), we need only to verify (2.3) with σ^2 replaced by $\hat{\sigma}^2$. Obviously, it is enough to obtain

$$\sup_{h\in H_n} \frac{|\hat{\sigma}^2 - \sigma^2| n^{-1} \operatorname{tr} M_n(h)}{R_n(h)} \to 0,$$

which holds because $\operatorname{tr} M_n(h) = h \leq nR_n(h)\sigma^{-2}$. □

Shibata (1981) demonstrated the asymptotic optimality property of a related selection procedure, final prediction error (FPE) criterion, which selects $\hat{h}$ by minimizing $n^{-1}\|\mathbf{y}_n - \hat{\mu}_n(h)\|^2(n + 2h)$, under the normality assumption of e_1, condition (1.3) and the assumption that the rank of the largest model considered p_n is of order $o(n)$. 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. However, this equivalence crucially depends on the assumption $p_n = o(n)$; for example if $p_n = n$, then FPE always selects $h = p_n$, the full rank model, which is obviously an overfitting.

2.2. *Nearest-neighbor nonparametric regression—Example* 2 (*continued*). Condition (A.1) is satisfied if we have the following conditions on the weight

function:

there exists a positive number δ' such that

$$w_{n,h}(1) \leq 1 - \delta', \tag{2.7}$$

for any $n, h \geq 2$;

(2.8) for any n, h and i, $w_{n,h}(i) \geq w_{n,h}(i+1) \geq 0$;

$$\sum_{i=1}^{h} w_{n,h}(i) = 1. \tag{2.9}$$

The proof appeared in Li (1985), Lemma 4.1. Condition (A.3) will be satisfied if the optimal rate of convergence is slower than n^{-1}, i.e.,

$$\lim_{n\to\infty} \Big(\inf_{h\in H_n} R_n(h) \Big) n^{1-1/m} = \infty. \tag{A.3''}$$

We summarize the result by the following.

COROLLARY 2.3. *For the nearest-neighbor nonparametric regression problem, C_L is asymptotically optimal if* (A.2), (A.3'') *and* (2.7)–(2.9) *hold.*

REMARK 2.1. The constraints on the weight function (2.7)–(2.9) are only used to guarantee (A.1). As mentioned before, for $\lambda(M(h))$ larger than 1 the resulting estimate is inadmissible. Thus, in practice, we should check to see if $\lambda(M(h))$ is too large. This comment applies to any nonparametric regression estimate, including kernel estimates. The smoothing spline estimate always satisfies the condition $\lambda(M(h)) \leq 1$, an advantage over its rivals.

3. Generalized cross-validation. Most literature relates GCV to C_L via Taylor expansion. For h such that $n^{-1}\operatorname{tr} M_n(h)$ is small, $(1 - n^{-1}\operatorname{tr} M_n(h))^{-2}$ is approximately equal to $1 + 2n^{-1}\operatorname{tr} M_n(h)$. Thus, it has been claimed that minimizing (1.2) is asymptotically equivalent to minimizing

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

which was again claimed to be asymptotically equivalent to C_L because of the consistency assumption that $n^{-1}\|\mathbf{y}_n - \hat{\mu}_n(h)\|^2 \to \sigma^2$. These heuristics suffer from two basic logic difficulties:

(i) by assuming $n^{-1}\operatorname{tr} M_n(h) = o(1)$, we already impose some artificial condition on the selection class, violating the spirit of being data-driven;

(ii) by assuming consistency, we have placed too much confidence on a procedure whose asymptotic property is still to be investigated.

Finally, one can not justify the asymptotic equivalence by simply demonstrating that the difference between two minimization criteria is of order $o(1)$. As a matter of fact, there do exist examples, where C_L is consistent and asymptotically optimal while GCV is consistent but not asymptotically optimal [see Li, (1986)].

Li (1985) pointed out a direct way to relate GCV with C_L via the notion of nil-trace estimate. Consider the linear estimate

$$\bar{\mu}_n(h) = -\alpha \mathbf{y}_n + (1+\alpha)\hat{\mu}_n(h), \tag{3.1}$$

with

$$\alpha = n^{-1}\operatorname{tr} M_n(h)/\big(1 - n^{-1}\operatorname{tr} M_n(h)\big). \tag{3.2}$$

If we apply the C_L procedure to select an estimate from the class $\{\bar{\mu}_n(h): h \in H_n\}$, then it is easy to check that the resulting minimization procedure is exactly the same as GCV, (1.2). Note that the matrix associated with $\bar{\mu}_n(h)$,

$$\bar{M}_n(h) = -\alpha I + (1+\alpha)M_n(h), \tag{3.3}$$

has trace 0. Intuitively, for those h with small $n^{-1}\operatorname{tr} M_n(h)$ [as is the case for which $\hat{\mu}_n(h)$ is a good candidate, at least being consistent], the nil-trace estimates $\bar{\mu}_n(h)$ are close to the original estimates $\mu_n(h)$. On the other hand, for those h with large $n^{-1}\operatorname{tr} M_n(h)$ (poor candidate), the corresponding nil-trace estimates $\bar{\mu}(h)$ are expected to become even poorer because of the negative weight for $\mathbf{y}_n$ in the expression of (3.1). It is important to point out that we are not recommending the use of (3.1) to replace the original estimate as a final estimate. It is only a device aiming at some understanding about GCV. Efron (1986) commented that despite the name, GCV is nearly a member of C_L. Our nil-trace estimate approach supports this comment.

The following theorem provides conditions to justify the replacement of the original class by the nil-trace class.

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

$$L_n(\hat{h}_n) = O_p(1), \tag{3.4}$$

$$\lim_{n\to\infty} P\big\{\big|n^{-1}\operatorname{tr} A_n(\hat{h}_n)\big| > \delta\big\} = 1, \quad \textit{for some } \delta > 0 \tag{3.5}$$

and

$$\big(n^{-1}\operatorname{tr} M_n(\hat{h}_n)\big)^2 / n^{-1}\operatorname{tr} M_n(\hat{h}_n)M_n'(\hat{h}_n) \to 0, \quad \textit{in probability}, \tag{3.6}$$

we have

$$n^{-1}\big\|\bar{\mu}_n(\hat{h}_n) - \hat{\mu}_n(\hat{h}_n)\big\|^2 / R_n(\hat{h}_n) \to 0. \tag{3.7}$$

Conditions (3.5) and (3.6) can be interpreted as follows. First, observe that the ith diagonal element $m_i(h)$ of $M_n(h)$ is the weight on ith observation itself for estimating its own mean μ_i. Call these diagonal elements self-weights. Clearly the average variance of $\hat{\mu}_n(h)$, which equals $\sigma^2 n^{-1}\operatorname{tr} M_n(h)M_n'(h)$, comes partly from these weights. In fact, since $(n^{-1}\operatorname{tr} M_n(h))^2 \le n^{-1}\sum_{i=1}^n (m_i(h))^2$, we see that if the portion of variance contributed by self-weights accounts for only a small percentage of the average variance of $\hat{\mu}_n(h)$, then (3.6) holds. Similarly, (3.5) holds if self-weights are not too close to one, which is quite easy to achieve.

We are ready to apply C_L to the nil-trace class. Then, using Theorems 2.1 and 3.1, we establish the asymptotic optimality for GCV.

THEOREM 3.2. *Assume that* (A.1)–(A.3) *and the following conditions hold*:

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

for any sequence $\{h_n \in H_n\}$ *such that*

$$\text{(A.5)} \qquad n^{-1}\operatorname{tr} M_n(h_n)M_n'(h_n) \to 0,$$

we have $\left(n^{-1}\operatorname{tr} M_n(h_n)\right)^2 / n^{-1}\operatorname{tr} M_n(h_n)M_n'(h_n) \to 0$;

$$\text{(A.6)} \qquad \sup_{h \in H_n} n^{-1}\operatorname{tr} M_n(h) \le \gamma_1, \quad \textit{for some } 1 > \gamma_1 > 0;$$

$$\text{(A.7)} \qquad \sup_{h \in H_n} \left(n^{-1}\operatorname{tr} M_n(h)\right)^2 / n^{-1}\operatorname{tr} M_n(h)M_n'(h) \le \gamma_2, \quad \textit{for some } 1 > \gamma_2 > 0.$$

Then $\hat{h}_G$ *is asymptotically optimal.*

(A.4) assumes only the existence of consistent selection procedure when μ_n is known. The interpretation of conditions (A.5)–(A.7) is the same as that for (3.1)–(3.2).

The following two subsections demonstrate how to apply Theorem 3.2.

3.1., *Model selection.* Since $\operatorname{tr} M_n(h) = \operatorname{tr} M_n(h)M_n(h) = h$, (A.5) obviously holds. (A.6) is the same as (A.7), requiring that the largest model has rank $p_n \le n\gamma$ for some $0 < \gamma < 1$. But this constraint can be easily removed by the following argument.

First, let h^* be the minimizer of $\inf_{h \in H_n} L_n(h)$. Equation (2.4), 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 $H_n' = H_n \cap \{h\colon\ h \le n\gamma\}$, 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 (1985) proved that $\hat{\mu}(\hat{h}_G)$ is consistent [i.e., $L_n(\hat{h}_G) \to 0$] providing that the following addition condition on the random errors holds:

$$\text{(3.8)} \qquad \begin{aligned} &\text{there exists a constant } k' \text{ so that for any } a \ge 0, \\ &\sup_{x \in R} P\{x - a \le e_1 \le x + a\} \le k'a. \end{aligned}$$

Equation (3.8) is satisfied if e_1 has a bounded density. Using this consistency result and the previous arguments for h^*, 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 $p_n \le n\gamma$. The following corollary conveys the result we have established.

COROLLARY 3.1. *For the model selection problem of Example* 1, $\hat{h}_G$ *is asymptotically optimal if* (A.2) *with* $m = 2$, (A.3′), (A.4) *and* (3.8) *hold.*

REMARK 3.1. Breiman and Freedman (1983) show the asymptotic optimality of S_p under a different set of conditions. Their conditions exclude, for instance, the application in selecting polynomial regression models.

3.2. *Nearest-neighbor nonparametric regression.* Observe that

$$n^{-1}\operatorname{tr} M_n(h) = w_{n,h}(1)$$

and that under (2.7),

$$n^{-1}\operatorname{tr} M_n(h)M_n'(h) = \sum_{i=1}^{h} w_{n,h}(i)^2 \geq w_{n,h}(1)^2 + h^{-1}(1 - w_{n,h}(1))^2.$$

Thus, it is clear that the following condition implies (A.5):

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

This condition was used in Li (1984) 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 $\|\mathbf{y}_n - \hat{\mu}_n(1)\|^2 = 0$ and $1 - n^{-1}\operatorname{tr} M_n(1) = 0$]. It reduces to condition (2.7).

Finally it is obvious that (3.9) and (2.7) imply (A.7). Therefore we obtain the following desired result.

COROLLARY 3.2. *Suppose that the weight functions satisfy the regularity conditions of* (2.5)–(2.7) *and* (3.9). *Then* $\hat{h}_G$ *is asymptotically optimal if* (A.2), (A.3″) *and* (A.4) *hold and* $H_n = \{2, \ldots, n\}$.

REMARK 3.2. As argued in Section 2.2, condition (2.8) may not be necessary providing that (A.1) is satisfied.

4. Stein estimates. Intuitively, the replacement of $\hat{\mu}_n(h)$ by $\bar{\mu}_n(h)$ does not seem appropriate if $n^{-1}\operatorname{tr} M_n(h)$ is not negligible because the weight on $\mathbf{y}_n$ is always negative. This is a weak point for considering $\bar{\mu}_n(h)$. A better way of replacement is by means of Stein estimates, defined by

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

which has an (approximately) unbiased risk estimate

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

The original version of these quantities, given in Stein (1981), was a little complicated. Note that $\tilde{\mu}_n(h)$ is also a linear combination of $\mathbf{y}_n$ and $\hat{\mu}_n(h)$. But unlike nil-trace estimate $\bar{\mu}_n(h)$, the weights are now data-dependent. They can always be made positive by a slight modification as was suggested in Stein

(1981). But for our purpose, this modification is not necessary since we are studying asymptotics and the weights will be positive with probability nearly one for large sample size. Stein estimates possess the nice property that as estimates of μ_n they dominate the raw data $\mathbf{y}_n$ for normal errors under some mild assumption about the largest characteristic root of $A_n(h)$. Li and Hwang (1984) studied the asymptotic behavior of $\tilde{\mu}_n(h)$ for nonparametric regression problems. Basically, $\tilde{\mu}_n(h)$ and $\hat{\mu}_n(h)$ will be very close to each other providing that $\hat{\mu}_n(h)$ is close to the true value μ_n. Hence using Stein estimates, we do not lose 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 for Stein estimates may not be as big as the linear ones which usually have unbounded risks. This justifies the replacement of $\hat{\mu}_n(h)$ by $\tilde{\mu}_n(h)$. Now it is easy to see the interesting consequence that the natural way of selecting $\tilde{\mu}_n(h)$, minimizing $\mathrm{SURE}_n(h)$, is exactly the same as selecting $\hat{\mu}_n(h)$ by GCV. Li (1985) argued that $\mathrm{SURE}_n(h)$, initially proposed as an estimate of the risk of the 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 $En^{-1}\|\mu_n - \tilde{\mu}_n(h)\|^2$]. In addition, the consistency is uniform in $\mu_n \in R^n$. The consistency of the Stein estimate selected by GCV, $\tilde{\mu}_n(\hat{h}_G)$, was also established there. The following theorem strengthens this result by proving the asymptotic efficiency of $\tilde{\mu}_n(\hat{h}_G)$.

THEOREM 4.1. *Under the assumptions of Theorem* 3.2, *we have*

$$n^{-1}\left\|\tilde{\mu}_n(\hat{h}_G) - \mu_n\right\|^2 / \inf_{h\in H_n} L_n(h) \to 1 \tag{4.1}$$

and

$$\inf_{h\in H_n} n^{-1}\left\|\tilde{\mu}_n(h) - \mu_n\right\|^2 / \inf_{h\in H_n} L_n(h) \to 1, \tag{4.2}$$

in probability.

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

5. Cross-validation. Let $\mu_n^c(h)$ denote the delete-one estimate of μ_n, $\tilde{M}_n(h)\mathbf{y}_n$. The exact form of $\tilde{M}_n(h)$ can only be written down case by case. However, a useful common feature is that all $\tilde{M}_n(h)$ have diagonal elements identical to zero. From this, we see that cross-validation is just the C_L procedure applied to the class of delete-one estimates $\{\mu_n^c(h)\colon h \in H_n\}$. The replacement of the original estimate $\hat{\mu}_n(h)$ by $\mu_n^c(h)$ can be justified intuitively by arguing that for large n the delete-one estimate, which is based on a sample of size $n-1$, should be nearly the same as the original estimate. The following theorem makes this type of argument more vigorous. Define

$$\tilde{L}_n(h) = n^{-1}\left\|\mu_n - \mu_n^c(h)\right\|^2,$$
$$\tilde{R}_n(h) = E\tilde{L}_n(h).$$

THEOREM 5.1. *Assume that* (A.1)–(A.4) *and the following conditions hold*:

$$\limsup_{n\to\infty} \sup_{h\in H_n} \lambda\big(\tilde{M}_n(h)\big) < \infty, \tag{A.8}$$

$$\sum_{h\in H_n} \big(n\tilde{R}_n(h)\big)^{-m} \to 0, \tag{A.9}$$

for any sequence $\{h_n\}$, $h_n \in H_n$, *we have*

$$\tilde{R}_n(h_n)/R_n(h_n) \to 1 \tag{A.10}$$

if either $R_n(h_n) \to 0$ *or* $\tilde{R}_n(h_n) \to 0$.

Then $\hat{h}_c$ *is asymptotically optimal.*

5.1. *Model selection.* For the model selection problem, we have discussed the conditions (A.1)–(A.3) in Section 2.1. Briefly speaking, (A.1) holds automatically and (A.3′) implies (A.3) with $m = 2$. In Section 6, we shall show that (A.8)–(A.10) hold under the following additional assumptions ($\bar{\lambda}(\cdot)$ denotes the maximum diagonal element of a matrix):

$$\limsup_{n\to\infty} \sup_{h\in H_n} \bar{\lambda}(M_n(h)) < 1; \tag{5.1}$$

there exists some positive constant Λ such that for any n, h, $\bar{\lambda}(M_n(h)) \le \Lambda h n^{-1}$. (5.2)

Condition (5.1) is a weak assumption, assuming only that the self-weights are bounded away from 1. Note that one property of a projection matrix is that $\bar{\lambda}(M_n(h)) \le 1$. If the ith diagonal element is close to 1, then all other elements in the ith row have to be close to zero, meaning that $\hat{\mu}_n(h)$ estimates μ_i almost only by y_i itself. Hence for such cases the delete-one estimate (which does not use y_i at all in estimating μ_i) clearly is not close to the initial estimate $\hat{\mu}_n(h)$ [for the extreme case that $\bar{\lambda}(M_n(h)) = 1$, the delete-one estimate is undefined].

Condition (5.2) excludes extremely unbalanced designs. To see this, we need the following view of the self-weights: The ith diagonal element of $M_n(h)$ times σ^2 equals the variance of the least-squares estimate (based on model h) of μ_i [this is simply because of the property of projection matrix $M_n(h)M_n(h) = M_n(h)$]. Thus if $\bar{\lambda}(M_n(h))\sigma^2$ is too large compared with the average variance $hn^{-1}\sigma^2$, then some μ_i's are estimated much less accurately than others, an indication of severe unbalancedness. Severe unbalancedness may also incur nonrobustness [see, for example, Box and Draper (1975) and Belsley, Kuh and Welsch (1980)].

We conclude this section by summarizing the result for model selection as follows.

THEOREM 5.2. *For the model selection problem of Example* 1, *assume that* (A.2) *with* $m = 2$, (A.3′), (A.4), (5.1) *and* (5.2) *hold. Then the procedure of cross-validation is asymptotically optimal.*

5.2. *Nearest-neighbor nonparametric regression.* Using Theorem 5.1, we may establish the asymptotic optimality of cross-validation in nearest-neighbor nonparametric regression.

THEOREM 5.3. *Under the assumptions of Corollary* 3.2, $\hat{h}_c$ *is asymptotically optimal.*

REMARK 5.1. Another way of defining the delete-one estimate can be given by letting $\hat{y}_{-i} = \sum_{j=1}^{h} w_{n-1,h}(j) y_{i(j+1)}$. If the weight function depends only on h but not on n, then we have the same delete-one estimate as before. In general, it seems that we need some conditions on the relationship between $w_{n-1,h}(j)$ and $w_{n,h}(j)$ to guarantee that our asymptotic setting provides a reasonable embedding of weight functions. For instance we may want to assume that $\sup_{1 \le j \le h} |w_{n-1,h}(j)/w_{n,h}(j) - 1| \to 0$ if $h \to \infty$.

6. Proofs.

PROOF OF THEOREM 2.1. We shall prove (2.2) first. Given any $\delta > 0$, by Chebyshev's inequality we have

$$P\left\{ \sup_{h \in H_n} n^{-1} \left| \langle \mathbf{e}_n, A_n(h)\mu_n \rangle \right| / R_n(h) > \delta \right\} \le \sum_{h \in H_n} \frac{n^{-2m} E\left[\langle e_n, A_n(h)\mu_n \rangle^{2m} \right]}{\delta^{2m} R_n(h)^{2m}},$$

which, by Theorem 2 of Whittle (1960), is no greater than

$$C\delta^{-2m} \sum_{h \in H_n} n^{-2m} \| A_n(h)\mu_n \|^{2m} R_n(h)^{-2m},$$

for some constant $C > 0$. Now since $n^{-1}\|A_n(h)\mu_n\|^2 \le R_n(h)$, the last expression does not exceed $C\delta^{-2m}\sum_{h \in H_n}(nR_n(h))^{-m}$, which tends to 0 by (A.3). Thus (2.2) is proved. Equation (2.3) can be established in a similar manner, by noting that, as an application of Theorem 2 of Whittle (1960),

$$E\left(\sigma^2 \operatorname{tr} M_n(h) - \langle \mathbf{e}_n, M_n(h)\mathbf{e}_n \rangle\right)^{2m} \le C'(\operatorname{tr} M_n(h) M_n'(h))^m,$$

for some $C' > 0$ and that $\sigma^2 n^{-1} \operatorname{tr} M_n(h) M_n'(h) \le R_n(h)$. Finally, it is clear that (2.4) will follow from the following two statements:

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

and

$$\sup_{h \in H_n} n^{-1} \left| \|M_n(h)\mathbf{e}_n\|^2 - \sigma^2 \operatorname{tr} M_n(h) M_n'(h) \right| / R_n(h) \to 0. \tag{6.2}$$

Since $\langle A_n(h)\mu_n, M_n(h)\mathbf{e}_n \rangle = \langle M_n'(h)A_n(h)\mu_n, \mathbf{e}_n \rangle$ and $\|M_n'(h)A_n(h)\mu_n\|^2 \le \lambda(M_n(h))^2 \|A_n(h)\mu_n\|^2$, the proof of (6.1) will be the same as that of (2.2) in view of (A.1). Similarly, write $\|M_n(h)\mathbf{e}_n\|^2 = \langle M_n'(h)M_n(h)\mathbf{e}_n, \mathbf{e}_n \rangle$ and observe that $\operatorname{tr}(M_n'(h)M_n(h))^2 \le \lambda(M_n(h))^2 \operatorname{tr} M_n'(h)M_n(h)$. We see that (6.2) can be proved exactly as (2.3). This completes the proof of Theorem 2.1. □

PROOF OF THEOREM 3.1. Observe that $\bar{\mu}_n(\hat{h}_n) - \hat{\mu}_n(\hat{h}_n) = \hat{\alpha}(\hat{\mu}_n(\hat{h}_n) - \mathbf{y}_n)$, where $\hat{\alpha}$ equals (3.2) with h replaced by $\hat{h}_n$. By (3.5), (3.6) and the fact that

$R_n(h) \geq \sigma^2 \operatorname{tr} M_n(h)M_n'(h)$, we see that $\hat{\sigma}^2/R_n(\hat{h}_n)$ converges to 0 in probability. Finally, since $\|\hat{\mu}_n(\hat{h}_n) - \mathbf{y}_n\|^2 \leq 2\|\hat{\mu}_n(\hat{h}_n) - \mu_n\|^2 + 2\|\mathbf{e}_n\|^2$, the desired result follows from (3.4). □

PROOF OF THEOREM 3.2. First we shall show how to apply Theorem 2.1 to the class of nil-trace estimators to obtain the following crucial result (6.5). Set $\overline{L}_n(h) = n^{-1}\|\bar{\mu}_n(h) - \mu_n\|^2$ and $\overline{R}_n(h) = E\overline{L}_n(h)$. A simple computation leads to

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

and

$$n\overline{R}_n(h) = \frac{\|A_n(h)\mu_n\|^2 + \sigma^2 \operatorname{tr} M_n(h)M_n'(h) - \sigma^2 n^{-1}(\operatorname{tr} M_n(h))^2}{(n^{-1}\operatorname{tr} A_n(h))^2}.$$

Now, by (A.6) and (A.7), it is easy to see that there exist positive constants c_1, c_2 such that for any n, $h \in H_n$, we have

$$c_1 \leq \operatorname{tr} \overline{M}_n(h)\overline{M}_n'(h)/\operatorname{tr} M_n(h)M_n'(h) \leq c_2, \tag{6.3}$$

$$c_1 \leq \overline{R}_n(h)/R_n(h) \leq c_2. \tag{6.4}$$

From this, it follows that (A.1) and (A.3) also hold with $R_n(h)$ and $M_n(h)$ replaced by $\overline{R}_n(h)$ and $\overline{M}_n(h)$, respectively. Hence, by Theorem 2.1, we have

$$\overline{L}_n(\hat{h}_G)/\inf_{h \in H_n} \overline{L}_n(h) \to 1. \tag{6.5}$$

In fact, the following analogue of (2.4) also holds:

$$\sup_{h \in H_n} |\overline{L}_n(h)/\overline{R}_n(h) - 1| \to 0. \tag{2.4'}$$

Next let h_* be the minimizer of $L_n(h)$ over $h \in H_n$. From (6.5), it is clear that (1.6) will hold for $\hat{h} = \hat{h}_G$ if we can verify that

$$\overline{L}_n(h_*)/L_n(h_*) \to 1 \tag{6.6}$$

and

$$\overline{L}_n(\hat{h}_G)/L_n(\hat{h}_G) \to 1. \tag{6.7}$$

Theorem 3.1 can be used to prove (6.6) and (6.7). To see this, first observe that (3.5) always holds because of (A.6). Next, from (2.4') and (A.4), it follows that $n^{-1}\operatorname{tr} M_n(h_*)M_n(h_*)' \to 0$. Hence, by (A.5), (3.6) holds for $\hat{h}_n = h_*$. Finally, (3.4) with $\hat{h}_n = h_*$ is weaker than (A.4). Therefore, Theorem 3.1 applies for $\hat{h}_n = h_*$, and (6.6) follows as a simple consequence of (3.7) and (2.4). Turning to the proof of (6.7), we need only show that

$$n^{-1}\operatorname{tr} M_n(\hat{h}_G)M_n'(\hat{h}_G) \to 0, \tag{6.8}$$

and (3.4) holds for $\hat{h}_n = \hat{h}_G$. Now by (6.6) and (A.4) we see that $\overline{L}_n(h_*) \to 0$,

which implies

$$\bar{L}_n(\hat{h}_G) \to 0, \tag{6.9}$$

because of (6.5). Then from (2.4′) we conclude that $\bar{R}_n(\hat{h}_G) \to 0$, which in turn implies $n^{-1}\operatorname{tr}\bar{M}_n(\hat{h}_G)\bar{M}_n'(\hat{h}_G) \to 0$. In view of (6.3), we have established (6.8). Finally, (3.4) follows from (6.9), (2.4′), (6.4) and (2.4). This completes the proof of Theorem 3.2. □

PROOF OF THEOREM 4.1. To prove (4.1), by Theorem 3.2 it suffices to show that

$$n^{-1}\left\|\tilde{\mu}_n(\hat{h}_G) - \hat{\mu}_n(\hat{h}_G)\right\|^2 / L_n(\hat{h}_G) \to 0. \tag{6.10}$$

First, observe that

$$\begin{aligned} n^{-1}\left\|\tilde{\mu}_n(\hat{h}_G) - \tilde{\mu}_n(\hat{h}_G)\right\|^2 &= \left(\frac{n^{-1}\sigma^2\operatorname{tr}A_n(\hat{h}_G)}{n^{-1}\left\|A_n(\hat{h}_G)\mathbf{y}_n\right\|^2} - 1\right)^2 n^{-1}\left\|A_n(\hat{h}_G)\mathbf{y}_n\right\|^2 \\ &= \Big[\left(\sigma^2 - n^{-1}\|\mathbf{e}_n\|^2\right) - L_n(\hat{h}_G) \\ &\quad - 2n^{-1}\left\langle \mathbf{e}_n, \mu_n - \hat{\mu}_n(\hat{h}_G)\right\rangle \\ &\quad - n^{-1}\sigma^2\operatorname{tr}M_n(\hat{h}_G)\Big]^2 / n^{-1}\left\|A_n(\hat{h}_G)\mathbf{y}_n\right\|^2, \end{aligned}$$

and that

$$n^{-1}\left\|A_n(\hat{h}_G)\mathbf{y}_n\right\|^2 = n^{-1}\left\|\mathbf{e}_n + \left(\mu_n - \hat{\mu}_n(\hat{h}_G)\right)\right\|^2 \to \sigma^2,$$

because of the consistency of $\hat{\mu}_n(\hat{h}_G)$. Therefore, to prove (6.10) it is enough to verify

$$\left(\sigma^2 - n^{-1}\|\mathbf{e}_n\|^2\right)^2 / L_n(\hat{h}_G) \to 0, \tag{6.11}$$

$$\left(n^{-1}\left\langle \mathbf{e}_n, \mu_n - \hat{\mu}_n(\hat{h}_G)\right\rangle\right)^2 / L_n(\hat{h}_G) \to 0 \tag{6.12}$$

and

$$\left(n^{-1}\operatorname{tr}M_n(\hat{h}_G)\right)^2 / L_n(\hat{h}_G) \to 0. \tag{6.13}$$

Now (A.3′), which is weaker than (A.3), and (2.4) imply that $nL_n(\hat{h}_G) \to \infty$. Thus, (6.11) follows from the central limit theorem. Next, as was proved in the proof of Theorem 3.2, (3.6) holds for $\hat{h}_n = \hat{h}_G$. This together with (2.4) implies (6.13). Finally to prove (6.12), it suffices to show

$$\left(n^{-1}\left\langle \mathbf{e}_n, A_n(\hat{h}_G)\mu_n\right\rangle\right)^2 / R_n(\hat{h}_G) \to 0 \tag{6.14}$$

and

$$\left(n^{-1}\left\langle \mathbf{e}_n, M_n(\hat{h}_G)\mathbf{e}_n\right\rangle\right)^2 / R_n(\hat{h}_G) \to 0. \tag{6.15}$$

It is not difficult to see that (6.14) follows from (2.2), and that (6.15) follows from (2.3), (6.13) and (2.4). This completes the proof of (4.1). The proof of (4.2) is similar; namely, to establish (6.10) with $\hat{h}_G$ replaced by the minimizer of $\min_{h\in H_n} n^{-1}\|\tilde{\mu}_n(h) - \mu_n\|^2$, which can be carried out as before. □

PROOF OF THEOREM 5.1. Using Theorem 2.1, we have

$$\tilde{L}_n(\hat{h}_c)/\inf_{h\in H_n}\tilde{L}_n(h) \to 1.$$

In addition, the analogue of (2.4) implies

$$\tilde{L}_n(\hat{h}_c)/\tilde{R}_n(\hat{h}_c) \to 1$$

and

$$\tilde{L}_n(h_*)/\tilde{R}_n(h_*) \to 1,$$

when h_* is the minimizer of $\inf_{h\in H_n} L_n(h)$. In the proof of Theorem 3.2, we have shown that $R_n(h_*) \to 0$, which, by (A.10), implies that $\tilde{R}_n(h_*)/R_n(h_*) \to 1$. Thus, $\tilde{L}_n(h_*)/L_n(h_*) \to 1$. Since $\tilde{L}_n(\hat{h}_c) \le \tilde{L}_n(h_*)$, it suffices to show that $\tilde{L}_n(\hat{h}_c)/L_n(\hat{h}_c) \to 1$. Now since $\tilde{L}_n(h_*) \to 0$, we have $\tilde{L}_n(\hat{h}_c) \to 0$, implying that $\tilde{R}_n(\hat{h}_c) \to 0$, which, due to (A.10), implies $\tilde{R}_n(\hat{h}_c)/R_n(\hat{h}_c) \to 1$. Therefore, we have $\tilde{L}_n(\hat{h}_c)/L_n(\hat{h}_c) \to 1$ as desired. □

PROOF OF THEOREM 5.2. It suffices to establish (A.8)–(A.10). First, using the simple inequalities that $\lambda(AB) \le \lambda(A)\lambda(B)$ and $\lambda(A + B) \le \lambda(A) + \lambda(B)$ for any $n \times n$ matrices A, B, we see from (1.4) that (A.8) follows from (A.1) and (5.1).

Next we shall prove (A.10). Observe that from (2.5) we have $R_n(h_n) \ge \sigma^2 h_n n^{-1}$ and $\tilde{R}_n(h_n) \ge \sigma^2 h_n n^{-1}$ [note that the Gauss–Markov theorem implies the variance part in $R_n(h_n)$ is no larger than the variance part in $\tilde{R}_n(h_n)$]. Thus, either $R_n(h_n) \to 0$ or $\tilde{R}_n(h_n) \to 0$ implies that $h_n n^{-1} \to 0$, which in turn yields $\bar{\lambda}(M_n(h_n)) \to 0$, due to (5.2). Thus, $\bar{\lambda}(D_n(h_n)) = 1 + o(1) = \underline{\lambda}(D_n(h_n))$, where $\underline{\lambda}(\cdot)$ denotes the minimum diagonal element of a matrix. Therefore,

$$\begin{aligned} n\tilde{R}(h_n) &= \left\|\left(\tilde{M}_n(h_n) - I\right)\mu_n\right\|^2 + \sigma^2 \operatorname{tr} \tilde{M}_n(h_n)\tilde{M}_n'(h_n) \\ &= (1 + o(1))\left\|\left(M_n(h) - I\right)\mu_n\right\|^2 + \sigma^2(1 + o(1))\operatorname{tr} M_n(h)M_n'(h) \\ &= (1 + o(1))nR(h_n), \end{aligned}$$

proving (A.10). Here, the second equality is based on (1.4).

To prove (A.9), we first let $\tilde{h}_n$ be the minimizer of $\min_{h\in H_n}\tilde{R}_n(h)$. Since (2.6) still holds with $R_n(h)$ replaced by $\tilde{R}_n(h)$, it suffices to show that $n\tilde{R}_n(\tilde{h}_n) \to \infty$. Suppose that this is not true. Then since $n\tilde{R}_n(\tilde{h}_n) \ge \sigma^2\tilde{h}_n$, $\tilde{h}_n$ is bounded, implying $n^{-1}\tilde{h}_n \to 0$, again. Therefore, we can show that $n\tilde{R}(h_n) = (1 + o(1))nR(h_n)$. Since $R(\tilde{h}_n) \ge \inf_{h\in H_n} R_n(h)$, by (A.4) and (2.4) we see that $nR_n(\tilde{h}_n) \to \infty$, which implies $n\tilde{R}_n(\tilde{h}_n) \to \infty$. Hence, in any case, we have shown (A.9) holds. This completes the proof of Theorem 5.2. □

PROOF OF THEOREM 5.3. Again we only need to establish (A.8)–(A.10).

To prove (A.8), first let $T_n(h) = \tilde{M}_n(h) + w_{n,h}(1)I$ and observe that $\lambda(\tilde{M}_n(h)) \le \lambda(T_n(h)) + w_{n,h}(1)$. Now apply Lemma 4.1 of Li (1985) by treating $T_n(h)$ as $M_n(h)$ [since $T_n(h)$ corresponds to a nearest-neighbor estimate with a different weight function that satisfies (4.6) and (4.7)]. We see that $\lambda(T_n(h))$ is bounded, which proves (A.8).

Next, we claim that (A.10) implies (A.9). To see this, first observe that (A.9) follows from (A.3″) with $R_n(h)$ replaced by $\tilde{R}_n(h)$. Denote the minimizers of $\min_{h \in H_n} \tilde{R}_n(h)$ and $\min_{h \in H_n} R_n(h)$ by $\tilde{h}_*$ and h_*, respectively. Since $R_n(\tilde{h}_*) \ge R_n(h_*)$, it suffices to show that $\tilde{R}_n(\tilde{h}_*) = (1 + o(1))R_n(\tilde{h}_*)$. Hence, by (A.10), we need only to verify that $\tilde{R}_n(\tilde{h}_*) \to 0$. Now, since $R_n(h_*) \to 0$, by (A.10) again, we see that $\tilde{R}_n(h_*) \to 0$, which implies the desired result $\tilde{R}_n(\tilde{h}_*) \to 0$ because $\tilde{R}_n(\tilde{h}_*) \le \tilde{R}_n(h_*)$. Thus, (A.9) will hold if (A.10) is satisfied.

It remains to verify (A.10). Let $f_\infty = \sup_{\mathbf{x} \in X} f(\mathbf{x})$. From the definition of $\tilde{M}_n(h)\mathbf{y}_n$, we see that

$$
\begin{aligned}
& n^{-1}\left\|\left(M_n(h) - \tilde{M}_n(h)\right)\mu_n\right\|^2 \\
& \quad = (1/n)\sum_{i=1}^{n}\Bigg(\sum_{j=2}^{h}\left(w_{n,h}(j) - w_{n,h}(j-1)\right)f(\mathbf{x}_{i(j)}) \\
& \qquad\qquad + w_{n,h}(1)f(\mathbf{x}_{i(1)}) - w_{n,h}(h)f(\mathbf{x}_{i(h+1)})\Bigg)^2 \\
& \quad \le \left(\sum_{j=2}^{h}\left(w_{n,h}(j-1) - w_{n,h}(j)\right) + w_{n,h}(1) + w_{n,h}(h)\right)^2 f_\infty^2 \\
& \quad \le 4w_{n,h}(1)^2 f_\infty^2,
\end{aligned}
$$

where the first inequality follows from (2.8). Now compare

$$
\tilde{R}_n(h) = n^{-1}\left\|\mu_n - \tilde{M}_n(h)\mu_n\right\|^2 + \sigma^2\sum_{i=1}^{h} w_{n,h}(i)^2
$$

with

$$
R_n(h) = n^{-1}\left\|\mu_n - M_n(h)\mu_n\right\|^2 + \sigma^2\sum_{i=1}^{h} w_{n,h}(i)^2.
$$

We see that the desired result $\tilde{R}_n(h)/R_n(h) \to 1$ will hold if

$$
w_{n,h}(1)^2 \Big/ \sum_{i=1}^{h} w_{n,h}(i)^2 \to 0. \tag{6.16}
$$

Since $\sum_{i=1}^{h} w_{n,h}(i)^2 \ge h^{-1}$, by (3.8) we see that (6.16) holds if $h \to \infty$. Finally, our assumption that either $\tilde{R}_n(h) \to 0$ or $R_n(h) \to 0$ implies that $h^{-1} \to 0$, as desired. The proof of this theorem is now complete. □

REFERENCES

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

BELSLEY, D. A., KUH, E. and WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

BOX, G. and DRAPER, N. (1975). Robust designs. *Biometrika* **62** 347–352.

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

COHEN, A. (1966). All admissible linear estimates of the mean vector. *Ann. Math. Statist.* **37** 458–463.

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.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461–470.

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

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

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

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

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

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

THOMPSON, M. (1978). Selection of variables in multiple regression. *Internat. Statist. Rev.* **46** 1–20, 129–146.

WAHBA, G. and WOLD. S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Comm. Statist.* **4** 1–17.

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, CALIFORNIA 90024