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

Short title: Generalized Cross-Validation

by

Ker-Chau Li
Department of Statistics
Purdue University

Technical Report #83-34

This paper is dedicated to the memory of Professor Jack Kiefer - advisor, teacher, and above all, friend.

August 1983

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

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

Keywords and Phrases: C_L; C_p; cross-validation; generalized cross-validation; model selection; nearest neighbor estimates; nil-trace linear estimates; ridge regression; smoothing splines; Stein estimates; Stein's unbiased risk estimates.

SUMMARY

This paper concerns the method of generalized cross-validation (G.C.V.), a promising way of choosing between linear estimates. Two new interpretations of G.C.V. will be given. One relates Mallows' C_p and C_L statistics while the other is based on Stein estimates and the associated unbiased risk estimates (Stein 1981). The latter approach turns out more useful. A number of consistency results are thereby 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). Valid confidence sets can be constructed as an immediate consequence of this uniform consistency at least for large sample sizes. We also discuss the case of unknown sampling error where three options are given with their consistency properties examined. Finally, we propose a variant of G.C.V. 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 among those being tentatively proposed. Quite often after some preliminary inspection on the given estimation problem, a statistician may suggest a certain class of estimators. Different members in the class may look reasonable under different plaussible conditions. Instead of enforcing any subjectivity to select one of them, it is often desirable to let the data speak for themselves. The generalized cross-validation (G.C.V.) of Craven and Wahba (1979) is one of many such promising data-driven techniques of selection. While the extension to the choice among non-linear estimators has been under way (Wahba 1982), we shall nevertheless focus our study on the linear ones in this paper.

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

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

and assume that ε_i's have mean 0 and common variances σ^2. To estimate $\underset{\sim}{\mu}_n = (\mu_1, \ldots, \mu_n)'$, a class of linear estimators $\hat{\underset{\sim}{\mu}}_n(h)$, indexed by h, is proposed. Let H_n be the index set and $M_n(h)$ be the n x n matrix associated with $\hat{\underset{\sim}{\mu}}_n(h)$ such that

$$\hat{\underset{\sim}{\mu}}_n(h) = M_n(h)\underset{\sim}{y}_n \tag{1.2}$$

where $\underset{\sim}{y}_n = (y_1, \ldots, y_n)'$. G.C.V. chooses h by minimizing the quantity

$$(1.3) \qquad GCV_n(h) = \frac{n^{-1}||y_n - \hat{\mu}_n(h)||^2}{(1 - n^{-1}\text{tr } M_n(h))^2}.$$

Here $||\cdot||$ denotes the Euclidean norm of R^n and tr $M_n(h)$ is the trace of $M_n(h)$.

For illustrations, consider the following examples.

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 $0 \le x_1 < x_2 < \ldots x_n < 1$. Due to the coninuity of f, it is reasonable to estimate μ_i by $(2h+1)^{-1} \sum_{j=-h}^{h} y_{i+j}$, for some $h \le (n-1)/2$. Here we identify y_{-j} with y_{n-j}. It is clear that the rows of $M_n(h)$ in this case are certain permutations of the row vector $(2h+1)^{-1}(1,1,\ldots,1,0,0,\ldots,0)$ with 2h+1 ones. (1.3) is reduced to

$$GCV_n(h) = (2h+1)^2(2h)^{-2}n^{-1}||y_n - \hat{\mu}_n(h)||^2$$

Example 2. Model selection.

Suppose associated with y_i there are p_n explanatory variables $x_{i1}, x_{i2}, \ldots, x_{ip_n}$, arranged in the decreasing order of importance. To estimate μ_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 β_j being unknown parameters, and then use the least squares estimator

$$(1.4) \qquad \hat{\mu}(h) = X_h(X_h'X_h)^{-1}X_h'y_n \quad ,$$

where X_h is the n x h design matrix. Now $M_n(h) = X_h(X_h'X_h)^{-1}X_h$ is a pro-

jection matrix with rank h. (1.3) becomes

$$GCV_n(h) = n(n-h)^{-2}||y_n - \hat{\mu}_n(h)||^2. \tag{1.5}$$

In the above examples H_n are discrete. For continuous H_n, see the following

Example 3. Ridge regression.

Assume the regression model

$$\mu_i = \sum_{j=1}^{p_n} x_{ij}\beta_j \quad , \quad i = 1,2,\ldots,n, \tag{1.6}$$

with the n x p_n design matrix $X = (x_{ij})$. When the information matrix X'X is nearly degenerated, the ridge regression method is often advocated. Denoting the p_n x p_n identity matrix by I_{p_n}, the ridge regression estimate of $\beta = (\beta_1,\ldots,\beta_{p_n})'$ is $(X'X + hI_{p_n})^{-1}X'y_n$ and μ_n is estimated by

$$\hat{\mu}_n(h) = X(X'X + hI_{p_n})^{-1}X'y_n. \tag{1.7}$$

Here the ridge parameter h is a non-negative number to be chosen. The trace of $M_n(h) = X(X'X + hI_{p_n})^{-1}X'$ is often obtained by the singular value decomposition (Golub and Reinsch 1870). In particular, let X = UDV with U and V as orthogonal matrices with ranks n and p_n respectively, and D being an n x p_n matrix having $\lambda_{1,n}^{1/2} \geq \lambda_{2,n}^{1/2} \geq \ldots \geq \lambda_{p_n,n}^{1/2} \geq 0$ for the i,i^{th} entries, $i = 1,2,\ldots, \min\{n,p_n\}$, and 0 elsewhere. A straight-forward manipulation yields

$$\text{tr } M_n(h) = \sum_{i=1}^{p_n} \lambda_{i,n}(h + \lambda_{i,n})^{-1} \tag{1.8}$$

and

$$(1.9) \qquad ||y_n - \hat{\mu}_n(h)||^2 = \sum_{i=1}^{n} h^2(h + \lambda_{i,n})^{-2}\overline{y}_i^2$$

where $\overline{y}_n = (\overline{y}_1,\ldots,\overline{y}_n)' = U'y_n$. (1.3) becomes

$$(1.10) \qquad GCV_n(h) = [n^{-1}\sum_{i=1}^{n}(h + \lambda_{i,n})^{-2}\overline{y}_i^2\ ] \ /\ (n^{-1}\sum_{i=1}^{n}(h + \lambda_{i,n})^{-1})^2$$

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

Example 4. Smoothing splines.

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

$$(1.11) \qquad \underset{f \in W_2^k[0,1]}{\text{Min}} \quad n^{-1}\sum_{i=1}^{n}(y_i - f(x_i))^2 + h\int_0^1 f^{(k)}(x)^2dx.$$

Here the smoothing parameter h is a non-negative 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 $\hat{\mu}_n = (\hat{f}_h(x_1),\ldots,\hat{f}_h(x_n)) = M_n(h)y_n$ has been studied extensively (Reinsch 1967, Demmler and Reinsch 1975, Wahba 1975, 1978, Craven and Wahba 1979, Speckman 1981 a,b, 1982 etc.). To implement G.C.V., one may either employ the fast algorithm of Utreras (1979 , 1980) or carry out a singular value decomposition (Craven and Wahba, 1979).

G.C.V. was firstly proposed to choose the smoothing parameter for a smoothing spline (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). In these papers, G.C.V. was viewed as a rotation-invariant version of the (ordinary) cross-validation (C.V.) of Stone (1974) and Geisser (1975), namely Allen's PRESS (Allen 1974). G.C.V. is just the C.V. after the data $\underset{\sim}{y}_n$ being transformed suitably so that the value in each coordinate when being excluded from the data set may be easier to predict from others. This transformation is invariant in certain sense which was argued to be desirable. However, since it involves circulant matrices, further studies on the probability structure of the transformed complex-valued data set seem to be necessary in order to get a better understanding of G.C.V.. Alternatively, in the cases of the ridge regression and smoothing splines, G.C.V. may be written as weighted version of C.V.: a suitably weighted sum of prediction squared errors. But these weights may as well appear arbitrary to some people. The two G.C.V. theorems in Golub, Heath and Wahba, seem to be most useful and persuading: the first one compares the expected value of (1.3) with the mean squared error for the linear estimator $\hat{\underset{\sim}{\mu}}_n(h)$; the second one justifies G.C.V. from a Bayesian viewpoint in the content of ridge regression. In fact, certain asymptotic optimality of G.C.V. in choosing smoothing splines had been obtained through the use of the first G.C.V. theorem or its strengthened version, intertwining with the rather complicated eigenvalue theory associated with the smoothing splines (Craven and Wahba 1979, Speckman 1982). However, these results required some artificial restrictions on the range of h. Hence,

this leaves something to be desired. At least we would like to have a consistency result without putting any conditions on h.

The goals of this paper are multifold: (i) seeking natural ways to understand G.C.V.; (ii) establishing general consistency results; (ii) assessing the performance of the G-cross-validated estimate; (iv) constructing valid confidence sets.

After a brief review of some relevent data-driven techniques, Section 2 provides two new viewpoints of G.C.V.: (i) G.C.V. is just the procedure of Mallows' C_L (Mallows 1973) applied to a class of suitably-constructed nil-trace linear estimates (NTLE, hereafter), linear estimates with the associated matrices possessing nil traces, that approximates the original class of estimates $\{\hat{\underset{\sim}{\mu}}_n(h): h \in H_n\}$; (ii) G.C.V. is equivalent to the procedure of minimizing simplified versions of Stein's unbiased risk estimates (SURE, hereafter) for the Stein's estimates (Stein 1981) associated with the original linear estimates $\hat{\underset{\sim}{\mu}}_n(h)$. The first viewpoint mimics one feature of C.V.. In fact, as will be demonstrated later, C.V. is just the C_L applied to a class of zero-diagonal linear estimates, linear estimates with the associated matrices being all zeros along the diagonals. By utilizing nil-trace (or zero-diagonal) estimates, G.C.V. and C.V. circumvate one disturbing feature of directly applying C_L: the dependence on σ^2.

The second viewpoint of G.C.V. turns out more useful in our development. In Section 3, we argue that SURE estimates the <u>true squared error loss</u> of the Stein estimates, <u>not the risks</u>! Under some conditions on ε_i's, we show that SURE is always a consistent estimate of the squared error of the Stein estimate, although sometimes this squared error does not converge at all when the sample size n tends to infinity. Furthermore this consistency

property of SURE has a novel feature: it is uniform over all $\underset{\sim}{\mu}_n$. More precisely, consider the probability that SURE will be within the given ε - neighborhood of the true loss. This probability of course depends on true $\underset{\sim}{\mu}_n$. We prove that the minimum of these probabilities over $\underset{\sim}{\mu}_n \in R^n$ will tend to 1 asymptotically. This seems to be the greatest advantage for considering the combination of Stein estimate and SURE. It certainly can not be enjoyed by any version of C_L (including C.V.) which aims at estimating the risk of some linear estimate that typically possesses unbounded risks over $\underset{\sim}{\mu}_n \in R^n$.

In Section 4, we shall obtain a number of desirable consistency results for the G-cross-validated Stein estimate (GCVSE hereafter) and the SURE associated with GCVSE (GCVSURE hereafter) in the problems of nonparametric regression with nearest neighbor estimates (Section 4.2a), the model selection (Section 4.2.b), the ridge regression (Section 4.3.a), and the smoothing splines (Section 4.3.b). These results will be derived under some conditions on ε_i's and some additional conditions on the weight sequences (for nearest neighbor 1.12 estimates) or the number of models to be chosen and the associated dimensions (for model selection). However, there are absolutely no assumptions explicitly made about $\underset{\sim}{\mu}_n$. It turns out that GCVSE is always consistent whenever given $\underset{\sim}{\mu}_n$ it is possible to choose deterministically a linear estimate $\underset{\sim}{\mu}_n(h)$ from the given class to estimate $\underset{\sim}{\mu}_n$ consistently. Even more appealingly, we prove that GCVSURE is always uniformly consistent. Thus the performance of GCVSE can be assessed satisfactorily by GCVSURE.

As an immediate consequence of the uniform consistency for GCVSURE we are able to obtain a valid confidence set of $\underset{\sim}{\mu}_n$ at least for large n. This will be illustrated for the model selection problem of Example 2 in Section 5.

Although this procedure is apparently conservative and improvements should be feasible, it nevertheless turns out to be the first mathematically valid one utilizing the data-driven techniques. On the other hand, as pointed out before, techniques like C_L or C.V. that concern linear estimates have the inherent difficulties of establishing uniform consistency. Unless these difficulties are removed, valid confidence sets based on these procedures are unlikely (although not totally impossible) to obtain. One may also try to use other sample-reuse methods such as boostrapping or jackknifing to construct some kinds of confidence sets. But again the probability of coverage is hard to guarantee. So far it seems to the author that the only rigorously justified procedure is the method of model-robust confidence set due to Knafl, Sacks, and Ylvisaker (1982) utilizing the appoximate linear model of Sacks and Ylvisaker (1978). When applied to the nonparametric regression, this method may also produce the confidence band for the whole function. But on balance, certain information about the function to be estimated is required; e.g., an upper bound for the size of the second derivatives. Practically, all methods mentioned above should be attempted and the procedure we here introduce naturally draws a valid benchmark for comparisons. Theoretically, one would hope that certain combination of model robust confidence set and GCVSURE may lead to a better solution. Further investigation along this direction seems worthwhile.

Section 6 is devoted to the situation where the variance σ^2 is unknown. Three options are discussed: (i) estimating σ^2 and using GCVSE and GCVSURE (Section 6.1); (ii) returning to the original linear estimates after G.C.V. (Section 6.2); (iii) utilizing the G-cross-validated nil-trace linear estimates, abbreviated by GCVNTLE (Section 6.3). The first option is parti-

cularly useful when there are large degrees of freedom to estimate σ^2: consistency and uniform consistency results of Section 4 are preserved. The second and the third options do not require the estimation of σ^2. We shall show that the consistency of GCVSE implies the consistency of GCVNTLE. The second option amounts to the common practice of G.C.V. and the resulted estimates seem to be the easiest to interpret. However, the consistency is not always guaranteed. For instance, in the ridge regression problem, some restrictions on the eigenvalues of the information matrix are imposed. One of them seems to be unavoidable in view of a counterexample provided there. Fortunately for many setups including Examples 1, 2 and 4, the consistency can be established without much loss of generality. Our results here seem to be the first ones rigorously proving the consistency of the common practice of GCV in these contexts. Intuitive ways of assessing the performance of estimates of options (ii) and (iii) will be mentioned. However, unlike option (i), it seems difficult to establish the uniform consistency for these assessments.

From Section 2 to Section 6, we implicitly assume that $\underset{\sim}{\mu}_n$ could be any vector in R^n. Section 7 discusses a natural variant of G.C.V. when $\underset{\sim}{\mu}_n$ is known to be in a proper linear subspace of R^n. Thus for the Examples 2 and 3, depending on our belief we may have two ways to conduct GCV: (i) the usual version (i.e., minimizing (1.3)) which assumes that no model with rank less than n is completely appropriate, and (ii) the variant version (i.e., minimizing (7.1) of Section 7) which assumes that some model is absolutely correct. Further comparison study about the performance of these two methods may be necessary for the intermediate case that some model though not absolutely correct may be appropriate. Practically, for diagnosis, one certainly should try both methods.

All technical proofs will be given in Section 8.

To close this section, we remark that almost every author referred to in our paper has realized one way or another that the shrinkage phenomena of Stein estimates are relevent in choosing estimates. However none of them directly employ the Stein estimates as we do here. Particularly, in ridge regression our approach is completely different from others (i.e., Casella (1980) and the reference given there) aiming at the minimax estimation.

2. Two viewpoints of G.C.V.: heuristics.

To better understand our motivation, let us briefly review some relevent methods first.

Mallows (1973) introduced the famous statistics C_p and C_L for selecting a model and a linear estimator, respectively. Define $C_p(h)=\sigma^{-2}||y_n-\hat{\mu}_n(h)||^2 - n+2h$ and $C_L(h) = \sigma^{-2}||y_n-\hat{\mu}_n(h)||^2-n + 2tr\ M_n(h)$. $C_L(h)$ is an unbiased estimate of the scaled sum of squared errors for the linear estimate $\hat{\mu}_n(h)$ and $C_p(h)$ is simply $C_L(h)$ when $\hat{\mu}_n(h)$ corresponds to a least squares estimate with h parameters in the associated linear model. For the case of unknown σ^2, it should be replaced by an estimate. This however is rather unpleasant particularly when the degrees of freedom for estimating σ^2 are not large. The instability of this estimation may greatly endanger the performance of C_L (or C_p). Nevertheless if the traces of all $M_n(h)$ are zeros, then there will be no need to estimate σ^2. $C_L(h)$ reduces to $\sigma^{-2}||y_n - \hat{\mu}_n(h)||^2 - n$ and now σ^2 has no influence on the choice of h. Unfortunately no such nil-trace linear estimates have been advocated for practical uses.

Another promising data-driven technique to be discussed is C.V.. C.V. selects h by minimizing the sum of squared prediction errors for y_i, $1 \leq i \leq n$, with y_i itself being excluded from the data set. A rigorous definition requires the specification of estimators to be used for sample size 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 $\sum_{j=1}^{n} \tilde{m}_{ij}(h) y_j$ with $\tilde{m}_{ii}(h)$ being 0. Then the sum of the squared prediction errors is $||\underset{\sim}{y}_n - \tilde{M}_n(h)\underset{\sim}{y}_n||^2$ where $\tilde{M}_n(h) = (\tilde{m}_{ij}(h))$. The common practice of C.V. comprises the minimization of this quantity (let $\tilde{h}$ denote the minimizer) and the advocation of $\hat{\underset{\sim}{\mu}}_n(\tilde{h})$. Observe that the trace of $\tilde{M}_n(h)$ is nil since all diagonal elements are zero. Thus an interesting connection of C.V. to C_L is in order. C.V. is just the C_L on some zero-diagonal linear estimators that approximate the original ones. Thus the difficulty of estimating σ^2 is circumvated. Clearly the success of C.V. may in part depend on the closeness between $\tilde{M}_n(h)$ and $M_n(h)$. In the case of kernel nonparametric regression with one explanatory variable and equi-spaced design points, Wong (1982) established the consistency of C.V.. Li (1982) obtained similar results for nearest neighbor estimates without the restrictions on the number of the explanatory variables and on the structure of design points.

Our first approach to derive G.C.V. is based on the construction of a suitable class of nil-trace linear estimates approximating the original class. For any linear estimate $\hat{\underset{\sim}{\mu}}_n(h)$, consider the linear combination $\bar{\underset{\sim}{\mu}}_n(h) = -\alpha\, \underset{\sim}{y}_n + (1+\alpha)\, \hat{\underset{\sim}{\mu}}_n(h)$ with $\alpha = \text{tr } M_n(h)/(n - \text{tr } M_n(h))$. It is easy to check that the matrix associated with $\bar{\underset{\sim}{\mu}}_n(h)$, $-\alpha I_n + (1+\alpha)M_n(h)$, has trace 0. Now consider the class of NTLE, $\{\bar{\underset{\sim}{\mu}}_n(h): h \in H_n\}$. The C_L procedure on this class amounts to minimizing

$$n^{-1}||y_n - \bar{\mu}_n(h)||^2 = (1+\alpha)^2\, n^{-1}||(I_n - M_n(h))y_n||^2$$

$$= \frac{n^{-1}||y_n - \hat{\mu}_n(h)||^2}{(1 - n^{-1}\mathrm{tr}\, M_n(h))^2}$$

$$= (1.3) \quad .$$

This is exactly G.C.V.! It remains to justify the use of $\{\bar{\mu}_n(h)\}$ to approximate $\{\hat{\mu}_n(h)\}$. This is done in the following theorem from an asymptotic viewpoint.

Theorem 2.1. For any sequence $\{h_n\}$ such that the original linear estimate $\hat{\mu}_n(h_n)$ is consistent in the mean square sense, namely

$$(2.1) \qquad E\, n^{-1}||\mu_n - \hat{\mu}_n(h_n)||^2 \longrightarrow 0, \text{ as } n \to \infty,$$

the corresponding NTLE, $\bar{\mu}_n(h_n)$, will also be consistent, i.e.,

$$(2.2) \qquad n^{-1}||\mu_n - \bar{\mu}_n(h_n)||^2 \longrightarrow 0, \text{ in probability, as } n \to \infty.$$

Moreover, the convergent rate of (2.2) is at least as fast as that of (2.1). If in addition the following condition holds:

$$(2.3) \qquad n^{-1}\mathrm{tr}\, M_n^2(h_n)/(n^{-1}\mathrm{tr}\, M_n(h_n))^2 \longrightarrow \infty,$$

then $\bar{\mu}_n(h_n)$ is asymptotically indifferent from $\hat{\mu}_n(h_n)$ in the sense that

$$(2.4) \qquad ||\bar{\mu}_n(h_n) - \hat{\mu}_n(h_n)||^2 / E||\hat{\mu}_n(h_n) - \mu_n||^2 \longrightarrow 0 \text{ in probability.}$$

(2.3) is frequently satisfied by good estimates in the ill-posed problems; for instance, Examples 1, 3, 4 (See also Golub, Heath and Wahba). Theorem 2.1 guarantees that for large sample size good candidates in the class of NTLE are at least as good as the good ones in the original class. Here note that

nothing is said about the bad estimates. However, since our purpose is to select a good one, essentially we do not lose anything. Nevertheless one thing worth reminding is that this justification is only asymptotically valid. With a small sample size, it is hard to see any appropriate interpretation. This is quite different from C.V.. Although, C.V. requires the definition at sample size n - 1 (hence a general statement like Theorem 2.1 is unlikely) and its rigorous justification is also only asymptotically valid, it seems much easier to accept by practicians.

With the admission of the weakness of the above approach for small sample sizes, we have to seek a different way, hoping that it will be valid regardless of the sample size.

We start with (1.1) and temporarily assume the normality of ε_i's. Consider the case where $M_n(h)$ is symmetric first. Define the Stein estimate associated with $M_n(h)$ by

$$\tilde{\underset{\sim}{\mu}}_n^0(h) = \underset{\sim}{y}_n - \frac{\sigma^2}{\underset{\sim}{y}_n' B_n(h) \underset{\sim}{y}_n} A_n(h)\underset{\sim}{y}_n \quad , \tag{2.5}$$

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

$$B_n(h) = (\text{trace } A_n(h) \cdot I_n - 2A_n(h))^{-1} A_n(h)^2 \, . \tag{2.6}$$

Here the largest characteristic root of $A_n(h)$ is assumed to be less than half of the trace of $A_n(h)$, which often is the case (see Li and Hwang 1982).

Stein showed that $\tilde{\underset{\sim}{\mu}}_n^0(h)$ dominates $\underset{\sim}{y}_n$ under the usual squared error loss.

The relationship between $\tilde{\underset{\sim}{\mu}}_n^0(h)$ and $\hat{\underset{\sim}{\mu}}_n(h)$ of (1.2) was studied from an asymptotic viewpoint by Li and Hwang. The following main results they proved

will be useful for our development later on.

Theorem 2.2. For any sequence $\{h_n\}$ such that (2.1) holds, the associated Stein estimate $\tilde{\underset{\sim}{\mu}}_n^o(h_n)$ is consistent; i.e.,

$$(2.7) \qquad n^{-1}||\tilde{\underset{\sim}{\mu}}_n^o(h_n) - \underset{\sim}{\mu}_n||^2 \longrightarrow 0 \quad \text{in probability.}$$

Moreover, the convergent rate of (2.7) is no slower than that of (2.3) except for the pathological case that $||\underset{\sim}{\mu}_n - \hat{\underset{\sim}{\mu}}_n(h_n)||^2 \longrightarrow 0$, namely, the convergent rate of (2.3) is faster than n^{-1} (note that even if we know that $\mu_1 = \ldots = \mu_n$ and take $\hat{\underset{\sim}{\mu}}_n(h_n) = (\frac{1}{n}\sum_{i=1}^n y_i, \ldots, \frac{1}{n}\sum_{i=1}^n y_i)$, this rate is only n^{-1}!). In addition, if (2.3) holds, then $\tilde{\underset{\sim}{\mu}}_n^o(h_n)$ and $\hat{\underset{\sim}{\mu}}_n(h_n)$ are asymptotically indifferent; namely (2.4) holds with $\bar{\underset{\sim}{\mu}}_n(h_n)$ being replaced by $\tilde{\underset{\sim}{\mu}}_n^o(h_n)$. Therefore, just like NTLE, asymptotically good estimates in the class of Stein estimates $\{\tilde{\underset{\sim}{\mu}}_n^o(h); h \in H_n\}$ are at least as good as the good ones in the original class. But for finite sample sizes, Stein estimates enjoy the nice property that the linear ones do not possess: the boundedness of the risks. In fact, the form (2.1) (or the simplified version (3.1) of Section 3) of Stein estimate, suggests itself to be viewed as a model-robust alternative to the linear estimates $\hat{\underset{\sim}{\mu}}_n(h)$. This may be best illustrated in the model selection context of Example 2. If $\hat{\underset{\sim}{\mu}}_n(h)$ is a good estimate of $\underset{\sim}{\mu}_n$ (in the case that model h is appropriate) then the shrinkage factor $\sigma^2/y_n'B_n(h)y_n$ should be close to 1 and $\tilde{\underset{\sim}{\mu}}_n^o(h)$ would be about the same as $\hat{\underset{\sim}{\mu}}_n(h)$. Otherwise (in the case that model h is inappropriate), $\tilde{\underset{\sim}{\mu}}_n^o(h)$ shrinks $\hat{\underset{\sim}{\mu}}_n(h)$ towards the raw data to gain some model-robustness (for the distinction

between model-robustness and distributional robustness, see Huber (1975)). Therefore, typically no matter what the sample size is, we are justified to replace the original linear estimates by the Stein estimates.

Now define Stein's unbiased risk estimate for $\tilde{\underset{\sim}{\mu}}_n^o(h)$ by

$$\text{SURE}_n^o(h) = \sigma^2 - \frac{\sigma^4 ||A_n(h)\underset{\sim}{y}_n||^2}{n(\underset{\sim}{y}_n' B_n(h)\underset{\sim}{y}_n')^2} . \tag{2.8}$$

Stein showed that $\text{SURE}_n^o(h)$ is an unbiased estimate of the risk of $\tilde{\underset{\sim}{\mu}}^o(h)$; i.e.,

$$E\ \text{SURE}_n^o(h) = E\ n^{-1}||\underset{\sim}{\mu}_n - \tilde{\underset{\sim}{\mu}}^o(h)||^2$$

for any $\underset{\sim}{\mu}_n \in R^n$. To select a good h, it is natural to minimize $\text{SURE}_n^o(h)$ over $h \in H_n$. Equivalently, we may minimize

$$\frac{n(\underset{\sim}{y}_n' B_n(h)\underset{\sim}{y}_n)^2}{||A_n(h)\underset{\sim}{y}_n||^2} . \tag{2.9}$$

Suppose n is large enough so that the largest eigenvalue of $A_n(h)$ is negligible compared to the trace. Then we may approximate (2.6) by $B_n(h) \approx (\text{trace } A_n(h))^{-1} A_n^2(h)$. Substituting this quantity into (2.9), $GCV_n(h)$ of (1.3) is obtained!

For a general $M_n(h)$, Li and Hwang (1982) replaced (2.2) by

$$B_n(h) = (\text{trace } A_n(h) - 2\lambda(A_n(h)))^{-1} A_n(h)' A_n(h)$$

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 estimate becomes

$$(2.10)\qquad \mathrm{SURE}_n^{\Theta}(h) = \sigma^2 - \frac{\sigma^4}{n}\left\{\frac{2r\ \mathrm{tr}\ A_n(h)}{||A_n(h)\underset{\sim}{y}_n||^2} - \frac{r^2}{||A_n(h)\underset{\sim}{y}_n||^2} - \frac{4r(\underset{\sim}{y}_n'A_n(h)'A_n(h)'A_n(h)\underset{\sim}{y}_n)}{||A_n(h)\underset{\sim}{y}_n||^4}\right\}$$

with $r = \mathrm{tr}\ A_n(h) - 2\lambda(A_n(h))$.

This looks rather complicated. But observe that for any $\underset{\sim}{y}_n$,

$$\frac{|\underset{\sim}{y}_n'\ A_n(h)'A_n(h)'A_n(h)\underset{\sim}{y}_n|}{||A_n(h)\underset{\sim}{y}_n||^2} \leq \lambda(A_n(h)).$$

Therefore for large n since $\lambda(A_n(h))$ is negligible compared with $\mathrm{tr}\ A_n(h)$, (2.10) can be simplified and again this will lead to the method of G.C.V. .

In conclusion, we have seen that minimizing SURE is equivalent to the procedure of G.C.V. .

3. Estimating the true loss, not the risk!

In this section we assume that σ^2 is known.

Consider the following simplified version of Stein estimate and SURE:

$$(3.1)\qquad \tilde{\underset{\sim}{\mu}}_n(h) = \underset{\sim}{y}_n - \frac{\sigma^2\ \mathrm{tr}\ A_n(h)}{||A_n(h)\underset{\sim}{y}_n||^2}\ A_n(h)\underset{\sim}{y}_n\ ,$$

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

For brevity, we shall omit the index h and assume that $A_n \neq 0$ in this section.

Observe that to choose between estimates what one really needs to estimate is the true loss, (or better to say, the difference of the true losses for each pair of estimates), not the risk. Since SURE is primarily derived from the need of estimating the risk, its performance for estimating the true loss may be questionable. But surprisingly, we shall see that the opposite side turns out to be the case. We shall show that under some conditions about the ε_i's, SURE is always consistent in estimating the true loss; but it may be inconsistent in estimating the risk. The following example is illustrative.

Example 5. Take $A_n = \text{diag}\,(1, n^{-1/2}, n^{-1/2}, \ldots, n^{-1/2})$, and $\underset{\sim}{\mu}_n = (0,0,\ldots,0)'$. Then

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

which tends to $\dfrac{\sigma^2\varepsilon_1^2}{\varepsilon_1^2 + \sigma^2}$ (a random variable!) as $n \to \infty$. Hence SURE_n can not be a consistent estimate of the risk $E\,\frac{1}{n}\,||\tilde{\underset{\sim}{\mu}}_n - \underset{\sim}{\mu}_n||^2$ since the risk is a non-random number. On the other hand, the Theorem 3.1 below shows that SURE_n estimates the true loss $\frac{1}{n}\,||\tilde{\underset{\sim}{\mu}}_n - \underset{\sim}{\mu}_n||^2$ consistently.

In the following development, the normality assumption is no longer

required. Instead, we assume:

(A.1) The 4th moments of ε_i's are bounded by a constant m;

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

$$\sup_{X \in R} p\{x - a \leq \varepsilon_i \leq x + a\} \leq Ka \text{ , for any } i .$$

Looking at the forms of (3.1) and (3.2), is is clear that if $||A_n \underset{\sim}{y}_n||^2$ takes too small values, then $\tilde{\underset{\sim}{\mu}}_n$ and $SURE_n$ 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 ε_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.

The following theorem is the main result of this section.

Theorem 3.1. Assume (A.1) and (A.2) hold. Then for any $\delta > 0$ we have

$$\sup_{\underset{\sim}{\mu}_n \in R^n} p\{|\, SURE_n - n^{-1}||\tilde{\underset{\sim}{\mu}}_n - \underset{\sim}{\mu}_n||^2| \geq \delta\} \to 0 \text{ as } n \to \infty \tag{3.3}$$

Theorem 3.1 establishes the uniform consistency property for $SURE_n$ as an estimate of the true loss of Stein estimate $\tilde{\underset{\sim}{\mu}}_n$. No assumptions about the matrix M_n are required here. Roughly speaking, the boundedness of the risks of Stein estimates makes this theorem plaussible. Thus it does not seem likely for the same results to hold in the case of using linear estimates together with their unbiased risk estimates.

Let us briefly discuss the convergent rate. Observe that

$$n^{-1}||\tilde{\underset{\sim}{\mu}}_n - \underset{\sim}{\mu}||^2 = n^{-1}||\underset{\sim}{\varepsilon}_n - \frac{\sigma^2 \text{tr } A_n}{||A_n \underset{\sim}{y}_n||^2} A_n \underset{\sim}{y}_n||^2$$

$$= n^{-1}||\underset{\sim}{\varepsilon}_n||^2 - 2n^{-1} < \underset{\sim}{\varepsilon}_n, \frac{\sigma^2 \text{tr } A_n}{||A_n \underset{\sim}{y}_n||^2} A_n \underset{\sim}{y}_n > + \frac{\sigma^4 (\text{tr } A_n)^2}{n||A_n \underset{\sim}{y}_n||^2} .$$

From this, it follows that

$$(3.4) \qquad |\text{SURE}_n - n^{-1}||\tilde{\underset{\sim}{\mu}}_n - \underset{\sim}{\mu}_n||^2| \leq |n^{-1}||\underset{\sim}{\varepsilon}_n||^2 - \sigma^2| + \frac{2\sigma^2 |\text{tr } A_n|}{n||A_n \underset{\sim}{y}_n||^2} | < \underset{\sim}{\varepsilon}_n, A_n \underset{\sim}{\varepsilon}_n > - \sigma^2 \text{tr } A_n| + \frac{2\sigma^2 |\text{tr } A_n|}{n||A_n \underset{\sim}{y}_n||^2} | < \underset{\sim}{\varepsilon}_n, A_n \underset{\sim}{\mu}_n > | .$$

The first term of the right hand side converges to 0 at rate $n^{-1/2}$. Thus it seems that the convergent rate of the left hand side of (3.4) can not be faster than $n^{-1/2}$. On the other hand, one frequently has the convergent rate of $n^{-1}||\tilde{\underset{\sim}{\mu}}_n - \underset{\sim}{\mu}_n||^2$ faster than $n^{-1/2}$. For instance, in Example 1, with h_n being appropriately chosen, the convergent rate $n^{-2/3}$ can be achieved if f has a bounded first derivative. For optimal convergent rates in the non-parametric regression, see Stone (1981, 1983). This generates the following unsolved problem:

Does there exist an estimate of $n^{-1}||\tilde{\mu}_n - \mu_n||^2$ with the desired convergent rate if μ_n is associated with a fixed smooth f, such that at the same time (3.3) can still hold?

4. Consistency of GCVSE and uniform consistency of GCVSURE.

Assume σ^2 is known. We shall investigate the consistency problem of GCVSE and GCVSURE in this section. Certainly some assumptions will be needed on the three ingredients of the problem: the unknown parameters of interest μ_n, the class of matrices $\{M_n(h): h \in H_n\}$; and the distribution of random error ε_i. Our assumptions on the error distribution may be further weakened since here we only mean to keep the proofs simple. But the assumptions on the class of matrices will be mild enough to cover most cases under study. Since the consistency of GCVSURE will be uniform over $\mu_n \in R^n$, of course nothing on μ_n is assumed in its proof. On the other hand, it seems unlikely for GCVSE to be uniformly consistent; the consistency of GCVSE is tied to the sequence $\{\mu_n\}$. However, fortunately we do not need any explicit assumptions of $\{\mu_n\}$; otherwise, we may get into the trouble of justifying them. Specifically, what we need is that consistency can hold with a deterministic choice of estimates; namely, there exists a deterministic sequence $h_n \in H_n$, $n = 1,2,\ldots$, such that

$$n^{-1}||\hat{\mu}_n(h_n) - \mu_n||^2 \longrightarrow 0 \text{ in probability.} \tag{4.1}$$

Here h_n can even depend on μ_n! We shall show that GCVSE is consistent:

$$n^{-1}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2 \longrightarrow 0 \text{ , in probability,} \tag{4.2}$$

and that GCVSURE is uniformly consistent: for any $\delta > 0$, and $n \to \infty$,

$$(4.3) \qquad \sup_{h \in R^n} p\{|SURE_n(\hat{h}) - n^{-1}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2| \geq \delta\} \to 0.$$

where $\hat{h}$ is chosen by GCV; or equivalently, $\hat{h}$ achieves

$$(4.4) \qquad \min_{h \in H_n} SURE_n(h) \quad .$$

The most difficult step in obtaining these results is to establish the following: for any sequence $\{\mu_n\}$,

$$(4.5) \qquad \sup_{h \in H_n} |SURE_n(h) - \frac{1}{n}||\tilde{\mu}_n(h) - \mu_n||^2| \longrightarrow 0 \text{ in probability.}$$

Once (4.5) is established, then immediately $|SURE_n(\hat{h}) - \frac{1}{n}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2| \leq \sup_{h \in H_n} |SURE_n(h) - \frac{1}{n}||\tilde{\mu}_n(h) - \mu_n||^2| \longrightarrow 0$ in probability. Now if (4.3) does not hold, then there must be some $\delta > 0$ and some h_n such that $p\{|SURE_n(\hat{h}) - n^{-1}||\tilde{\mu}_n(h) - \mu_n||^2| \geq \delta\} \not\to 0$, contradicting the above statement. Thus we have seen that (4.5) implies (4.3). (4.2) also follows from (4.5) and (4.1) as to be elaborated below.

First observe that

$$n^{-1}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2 \leq SURE_n(\hat{h}) + |SURE_n(\hat{h}) - n^{-1}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2|$$

$$\leq SURE_n(h_n) + |SURE_n(\hat{h}) - n^{-1}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2|$$

(by (4.4))

$$\leq n^{-1}||\tilde{\mu}(h_n) - \mu_n||^2 + |SURE_n(h_n) -$$

$$n^{-1}||\tilde{\mu}_n(h_n) - \mu_n||^2| + |SURE_n(\hat{h}) - n^{-1}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2|.$$

The second and the third terms in the last expression tend to 0 because of (4.5). By (4.1) and Theorem 2.2 (which holds obviously for $\tilde{\underset{\sim}{\mu}}_n(h_n)$ in replacement of $\tilde{\underset{\sim}{\mu}}_n^o(h_n)$), the first term also tends to 0. Hence (4.2) is proved. We summarize what we have obtained by the following

Theorem 4.1. Under (4.5), (4.3) holds. Suppose in addition (4.1) holds. Then (4.2) holds.

However (4.5) does not always hold. Certain conditions on the class $\{M_n(h): h \in H_n\}$ are unavoidable. For instance, when H_n is discrete, the cardinality, $\#H_n$, may not be too large; when H_n is continuous, good analytical properties on the matrix valued function $M_n(\cdot)$ may be imposed. Instead of developing a general theorem to cover all situations, in what follows we shall look at a number of useful cases individually.

4.1. Bounded $\#H_n$.

The following theorem follows immediately from Theorem 3.1 (which implies (4.5)) and Theorem 4.1.

Theorem 4.2. Assume that (A.1) and (A.2) hold and that $\sup\{\#H_n: n = 1,2,\ldots,\} < \infty$. Then (4.3) holds. If in addition (4.1) holds, then (4.2) holds.

4.2. Finte $\#H_n$.

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') ε_i's have mean 0, common 2nd, 4th and 6th moments, and their 8th moments are bounded by a constant M.

The following theorem will be useful in verifying (4.5). Let $\underline{\lambda}'(A_n(h))$ 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 and δ only) such that for any $\underset{\sim}{\mu}_n \in R^n$,

$$(4.6) \qquad p\{\sup_{h\in H_n} |SURE_n(h) - \frac{1}{n}||\tilde{\underset{\sim}{\mu}}_n(h) - \underset{\sim}{\mu}_n||^2| \geq 2\delta\}$$

$$\leq p\{|n^{-1}||\underset{\sim}{\varepsilon}_n||^2 - \sigma^2| \geq \delta\} + C_1 n^{-2} \#H_n + C_2 \sum_{h\in H_n} \frac{[\lambda'(A_n(h))]^4}{(\mathrm{tr}\, A_n'(h)A_n(h))^2} .$$

4.2.a Nearest neighbor estimates in nonparametric regression.

Let p be a natural number and $\mathcal{X}$ be the compact closure of an open set in R^p. Suppose $y_1, y_2, \ldots, y_n$ are observed at levels $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots \underset{\sim}{x}_n \in \mathcal{X}$ with $\underset{\sim}{x}_i \neq \underset{\sim}{x}_j$ for $i \neq j$ such that the expected value μ_i of y_i is equal to $f(\underset{\sim}{x}_i)$ for an unknown continuous function f on $\mathcal{X}$. Let $\underset{\sim}{x}_{i(j)}$ denote the $j^{\underline{th}}$ nearest neighbor of $\underset{\sim}{x}_i$ in the sense that $||\underset{\sim}{x}_i - \underset{\sim}{x}_{i(j)}||$ is the $j^{\underline{th}}$ 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 $h\in H_n$, consider $\hat{\underset{\sim}{\mu}}_n(h)$, the h-nearest neighbor estimate of $\underset{\sim}{\mu}_n$. The $i^{\underline{th}}$ coordinate of $\hat{\underset{\sim}{\mu}}_n(h)$ is given by $\sum_{j=1}^{h} w_{n,h}(j) y_{i(j)}$ with $w_{n,h}(\cdot)$ being a non-negative weight function such that

$$(4.7) \qquad \sum_{i=1}^{h} w_{n,h}(i) = 1 \quad .$$

The rows of $M_n(h)$ are certain permutations of $(w_{n,h}(1),\ldots,w_{n,h}(h),0,\ldots,0)$ and the diagonal elements of $M_n(h)$ are equal to $w_{n,h}(1)$. To ensure (4.1), we assume that

$$(4.8) \qquad \text{there exists a sequence } \{h_n\} \text{ such that } h_n/n \to 0 \text{ and}$$

$$w_{n,h_n}(1) \to 0, \text{ as } n \to \infty .$$

It can be easily verified that (4.8) implies the consistency of $\hat{\underset{\sim}{\mu}}_n(h_n)$ (see, e.g., Li (1982)). Stone (1977) gives more general consistency results for nearest neighbor non-parametric regression.

In addition to (4.7) and (4.8), we have the following two mild restrictions on the weight function:

(4.9) There exists a positive number δ' such that $w_{n,h}(1) \leq 1 - \delta'$ for any n, $h \geq 2$.

(4.10) For any n,h and i, $w_{n,h}(i) \geq w_{n,h}(i+1)$.

To use Theorem 4.3, we need to have an upper bound for $\lambda'(A_n(h))$. This can be derived from following lemma.

Lemma 4.1. Under (4.7) and (4.10), there exists a constant Λ (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 $\lambda'(A_n(h)) \leq (1+\Lambda)$. Now using this bound and the observation that

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

we see that the right hand side of (4.6) is bounded by

$$p\{|n^{-1}||\underset{\sim}{\varepsilon}_n||^2 - \sigma^2| \geq \delta\} + C_1 n^{-1} + C_2(1+\Lambda)^4(1-\delta')^{-4}n^{-1}$$

which obviously tends to 0 as $n \to \infty$. Hence (4.5) is established. Now the

following main result for nearest neighbor nonparametric regression follows from Theorem 4.1.

Theorem 4.4. Under (A.1'), and (4.7) $\sim$ (4.10), the Steinized nearest neighbor estimate $\tilde{\mu}_n(\hat{h})$ with $\hat{h}$ being chosen by G.C.V. is consistent. Moreover, the associated Stein unbiased risk estimate $\text{SURE}_n(\hat{h})$ is a uniformly consistent estimate of the true loss $n^{-1}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2$.

4.2.b. 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 any 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 non-singular. Consider the least squares estimate $\hat{\mu}_n(h)$ of (1.4) and its Steinized version $\tilde{\mu}_n(h)$. When d(h) = n, define $\tilde{\mu}_n(h) = y_n$ and $\text{SURE}_n(h) = \sigma^2$. Here in advocating $\tilde{\mu}_n(h)$ we implicitly assume that none of models with ranks less than n are completely appropriate. Otherwise, we shall proceed differently; see Section 7 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 1982).

Recall that here $A_n(h)$ is a projection matrix with rank n - d(h). Thus the last term on the right hand side of (4.6) equals $C_2 \sum_{h\in H_n} (n-d(h))^{-2}$. Hence the left hand side of (4.6) will tend to 0 if $\#H_n/n^2 \to 0$ and $\sum_{h\in H_n} (n-d(h))^{-2} \to 0$. However occasionally some models in the suggested class H_n may have large numbers of parameters so that for these h, n-d(h)

may be quite small. If there are not too many such models in each H_n, Theorem 4.2 can be utilized to circumvate this difficulty. Specifically, we have the following

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

(4.11) for any positive number ε, 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 $\sum_{h \notin H_n'} (n-d(h))^{-2} \leq \varepsilon$.

Then $SURE_n(\hat{h})$ with $\hat{h}$ chosen by G.C.V. is uniformly consistent. Furthermore $\tilde{\mu}_n(\hat{h})$ is consistent whenever given μ_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, $\#H_n = p_n \leq n$. Put $H_n' = \{n, n-1, \ldots, n-k-1\} \cap H_n$. Then $\sum_{h \in H_n'} (n-d(h))^{-2} \leq \sum_{i=k}^{n} i^{-2}$. Since $\sum_{i=1}^{\infty} i^{-2}$ converges, it is easy to see that (4.11) can be satisfied for a suitably-chosen k. Hence Theorem 4.5 applies here.

4.3. Continuous H_n.

Two cases for $H_n = \{h: h \geq 0\}$ will be considered; namely, the ridge regression and the smoothing splines. In fact, the results for smoothing splines follow immediately from those for ridge regression. Hence we first look at

4.3.a. Ridge regression.

Consider the Steinized ridge regression estimate $\tilde{\mu}_n(h)$ associated with $\hat{\mu}_n(h)$ of (1.7). Here $\hat{\mu}_n(0)$ is defined by $\lim_{h \to 0} \tilde{\mu}_n(h)$ ($\neq y_n$ unless all $\lambda_{i,n}$, $i=1,\ldots,n$, are equal) and similarly for $\mathrm{SURE}_n(0)$. Again in advocating $\tilde{\mu}_n(h)$, we implicitly assume that the model (1.6) 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 δ_i's being nuisance parameters. This is the approximate linear model of Sacks and Ylvisaker although we do not specify a bound for δ_i's. If (1.6) is completely appropriate, we may proceed differently; see Section 7.

Since we shall work with the transformed data $\bar{y}_n$ (recall the definition from the line following (1.9)) and the independence for the components of $\bar{y}_n$ is desired, we shall impose the following:

(A.1'') $\quad \varepsilon_i$'s are i.i.d. $N(0,\sigma^2)$.

Under (A.1''), we see that $\bar{y}_i$, the $i^{\underline{th}}$ component of $\bar{y}_n$, satisfied

$$\bar{y}_i = \bar{\mu}_i + \bar{\varepsilon}_i$$

with $(\bar{\mu}_1,\ldots,\bar{\mu}_n)' = \bar{\mu}_n = U'\mu_n$, and $\bar{\varepsilon}_i$'s being i.i.d. $N(0,\sigma^2)$ again. Since Euclidean norm is invariant under orthogonal transformation, we rewrite all the relevant quantities in terms of $\bar{y}_n$, $\bar{\mu}_n$ and $\bar{\varepsilon}_n = (\bar{\varepsilon}_1,\ldots,\bar{\varepsilon}_n)'$. Put

$$\bar{M}_n(h) = \mathrm{Diag}\,(\lambda_1(\lambda_1+h)^{-1},\ldots,\lambda_n(\lambda_n+h)^{-1}), \tag{4.12}$$

and

$$\bar{A}_n(h) = I - \bar{M}_n(h) .$$

Here we abbreviate λ_i for $\lambda_{i,n}$. Then, we have

(4.13)
$$\bar{\hat{\mu}}_n(h) \equiv U'\hat{\mu}_n(h) = \bar{M}_n(h)\bar{y}_n ,$$

$$\bar{\tilde{\mu}}_n(h) \equiv U'\tilde{\mu}_n(h) = \bar{y}_n - \frac{\sigma^2 \operatorname{tr} \bar{A}_n(h)}{||\bar{A}_n(h)\bar{y}_n||^2}\bar{A}_n(h)\bar{y}_n ,$$

$$\overline{\mathrm{SURE}}_n(h) = \sigma^2 - \frac{\sigma^4(\operatorname{tr} \bar{A}_n(h))^2}{n||\bar{A}_n(h)\bar{y}_n||^2} ,$$

and

$$||\bar{\tilde{\mu}}_n(h) - \mu_n||^2 = ||\bar{\tilde{\mu}}_n(h) - \bar{\mu}_n||^2 .$$

Therefore thinking of the transformed data $\bar{y}_n$ as y_n, our ridge regression problem reduces to establishing (4.5) for the special case that $\hat{\mu}_n(h) = M_n(h)y_n$ with $M_n(h)$ taking the diagonal form of (4.12).

Lemma 4.2. Assume that $M_n(h)$ takes the diagonal form (4.12) for $h \geq 0$. Then under (A.1) and (A.2), (4.5) holds.

Based on this lemma, we immediately obtain

Theorem 4.6. Under (A.1''), for the ridge regression problem with ridge estimate (1.7), $\mathrm{SURE}_n(\hat{h})$ with $\hat{h}$ chosen by G.C.V. 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.

Our work here seems to be the first general asymptotic study in the ridge regression literature. The appealing features are (i) there is no need to specify how to build up the sequence of models for different sample sizes; (ii) no explicit assumptions are made about the asymptotic behavior of μ_n; (iii) p_n may be taken as ∞ providing that the associated summations

converge.

4.3.b. 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

$$S_n^k = \{f: f \in C^{2k-2}[0,1],\ f \text{ is a polynomial of degree } 2k-1 \text{ on } (x_i, x_{i+1}),\ i=1,\ldots,n-1, \text{ and } f^{(k)} \equiv 0 \text{ on } [0,x_1] \text{ and } [x_n,1]\}.$$

Consider the basis for S_n^k introduced by Demmler and Reinsch (1975) (see also, Speckman 1981 a, b; 1982) consisting of eigen functions $\{\phi_{jn}\}_{j=1}^n$ along with eigenvalues $\{p_{kn}\}_{k=1}^n$ satisfying

$$\frac{1}{n} \sum_{i=1}^{n} \phi_{jn}(x_i)\, \phi_{j'n}(x_i) = \delta_{jj'}, \tag{4.14}$$

$$\int_0^1 \phi_{jn}^{(k)}(x)\, \phi_{j'n}^{(k)}(x)\, dx = p_{jn}\delta_{jj'},$$

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

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

Here $\delta_{jj'}$ is the Kronecker delta. Using this basis, (1.11) is equivalent to

$$\min_{\underset{\sim}{C} \in R^n} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{n} C_j n^{-1/2} \phi_j(x_i)\Big)^2 + h \sum_{j=k+1}^{n} C_j^2 p_{jn} \tag{4.15}$$

Here $\underset{\sim}{C} = (C_1, C_2 \ldots, C_n)'$. Let U_n denote the n x n matrix with the ij$^{\text{th}}$ element $n^{-1/2}\, \phi_j(x_i)$. From (4.14), it follows that $U_n' U_n = I_n$. Put $\underset{\sim}{\bar{y}}_n = U_n' \underset{\sim}{y}_n$. Then (4.15) reduces to

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

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

$$C^* = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, (1+hp_{k+1,n})^{-1} \bar{y}_{k+1}, \ldots, (1+hp_{n,n})^{-1} \bar{y}_n)'$$

Put $\lambda_1 = \lambda_2 = \ldots = \lambda_k = \infty$ and $\lambda_i = p_{i,n}^{-1}$ for $i = k+1, \ldots, n$. We see that C^* takes the form $\tilde{\bar{\mu}}_n(h)$ of (4.13) when the first k diagonal elements in the matrix $\bar{M}_n(h)$ of (4.12) are interpreted as ones. Hence in terms of $\bar{y}_n$, 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 $\text{SURE}_n(\hat{h})$ with $\hat{h}$ chosen by G.C.V. is uniformly consistent in estimating the true loss. In addition, if given the true μ_n, there exists a sequence of nonnegative numbers $\{h_n\}$ such that the corresponding smoothing spline solution of (1.11) is consistent, then the G-cross-validated Steinized smoothing spline estimate $\tilde{\mu}_n(\hat{h})$ is consistent.

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

5. Confidence sets for model selection.

(4.3) guarantees that for large n we may obtain a valid confidence set based on GCVSE and GCVSURE. Let $p(n,\delta)$ denote the left hand side of (4.3). Then to construct a confidence set of μ_n with the coverage probability at least $1 - \alpha$, we may choose $\delta = \delta_n(\alpha)$ with $p(n,\delta_n(\alpha)) \leq \alpha$ to form the n-dimensional ball $\{\mu_n : n^{-1}||\mu_n - \tilde{\mu}_n(\hat{h})||^2 \leq \text{SURE}_n(\hat{h}) + \delta_n(\alpha)\}$.

(4.3) guarantees that we may have $\delta_n(\alpha) \to 0$ as $n \to \infty$. In fact, for all cases where Section 4 has studied, we can always obtain $\delta_n(\alpha)$ by the bounds used in the proofs of (4.5) (see Section 8). But these bounds are too crude. This section will demonstrate by one case that sometimes with a more careful evaluation of the probabilities involved we may obtain a better $\delta_n(\alpha)$. Note that in view of the discussion given at the end of Section 3, the convergent rate of $\delta_n(\alpha)$ is unlikely to be faster than $n^{-1/2}$. The $\delta_n(\alpha)$ obtained in the case studied here achieves this rate.

We shall consider the model selection problem of Example 2. To simplify a little bit the computation, we assume that $p_n \leq \frac{n}{2}$. We remind the readers again that implicitly we think that none of these models are completely appropriate; otherwise we should proceed differently (see Section 7). We assume further that ε_i's are i.i.d. $N(0,\sigma^2)$.

Now, since the models are nested, by standard methods we may transform the problem into the canonical form that the p_n columns of the matrix X are mutually orthogonal. Furthermore, the normality assumption allows us to transform y_n so that the relevant estimates take the following simple form:

for each $h \in H_n = \{1,2,\ldots,p_n\}$, $\hat{\mu}_n(h) = (y_1,\ldots,y_h,0,\ldots,0)'$,

$$\tilde{\mu}_n(h) = (y_1,y_2,\ldots,y_h, [1 - \sigma^2(n-h)(\sum_{i=h+1}^{n} y_i^2)^{-1}]y_{h+1},\ldots,[1-\sigma^2(n-h)(\sum_{i=h+1}^{n} y_i^2)^{-1}]y_n)'$$

and $$\text{SURE}_n(h) = \sigma^2 - \sigma^4(n-h)^2 n^{-1} (\sum_{i=h+1}^{n} y_i^2)^{-1}.$$

We have the following theorem to find $\delta_n(\alpha)$.

Theorem 5.1. With the definitions of $\tilde{\mu}_n(h)$, $\text{SURE}_n(h)$ given above, we have

$$(5.1) \qquad p\left\{\sup_{h\in H_n} \left| \mathrm{SURE}_n(h) - \frac{1}{n}\left|\left|\tilde{\mu}_n(h) - \mu_n\right|\right|^2\right| \geq \delta \right\}$$

$$\leq 8\sigma^4 n^{-1}\delta^{-2} + 64(1 + (2\sigma^4\delta^{-2})^{1/3})^3(n - p_n)^{-1} .$$

We now use this theorem to get a $\delta_n(\alpha)$ with the convergent rate $n^{-1/2}$. First, supposing that $\sqrt{2}\,\sigma^2 > \delta$, then one can get a simpler bound: $8\sigma^4 n^{-1}\delta^{-2} + 1024\,\sigma^4\delta^{-2}(n-p_n)^{-1}$. Then we may let $\delta_n(\alpha) = \alpha^{-1/2}(8n^{-1} + 1024(n-p_n)^{-1})^{1/2}\sigma^2 \leq 32\sqrt{2}\,\alpha^{-1/2}\,\sigma^2(n-p_n)^{-1/2}$. Now since $P_n \leq n/2$, $\delta_n(\alpha)$ is of order $n^{-1/2}$ as desired. However, in order that $\sqrt{2}\,\sigma^2 > \delta_n(\alpha)$ we need $n - p_n$ to be large, i.e., $n - p_n \geq 1024\alpha^{-1}$. With α = 10% and $p_n = n/2$, we need at least 20,000 observations to make this δ_n work! In addition, the right hand side of (5.1) is at least as large as $64(n - p_n)^{-1}$. Thus for $\alpha \leq 64(n - p_n)^{-1}$, this method breaks down completely. Hence it seems necessary to obtain a better evaluation of the left hand side of (5.1) for small n. Another possibility (a very challenging job!) may be to directly evaluate the probability that $\left|\mathrm{SURE}_n(\hat{h}) - \frac{1}{n}\left|\left|\tilde{\mu}_n(\hat{h}) - \mu_n\right|\right|^2\right| \geq \delta$. Finally, the heuristics of Section 8 of Stein 1981 may provide better insights for our problem.

6. Unknown variance of sampling error.

Three options will be discussed below when σ^2 is unknown.

6.1. Estimating σ^2.

As mentioned before G.C.V. does not require σ^2. Consequently, we may still use it to select $\hat{h}$. After $\hat{h}$ being chosen, we may estimate μ_n by the Stein estimate $\tilde{\mu}_n(\hat{h})$ with the unknown σ^2 substituted by a good estimate $\hat{\sigma}^2_n$.

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 $\text{SURE}_n(\hat{h},\hat{\sigma}_n)$ defined to be the $\text{SURE}_n(\hat{h})$ with σ substituted by $\hat{\sigma}_n$. The consistency and uniform consistency properties are preserved if $\hat{\sigma}_n^2$ is consistent. More precisely, we have

Theorem 6.1. Assume that $\hat{\sigma}_n^2$ is a consistent estimate of σ^2. Then $\tilde{\mu}_n(\hat{h},\hat{\sigma}_n)$ and $\text{SURE}_n(\hat{h},\hat{\sigma}_n)$ are consistent whenever given σ^2, $\tilde{\mu}_n(\hat{h})$ and $\text{SURE}_n(\hat{h})$ are consistent respectively. Moreover, if the distribution of $\hat{\sigma}_n^2$ does not depend on μ_n, then $\text{SURE}_n(\hat{h},\hat{\sigma}_n)$ is uniformly consistent, whenever given σ^2, $\text{SURE}_n(\hat{h})$ is uniformly consistent.

Perhaps the most natural case to have a $\hat{\sigma}_n^2$ whose distribution does not depend on μ_n is when there are replications of the data y_n. Another common situation is in ridge regression or model selection where one may have a completely appropriate model available and the residual sum of squares for the least squares estimates under this true model can be used to construct $\hat{\sigma}_n^2$ (in this case, a modification of G.C.V. is needed; see Section 7)). However, sometimes we may get $\hat{\sigma}_n^2$ that may depend on μ_n. For example, in Example 1, we may take $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n/2}(y_{2i-1} - y_{2i})^2$ for n even. Supposing that as n increases, the x values get dense in [0,1], it is easy to see that $\hat{\sigma}_n^2 \to \sigma^2$ if f is continuous. This method of constructing $\hat{\sigma}_n^2$ extends naturally to higher dimension cases. The pairs of differences $y_{2i-1} - y_{2i}$ may be so formed that the corresponding explanatory variables take values as close to each other as possible. Rice (1983) had considered such variance estimates in a study of utilizing the C_L procedure to select the bandwidth of a kernel non-parametric regression in R^1. He even suggested the use of higher order differences in replacement of the first order difference $y_{2i-1} - y_{2i}$

to reduce the bias of $\hat{\sigma}_n^2$.

6.2. Returning to the original linear estimates.

After G.C.V., the common practice is to return to the original linear estimates, namely $\hat{\underset{\sim}{\mu}}_n(\hat{h})$. We shall discuss the consistency problem of $\hat{\underset{\sim}{\mu}}_n(\hat{h})$ below.

To begin with, we have the following

Lemma 6.1 Assume that

(6.2.1) both $\tilde{\underset{\sim}{\mu}}_n(\hat{h})$ and $\mathrm{SURE}_n(\hat{h})$ are consistent.

Then $\hat{\underset{\sim}{\mu}}_n(\hat{h})$ is consistent if and only if

$$(6.2.2) \qquad n^{-1}\mathrm{tr}\, A_n(\hat{h}) \to 1, \text{ in probability.}$$

A technical step to establish (6.2.2) is to obtain the following statement:

$$(6.2.3) \qquad \lim_{n\to\infty} p\{ \frac{||A_n(\hat{h})\underset{\sim}{y}_n||^2}{||A_n(\hat{h})\underset{\sim}{\mu}_n||^2 + \sigma^2 \mathrm{tr}\, A_n^2(\hat{h})} \le 1 - \delta \} = 0, \text{ for}$$

any $\delta > 0$.

From (6.2.1) and (6.2.3), we get that for any $\delta_1, \delta_2 > 0$

$$1 = \lim_{n\to\infty} p\{ \sigma^2 \ge \frac{n||A_n(\hat{h})\underset{\sim}{y}_n||^2}{(\mathrm{tr}\, A_n(\hat{h}))^2} - \delta_1 \} \qquad \text{(by 6.2.1))}$$

$$\le \lim_{n\to\infty} p\{ \sigma^2 \ge \frac{||A_n(\hat{h})\underset{\sim}{y}_n||^2}{||A_n(\hat{h})\underset{\sim}{\mu}_n||^2 + \sigma^2 \mathrm{tr}\, A_n^2(\hat{h})} \cdot \frac{n\sigma^2 \mathrm{tr}\, A_n^2(\hat{h})}{(\mathrm{tr}\, A_n(\hat{h}))^2} - \delta_1 \}$$

$$\le \lim_{n\to\infty} p\{ \sigma^2 \ge (1 - \delta_2) \cdot \frac{n\sigma^2 \mathrm{tr}\, A_n^2(\hat{h})}{(\mathrm{tr}\, A_n(\hat{h}))^2} - \delta_1 \}$$

$$+ \lim_{n\to\infty} p\left\{ \frac{||A_n(\hat{h})y||^2}{||A_n(\hat{h})\underset{\sim}{\mu}_n||^2+\sigma^2 \operatorname{tr} A_n^2(\hat{h})} \leq 1 - \delta_2 \right\}$$

$$\leq \lim_{n\to\infty} p\left\{(1+\sigma^2/\delta_1)(1-\delta_2)^{-1} \geq \frac{n \operatorname{tr} A_n^2(\hat{h})}{(\operatorname{tr} A_n(\hat{h}))^2}\right\} \quad \text{(by (6.2.3))}.$$

Therefore, for any $\delta > 0$,

$$1 = \lim_{n\to\infty} p\left\{(1+\delta) \geq \frac{n \operatorname{tr} A_n^2(\hat{h})}{(\operatorname{tr} A_n(\hat{h}))^2}\right\} .$$

Since $n \operatorname{tr} A_n^2(\hat{h})$ is always no less than $(\operatorname{tr} A_n(\hat{h}))^2$, it follows that

$$\text{(6.2.4)} \qquad \frac{n \operatorname{tr} A_n^2(\hat{h})}{(\operatorname{tr} A_n(\hat{h}))^2} \longrightarrow 1 \text{ in probability.}$$

In many cases, we shall see that (6.2.4) implies (6.2.2) and hence the consistency of $\hat{\underset{\sim}{\mu}}_n(\hat{h})$. The simplest case is the model selection where all $A_n(h)$ are projection matrices i.e., $A_n^2(h) = A_n(h)$. In this case (6.2.4) and (6.2.2) are identical. Next, consider the case of nearest neighbor non-parametric regression (Section 4.2.a) where $n^{-1} \operatorname{tr} A_n^2(\hat{h}) = (1 - w_{n,\hat{h}}(1))^2 + \sum_{i=2}^{\hat{h}} w_{n,\hat{h}}^2(i)$ and $\operatorname{tr} A_n(\hat{h}) = n(1 - w_{n,\hat{h}}(1))$. (6.2.4) implies that $0 = \operatorname{plim}_{n\to\infty} \left[\sum_{i=2}^{\hat{h}} w_{n,\hat{h}}^2(i)\right] / (1 - w_{n,\hat{h}}(1))^2 \geq \operatorname{plim}_{n\to\infty} (\hat{h}-1)^{-1}$ (by Cauchy-Schwartz and (4.7)). Therefore $\hat{h} \to \infty$ as $n \to \infty$. Now assume that the weight functions satisfy the following regularity condition:

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

Then "$\hat{h} \to \infty$" implies "$w_{n,\hat{h}}(1) \to 0$" which in turn implies (6.2.2).

The cases of ridge regression and smoothing spline are a little bit complicated. We need the following condition on the eigenvalues $\lambda_{i,n}$:

(6.2.6) there exist p and q, $0 < p < q < 1$, such that

$$\overline{\lim_{n\to\infty}} \ \lambda_{[qn],n} / \lambda_{[pn],n} < 1$$

where [x] denotes the greatest integer $\leq x$.

Lemma 6.2 Assume (6.2.6) holds. Then (6.2.4) implies (6.2.2).

Condition (6.2.6) means that the asymptotic distribution of eigenvalues should not concentrate at only one point. Violating this condition, we might get inconsistency. This is demonstrated by the following

Example 6. Consider the canonical case with $X_n = \text{diag}\,(2,1,\ldots,1)$. Here $\lambda_{1,n} = 4$, $\lambda_{2,n} = \ldots = \lambda_{n,n} = 1$ and G.C.V. chooses h by minimizing

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

over $h \geq 0$. Straightforwardly, we inspect the derivative of the last expression with respect to h and conclude that

$$\hat{h} = \infty \text{ , if } r \geq 1 \text{ ,}$$

$$= (4r - 1)(1 - r)^{-1} \text{ , if } 1/4 < r < 1 \text{ ,}$$

$$= 0 \text{ , if } 0 \leq r \leq 1/4 \text{ ,}$$

where the random variable r is defined by $(n-1)^{-1} \sum_{i=2}^{n} y_i^2 / y_1^2$.

Now suppose $\mu_n = 0$ and ε_i's are i.i.d $N(0,\sigma^2)$. Clearly, $\lim_{n\to\infty} p\{\hat{h} = 0\}$

$= \lim_{n\to\infty} p\{ r \le 1/4 \} = \lim p\{ 4\sigma^2 \le \varepsilon_1^2 \} \approx 0.05.$

Since n^{-1} tr $A_n(0) = 0$, (6.2.2) does not hold. However, by Theorem 4.6., $\tilde{\mu}_n(\hat{h})$ and $SURE_n(\hat{h})$ are consistent when $\mu_n = 0$.

This example and the condition (6.2.6) indicate that the common practice of G.C.V. 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 methods of Sections 6.1 and 6.3 (as we shall see) do not have this problem.

To completely establish the consistency of $\hat{\mu}_n(\hat{h})$, it remains to verify (6.2.3). However sometimes we may need further conditions. Case by case, we state our conclusions in the following.

Theorem 6.2. For the nearest neighbor non-parametric regression problem, assume (A.1'), (4.7) $\sim$ (4.10), and (6.2.4). Then $\hat{\mu}_n(\hat{h})$ is consistent.

Theorem 6.3. For the model selection problem, under the same assumptions as those in Theorem 4.5, $\hat{\mu}_n(\hat{h})$ is consistent whenever given μ_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.

Theorem 6.4. For the ridge regression problem, under (A.1''), (6.2.6) and that

$$(6.2.7) \qquad \Big(\sum_{i=1}^{n} \lambda_{i,n}^{-2}\Big)\lambda_{n,n}^{2} \longrightarrow \infty, \quad \text{as } n \longrightarrow \infty,$$

$\hat{\mu}_n(\hat{h})$ is consistent whenever given μ_n, there exists a sequence of h_n such that $\hat{\mu}_n(h_n)$ is consistent.

Since as was shown before spline smoothing of Example 4 is a special case of ridge regression, we can use Theorem 6.4 to obtain the desired consistency results. All we need is to check (6.2.7). For instance, if x_i's are equi-spaced, then Craven and Wahba's result showed that (6.2.7) holds. Roughly we have $\lambda_{i,n} \approx ci^{-2k}$ for some constant c and hence

$$\Big(\sum_{i=1}^{n} \lambda_{i,n}^{-2}\Big)\lambda_{n,n}^{2} \approx \sum_{i=1}^{n} \Big(\frac{i}{n}\Big)^{4k} \approx n \cdot \int_0^1 x^{4k}dx \longrightarrow \infty.$$

Finally, in the model selection problem, besides C_p, there are other procedures closely related with G.C.V. . Hocking (1976) firstly proposed a criterion called S_p, which select h by minimizing $(n-1)(n-h)^{-1}(n-h-1)^{-1}$. $||y_n - \hat{\mu}_n(h)||^2$. Compared with (1.5), we see that their difference is only marginal. In fact in the proof of Theorem 6.3 (Section 8.12), we shall see that the consistency of $\tilde{\mu}_n(\hat{h})$ and $\text{SURE}_n(\hat{h})$ implies that $n-\hat{h} \to \infty$ (see (8.12.1)). Therefore S_p and G.C.V. are asymptotically equivalent under such circumstances. S_p was further studied by Thompson (1978) and Breiman and Freedman (1983). Breiman and Freedman established an asymptotic optimality for S_p in the setup of Example 2 under the assumptions that $p_n = \infty$, $H_n = \{1,2,\ldots,n/2\}$, all explanatory variables and random errors are jointly normal and that there are infinitely many non-zero β_j's. Shibata (1981)

used a different criterion: minimizing $n^{-1}(n+2h)||y_n - \hat{\mu}_n(h)||^2$. Clearly, if for all $h \in H_n$, $h << n$ in the sense that $\frac{h}{n} \to 0$, then this criterion is asymptotically equivalent to the G.C.V. . Otherwise they might be quite different. Shibata obtained an asymptotic optimality for his criterion (again in the setup of Example 2 with $p_n = \infty$) but the underlying assumption about H_n (i.e., $\max H_n = o(n)$) makes this selection procedure not completely driven by the data (the same criticism applys to the work of Breiman and Freedman but less seriously).

6.3. Utilizing NTLE.

If we accept the first viewpoint of G.C.V., then it seems reasonable to use $\bar{\mu}_n(\hat{h})$ to estimate μ_n, where $\hat{h}$ is again the minimizer of (1.3). The consistency problem of $\bar{\mu}_n(\hat{h})$ will be discussed below.

First, observe that if we replace $A_n(h) = I - M_n(h)$ in (3.1) and (3.2) by $I_n - (-\alpha I_n + (1+\alpha)M_n(h))$, then we get the same $\tilde{\mu}_n(h)$ and $\text{SURE}_n(h)$. This means that the simplified versions of Stein's unbiased estimate and unbiased risk estimate corresponding to NTLE, $\bar{\mu}_n(h)$, are the same as those corresponding to the original $\hat{\mu}_n(h)$. Thus we may pretend that $\tilde{\mu}_n(h)$ is indeed constructed from $\bar{\mu}_n(h)$. Now Lemma 6.1 amounts to saying say under (6.2.1), $\bar{\mu}_n(\hat{h})$ is consistent if and only if

$$(6.3.1) \qquad n^{-1}\,\text{tr}(I_n - (-\alpha I_n + (1+\alpha)M_n(\hat{h}))) \to 1 \quad \text{in probability.}$$

However, (6.3.1) always holds (in fact " $\to$ " is really " = ") because of the definition of NTLE. We state our conclusion in the following:

Theorem 6.3.1. Under (6.2.1), $\bar{\mu}_n(\hat{h})$ is consistent.

Therefore we have seen that from the consistency point of view, $\bar{\mu}_n(\hat{h})$ is more favorable than $\hat{\mu}_n(\hat{h})$ because no extra conditions are needed for $\bar{\mu}_n(\hat{h})$. But for sample size not too large we have to be careful since as mentioned before, the motivation of NTLE may be shaky. Another important problem remains, i.e., how to obtain an assessment for the performance of $\bar{\mu}_n(\hat{h})$ or $\hat{\mu}_n(\hat{h})$ that may have desirable properties like uniform consistency? It seems that one may use $n^{-1}||y_n - \bar{\mu}_n(\hat{h})||^2 - \sigma^2$ (or equivalently, the quantities of (1.3) $-\sigma^2$) as the error assessment for $\bar{\mu}_n(\hat{h})$ (and for $\hat{\mu}_n(\hat{h})$?) (of course, σ should be estimated). But the properties of this estimate still awaits further investigation.

7. A variant of G.C.V.

Suppose that μ_n is known to be in a proper linear subspace S_n of R^n with dimension s_n and that $\hat{\mu}_n(h)$ takes values only in S_n. Then one should not use $\tilde{\mu}_n(h)$ or $\bar{\mu}_n(h)$ since they may take values outside S_n. Natural ways to proceed in such circumstances seem to replace the raw data y_n by it's projection on S_n, denoted by y_n^* ; namely $y_n^* = P_n y_n$ where P_n denote the n x n projection matrix from R^n to S_n. The simplified version of Stein estimate (3.1) should be changed to

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

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

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

$$\text{SURE}_n^*(h) = \sigma^2 - \frac{\sigma^4(s_n - \text{tr } M_n(h))^2}{s_n||y_n^* - \hat{\mu}_n(h)||^2}$$

Here $\text{SURE}_n^*(h)$ is estimating the loss $s_n^{-1}||\mu_n - \tilde{\mu}_n^*(h)||^2$. Therefore in choosing h, one should minimize

$$\text{(7.1)} \qquad \text{GCV}_n^*(h) = \frac{s_n^{-1}||y_n^* - \hat{\mu}_n(h)||^2}{(1 - s_n^{-1} \text{ tr } M_n(h))^2} \quad ,$$

instead of (1.3).

On the other hand, $\bar{\mu}_n(h)$ should be changed to $\bar{\mu}_n^*(h) = -\alpha^* y_n^* + (1+\alpha^*)\hat{\mu}_n(h)$, with $\alpha^* = \text{tr } M_n(h) / (s_n - \text{tr } M_n(h))$. Applying C_L to $\bar{\mu}_n^*(h)$ amounts to minimizing

$$s_n^{-1}||y_n^* - \bar{\mu}_n^*(h)||^2 = (1-\alpha^*)^2 s_n^{-1}||y_n^* - \hat{\mu}_n(h)||^2 = (7.1).$$

In the problem of ridge regression or model selection, we thus have two options of G.C.V. to choose h: (i) the original one (minimizing (1.3)) which assumes that no model with rank less than n is completely appropriate, i.e. $\mu_n \in R^n$; (ii) the modified one (minimizing (7.1)) which assumes some model with rank $s_n < n$ is correct, i.e., $\mu_n \in S_n \subset R^n$. However in the case of replication, it seems that one should use (7.1); here y_n^* is just the sample average of replicated observations.

8. Proofs.

8.1. Proof of Theorem 2.1.

Observe that $n^{-1}||\bar{\mu}_n(h_n) - \hat{\mu}_n(h)||^2 = \alpha^2 n^{-1}||\hat{\mu}_n(h_n) - y_n||^2 \le$ $\alpha^2(n^{-1}||\hat{\mu}_n(h_n) - \mu_n||^2 + n^{-1}\sum_{i=1}^{n} \varepsilon_i^2)$. On the other hand, (2.1) implies

that the variance part $\sigma^2\, n^{-1}\, \mathrm{tr}\, M_n'(h_n)\, M_n(h_n)$ converges to 0. This in turn shows that $n^{-1}\, \mathrm{tr}\, M_n(h) \to 0$ since $(n^{-1}\mathrm{tr}\, M_n(h_n))^2 \le n^{-1}\, \mathrm{tr}\, M_n'(h_n)M_n(h_n)$. Now since $\alpha = n^{-1}\mathrm{tr}\, M_n(h_n) \,/\, (1 - n^{-1}\mathrm{tr}\, M_n(h_n))$, it is clear that $n^{-1}||\bar{\underset{\sim}{\mu}}_n(h_n) - \hat{\underset{\sim}{\mu}}_n(h_n)||^2 \to 0$ and hence that (2.2) holds. Other statements follow by similar arguments. □

8.2. Proof of Theorem 3.1.

In view of (3.4), it suffices to show that for any $\delta_1, \delta_2 > 0$, there exists an integer N such that when $n \ge N$,

$$(8.2.1) \qquad p\left\{ \frac{|\mathrm{tr}\, A_n|}{n} \cdot \frac{|<\underset{\sim}{\varepsilon}_n, A_n\underset{\sim}{\varepsilon}_n> - \sigma^2\, \mathrm{tr}\, A_n|}{||A_n\underset{\sim}{y}||^2} \ge \delta_1 \right\} \le \delta_2 ,$$

and

$$(8.2.2) \qquad p\left\{ \frac{|\mathrm{tr}\, A_n|}{n} \cdot \frac{|<\underset{\sim}{\varepsilon}_n, A_n\underset{\sim}{\mu}_n> |}{||A_n y||^2} \ge \delta_1 \right\} \le \delta_2 .$$

The following two lemmas will be useful. Recall that $\lambda'(A_n)$ denotes the maximum singular value of A_n.

Lemma 8.2.1. Assume that $\lambda'(A_n) = 1$. Then under (A.1) we have

$$\mathrm{Var} < \underset{\sim}{\varepsilon}_n, A_n\, \underset{\sim}{\mu}_n > \; = \sigma^2 ||A_n\, \underset{\sim}{\mu}_n||^2,$$

$$\mathrm{Var} < \underset{\sim}{\varepsilon}_n, A_n\, \underset{\sim}{\varepsilon}_n > \le m\, \mathrm{tr}\, A_n' A_n,$$

and

$$\mathrm{Var}\; ||A_n\, \underset{\sim}{y}_n||^2 \le 2m\, \mathrm{tr}\, A_n' A_n + 8\sigma^2 ||A_n\, \underset{\sim}{\mu}_n||^2.$$

Lemma 8.2.2. Suppose for any n, $\underset{\sim}{c}_n$ is a vector in R^n with $||\underset{\sim}{c}_n|| = 1$.

Then under (A.2) for any sequence of nonnegative numbers $\{a_n\}$ converging to 0, and any sequence of real numbers $\{b_n\}$, we have

$$\lim_{n\to\infty} p\{|c_n'\varepsilon_n + b_n| \le a_n\} = 0.$$

Assuming the validity of these lemmas, we proceed with the proof of Theorem 3.1. Without loss of generalty, we assume $\lambda'(A_n) = 1$. (8.2.1) and (8.2.2) will hold if there exists a positive number a_n such that

$$(8.2.3) \qquad p\{||A_n y||^2 \le a_n(||A_n\mu_n||^2 + \sigma^2 \operatorname{tr} A_n' A_n)\} \le \frac{\delta_2}{2},$$

$$(8.2.4) \qquad p\{\frac{|\operatorname{tr} A_n|}{n} \cdot |<\varepsilon_n, A_n\varepsilon_n> - \sigma^2 \operatorname{tr} A_n| \ge \delta_1 a_n \cdot$$

$$(||A_n\mu_n||^2 + \sigma^2 \operatorname{tr} A_n'A_n)\} \le \frac{\delta_2}{2},$$

and

$$(8.2.5) \qquad p\{\frac{|\operatorname{tr} A_n|}{n} \cdot |<\varepsilon_n, A_n\mu_n>| \ge \delta_1 a_n(||A_n\mu_n||^2 + \sigma^2 \operatorname{tr} A_n'A_n)\} \le \frac{\delta_2}{2}.$$

Using Chebychev inequality, the left hand side of (8.2.4) is no greater than

$$\left(\frac{\operatorname{tr} A_n}{n}\right)^2 \cdot \operatorname{Var} <\varepsilon_n, A_n\varepsilon_n> / \delta_1^2 a_n^2(||A_n\mu_n||^2 + \sigma^2 \operatorname{tr} A_n'A_n)^2 .$$

Now by Lemma 8.2.1. and the fact that $\left(\frac{\operatorname{tr} A_n}{n}\right)^2 \le \frac{\operatorname{tr} A_n' A_n}{n}$,

the above term does not exceed $n^{-1}a_n^{-2} \cdot m\sigma^{-2}\delta_1^{-2}$.

Take $c = (2m\,\sigma^{-2}\delta_1^{-2}\delta_2^{-1})^{1/2}$. We see that (8.2.4) holds if

(8.2.6) $\qquad a_n \geq cn^{-1/2}$

By a similar argument, we can show that (8.2.5) holds if

(8.2.7) $\qquad a_n \geq c' n^{-1/2}$

with $c' = (2\ \delta_1^{-2}\ \delta_2^{-1}\ )^{1/2}$. Now to get (8.2.3), we first set

(8.2.8) $\qquad a_n < 1/2$.

Then by Chebychev inequality again, the left hand side of (8.2.3) does not exceed var $||A_n\underset{\sim}{y}||^2$ / $(1-a_n)^2(||A_n\underset{\sim}{\mu}_n||^2 + \sigma^2 \text{tr } A_n'A_n)^2$. We can bound this quantity by $(8m\ \sigma^{-2} + 32)\ (||A_n\underset{\sim}{\mu}_n||^2 + \sigma^2\ \text{tr}\ A_n'A_n)^{-1}$ after utilizing Lemma 8.2.1 and some simple computation. Put $c'' = (16m\ \sigma^{-2} + 64)\ \delta_2^{-1}$. Then (8.2.3) holds if

(8.2.9) $\qquad ||A_n\underset{\sim}{\mu}_n||^2 + \sigma^2\ \text{tr}\ A_n'\ A_n \geq c''$.

It remains to take care of those n such that (8.2.9) does not hold. Let $\{n'\}$ denote the subsequence of $\{n\}$ such that (8.2.9) fails. Suppose $\{n'\}$ is a finite sequence. Then for n large enough, (8.2.9) holds and so do (8.2.6) $\sim$ (8.2.8) with, say, $a_n = \frac{1}{4}$. Consequently (8.2.1) and (8.2.2) are proved and Theorem 3.1 follows. Thus it remains to consider the case that n' is infinite.

Recall that we have assumed $\lambda'(A_n) = 1$. It follows that $A_n'A_n \geq \underset{\sim}{c}_n'\underset{\sim}{c}_n$ in the nonnegative definite sense where $\underset{\sim}{c}_n$ is an eigenvector with eigenvalue 1 for $A_n'A_n$ and $||\underset{\sim}{c}_n|| = 1$. Therefore, with $n = n'$, the left hand side of (8.2.3) does not exceed

$$p\{|c_{n'}'\varepsilon_{n'} + c_{n'}'\mu_{n'}| \le a_{n'}^{1/2}(||A_{n'}\mu_{n'}||^2 + \sigma^2 \operatorname{tr} A_{n'}'A_{n'})^{1/2}\}$$

$$\le p\{|c_{n'}'\varepsilon_{n'} + c_{n'}'\mu_{n'}| \le a_{n'}^{1/2} c''\}.$$

Now by Lemma 8.2.2., this last quantity can be made arbitrary small for n' large enough providing that we have set

(8.2.10) $\qquad a_n \to 0.$

Hence with (8.2.10) and (8.2.8) we have shown that (8.2.3) holds for n large enough. Finally observe that there exists $\{a_n\}$ such that (8.2.6) $\sim$ (8.2.8) and (8.2.10) hold for n large enough. This completes the proofs of (8.2.3) $\sim$ (8.2.5). Theorem 3.1 is established. □

Proof of Lemma 8.2.1. This is straightforward. One may have to use the inequality that $(A_n' A_n)^2 \le A_n'A_n$ which is implied by the assumption that $\lambda'(A_n) = 1$. The details will be omitted.

Proof of Lemma 8.2.2.

Write $c_n = (c_{1n},\ldots,c_{nn})'$. Without loss of generality, assume that $c_{1n} = \max_{1\le i\le n} c_{in}$. Given $\delta > 0$, we want to show that for large n, the probability of the event

(8.2.11) $\qquad \{|c_n'\varepsilon_n + b_n| \le a_n\}$

is no greater than δ. Rewrite (8.2.11) as

$$\{|\varepsilon_1 + c_{1n}^{-1}(\sum_{i=2}^{n} c_{in}\varepsilon_i + b_n)| \le a_n c_{1n}^{-1}\}$$

and consider the conditional probability that this event will happen given $\varepsilon_2, \varepsilon_3,\ldots, \varepsilon_n$. By (A.2) we see that this conditional probablity

does not exceed $k\, a_n c_{1n}$. Hence the unconditional probability of event (8.2.11) is no greater than $k\, a_n c_{1n}$. Therefore if for any n, we have

$$c_{1n} \geq k\, a_n \delta^{-1} \tag{8.2.12}$$

then the probability of event (8.2.11) does not exceed δ, as desired. Now, consider the subsequence $\{n'\}$ of $\{n\}$ for which (8.2.12) fails. Along this sequence, $c_{1n'} \to 0$ because $a_n \to 0$. Then it can be shown that $\underset{\sim}{c}'_{n'}, \underset{\sim}{\varepsilon}_{n'}$ is asymptotically normal with mean 0 and variance 1 by checking that the Linderberg-Feller condition is satisfied. Now it becomes trivial to show that for large n' the probability of event (8.2.11) (with n setting to n') does not exceed δ. This completes the proof of Lemma 8.2.2. □

8.3. Proof of Theorem 4.3.

We shall use the following lemma whose proof is detered.

Lemma 8.3.1. Assume (A.1') holds. Then there exist constants c' and c'' (depending on M only) such that

$$E(<\underset{\sim}{\varepsilon}_n, A\,\underset{\sim}{\varepsilon}_n> - \sigma^2 \operatorname{tr} A)^4 \leq c'(\operatorname{tr} A'A)^2\,, \tag{8.3.1}$$

$$E(<\underset{\sim}{\varepsilon}_n, A\underset{\sim}{\mu}_n>)^4 \leq m||A\underset{\sim}{\mu}_n||^4 \tag{8.3.2}$$

and

$$E(||A\underset{\sim}{y}_n||^2 - ||A\underset{\sim}{\mu}_n||^2 - \sigma^2 \operatorname{tr} A'A)^4 \leq c''(\lambda'(A))^4[\sigma^4(\operatorname{tr} A'A)^2 + ||A\underset{\sim}{\mu}_n||^4] \tag{8.3.3}$$

for any n x n matrix A and any $\underset{\sim}{\mu}_n \in R^n$.

To prove Theorem 4.3, first in view of (3.4), (8.2.3) ∿ (8.2.5), it suffices to show that

(8.3.4) $$C_1 n^{-2} \#H_n + C_2 \sum_{h\in H_n} [\lambda''(A_n(h))]^4 [\mathrm{tr}\, A_n'(h)\, A_n(h)]^{-2}$$

is no less than

(8.3.5) $$p\{ \text{for some } h\in H_n: \frac{2\sigma^2}{n} \mid \mathrm{tr}\, A_n(h) \mid \cdot \mid < \underset{\sim}{\varepsilon}_n, A_n(h)\underset{\sim}{\varepsilon}_n > - \sigma^2 \mathrm{tr}\, A_n(h) \mid \geq \frac{\delta}{4} \cdot (||A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \mathrm{tr}\,(A_n'(h)A_n(h)))\} +$$

$$p\{ \text{for some } h\in H_n: \frac{2\sigma^2}{n} \mid \mathrm{tr}\, A_n(h) \mid \cdot < \underset{\sim}{\varepsilon}_n, A_n(h)\underset{\sim}{\mu}_n > \mid \geq \frac{\delta}{4} (||A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \mathrm{tr}(A_n'(h)A_n(h))\} +$$

$$p\{ \text{for some } h\in H_n: ||A_n(h)\underset{\sim}{y}_n||^2 \leq 1/2\, (||A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \mathrm{tr}(A_n'(h)\, A_n(h))\}\,.$$

Clearly, the above expression does not exceed

$$\sum_{h\in H_n} p\{ \frac{16\sigma^8}{n^4} (\mathrm{tr}\, A_n(h))^4 (< \underset{\sim}{\varepsilon}_n, A_n(h)\underset{\sim}{\varepsilon}_n > - \sigma^2 \mathrm{tr}\, A_n(h))^4 \geq (\frac{\delta}{4})^4 \cdot (||A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \mathrm{tr}(A_n'(h)A_n(h)))^4\}$$

$$+ \sum_{h\in H_n} p\{ \frac{16\sigma^8}{n^4} (\mathrm{tr}\, A_n(h))^4 (< \underset{\sim}{\varepsilon}_n, A_n(h)\underset{\sim}{\mu}_n >)^4 \geq (\frac{\delta}{4})^4 \cdot (||A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \mathrm{tr}(A_n'(h)A_n(h)))^4\}$$

$$+ \sum_{h\in H_n} p\{ [||A_n(h)\underset{\sim}{y}_n||^2 - (||(A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \mathrm{tr}(A_n'(h)A_n(h))]^4 \geq \frac{1}{16} \cdot$$

$$(||A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \operatorname{tr}(A_n'(h)A_n(h)))^4\} .$$

Now, using Chebychev inequality and Lemma 8.3.1, we can bound the above expression by

$$\sum_{h\in H_n} 16^3\ \sigma^8\ \delta^{-4}\ n^{-4}(\operatorname{tr} A_n(h))^4 c'(\operatorname{tr}A_n'(h)\ A_n(h))^2(||A_n(h)\underset{\sim}{\mu}_n||^2 +$$

$$\sigma^2 \operatorname{tr}(A_n'(h)A_n(h)))^{-4}$$

$$+ \sum_{h\in H_n} 16^3\ \sigma^8\ \delta^{-4}n^{-4}(\operatorname{tr} A_n(h))^4\ \sigma^4||A_n(h)\underset{\sim}{\mu}_n||^4(||A_n(h)\underset{\sim}{\mu}_n||^2 +$$

$$\sigma^2 \operatorname{tr}(A_n'(h)A_n(h)))^{-4}$$

$$+ \sum_{h\in H_n} 16\ \ c''[\lambda'(A_n(h))]^4\ [\sigma^4(\operatorname{tr} A_n'(h)A_n(h))^2 +$$

$$||A_n(h)\underset{\sim}{\mu}_n||^4] \cdot (||A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \operatorname{tr}(A_n'(h)A_n(h)))^{-4} .$$

Deleting $||A_n(h)\underset{\sim}{\mu}_n||^2$ from the first summation term and utilizing the inequality $n^{-2}(\operatorname{tr} A_n(h))^{-2} \leq n^{-1}\operatorname{tr} A_n'(h)A_n(h)$, we see that the first summation is no greater than $16^3\ \delta^{-4}c'n^{-2}\#H_n$. A similar argument applying to the second summation yields the bound $16^3\ \sigma^8\ \delta^{-4}n^{-2}\#H_n$. Finally, it is clear that the third term does not exceed $16c''\sigma^{-4} \sum_{h\in H_n} [\lambda(A_n(h))]^4 \cdot$ $(\operatorname{tr}A_n'(h)A_n(h))^{-2}$. Now put $C_1 = 16^3\ \delta^{-4}(c' + \sigma^8)$ and $C_2 = 16c''\ \sigma^{-4}$ to get (8.3.4), the desired upper bound. The proof of Theorem 4.3 is complete.

□

Proof of Lemma 8.3.1.

Proof of (8.3.1). Since tr $(\frac{A+A'}{2})^2 \leq$ tr A'A, we may replace A by (A+A')/2 when A is asymmetric. Thus without loss of generality we may assume A is symmetric. In addition, we may assume $\lambda'(A) = 1$ because dividing both sides of (8.3.1) by a constant $(\lambda'(A))^4$ does not change the inequality. Let $A = (a_{ij})$. Then $E\ (< \underset{\sim}{\varepsilon}_n A \underset{\sim}{\varepsilon}_n > - \sigma^2 \text{tr } A)^4 = E(\sum_{i \neq j} a_{ij}\varepsilon_i\varepsilon_j + \sum_i a_{ii}(\varepsilon_i^2 - \sigma^2))^4 \equiv E(I + II)^4 = EI^4 + 4EI^3II + 6EI^2II^2 + 4EI\ II^3 + EII^4$.

The term EI^4 is the sum of items of the form $a_{i_1j_1}a_{i_2j_2}a_{i_3j_3}a_{i_4j_4}E\varepsilon_{i_1}\varepsilon_{j_1}\varepsilon_{i_2}\varepsilon_{j_2}\cdot \varepsilon_{i_3}\varepsilon_{j_3}\varepsilon_{i_4}\varepsilon_{j_4}$ with $i_1 \neq j_1$, $i_2 \neq j_2$, $i_3 \neq j_3$ and $i_4 \neq j_4$. Since $E\varepsilon_i = 0$, some of these items vanish automatically. Let α, β, γ, δ denote four different integers between 1 and n. Then what reamin in EI^4 are items with label $\{\{i_1,j_1\}, \{i_2,j_2\}, \{i_3,j_3\}, \{i_4,j_4\}\}$ taking one of the following forms or their permutation versions:

(i) $\{\{\alpha,\beta\},\{\alpha,\beta\},\{\gamma,\delta\},\{\gamma,\delta\}\}$,

(ii) $\{\{\alpha,\beta\},\{\beta,\gamma\},\{\gamma,\delta\},\{\delta,\alpha\}\}$,

(iii) $\{\{\alpha,\beta\},\{\alpha,\beta\},\{\beta,\gamma\},\{\gamma,\alpha\}\}$,

(iv) $\{\{\alpha,\beta\},\{\alpha,\beta\},\{\alpha,\gamma\},\{\alpha,\gamma\}\}$,

(v) $\{\{\alpha,\beta\},\{\alpha,\beta\},\{\alpha,\beta\},\{\alpha,\beta\}\}$.

The sum of items of form (i) does not exceed $M(\sum_{\alpha,\beta} a_{\alpha\beta}^2)(\sum_{\gamma,\delta} a_{\gamma\delta}^2)$, which in turn is no greater than $M(\text{tr } A^2)^2$. Denote $A^2 = (b_{ij})$, $A^3 = (c_{ij})$ and $A^4 = (d_{ij})$. The sum of items of form (ii) does not exceed

$$M \sum_{\alpha,\beta,\gamma,\delta} a_{\alpha\beta}\, a_{\beta\gamma}\, a_{\gamma\delta}\, a_{\delta\alpha} = M \sum_{\alpha,\beta,\gamma} a_{\alpha\beta}\, a_{\beta\gamma}\, [b_{\gamma\alpha} - a_{\gamma\alpha}\, a_{\alpha\alpha} - a_{\gamma\beta}\, a_{\beta\alpha} -$$

$$a_{\gamma\gamma} a_{\gamma\alpha}) = M \sum_{\alpha,\beta} a_{\alpha\beta}[(c_{\beta\alpha} - a_{\beta\alpha} b_{\alpha\alpha} - a_{\beta\beta} b_{\beta\alpha}) - (b_{\beta\alpha} - a_{\beta\alpha} a_{\alpha\alpha} - a_{\beta\beta} a_{\beta\alpha}) a_{\alpha\alpha} -$$

$$(b_{\beta\beta} - a_{\beta\alpha}^2 - a_{\beta\beta}^2) a_{\beta\alpha}] - M \sum_{\alpha,\beta,\gamma} a_{\alpha\beta} a_{\beta\gamma} a_{\gamma\gamma} a_{\gamma\alpha}. \quad \text{Now} \quad \sum_{\alpha,\beta} a_{\alpha\beta} c_{\beta\alpha} =$$

$\sum_{\alpha} d_{\alpha\alpha} - \sum_{\alpha} a_{\alpha\alpha} c_{\alpha\alpha} = \mathrm{tr} A^4 - \sum_{\alpha} a_{\alpha\alpha} c_{\alpha\alpha}$ does not exceed $(\mathrm{tr}\, A^2)^2 + (\sum_{\alpha} a_{\alpha\alpha}^2)^{1/2}$.

$(\sum_{\alpha} c_{\alpha\alpha}^2)^{1/2} \leq (\mathrm{tr}\, A^2)^2 + (\mathrm{tr}\, A^2)^{1/2} (\mathrm{tr}\, (A^3)^2)^{1/2} \leq (\mathrm{tr}\, A^2)^2 + \mathrm{tr}\, A^2$ (the last inequality is due to the assumption that $\lambda'(A) = 1$ which implies that $A^6 \leq A^2$). Now since "$\lambda'(A) = 1$" again implies "$\mathrm{tr}\, A^2 \geq 1$", we conclude that $\sum_{\alpha,\beta} a_{\alpha\beta} c_{\beta\alpha}$ $\leq 2(\mathrm{tr}\, A^2)^2$. Similar arguments can be applied to all other summations including those involved in the sum of forms (ii) $\sim$ (v) and those in the evaluation of EI^3II, EI^2II^2, etc. The details are omitted.

Proof of (8.3.2). This is straightforward.

Proof of (8.3.3). Observe that $E(||A\underset{\sim}{y}_n||^2 - ||A\underset{\sim}{\mu}_n||^2 - \sigma^2 \mathrm{tr}\, A'A)^4$

$= E(< \underset{\sim}{\varepsilon}_n, A'A\underset{\sim}{\varepsilon}_n > - \sigma^2 \mathrm{tr}\, A'A + 2 < \underset{\sim}{\varepsilon}_n, A'A\underset{\sim}{\mu}_n >)^4 \leq 8\, \{E(< \underset{\sim}{\varepsilon}_n, A'A\underset{\sim}{\varepsilon}_n > - \sigma^2 \mathrm{tr}\, A'A)^4$

$+ 16\, E < \underset{\sim}{\varepsilon}_n, A'A\underset{\sim}{\mu}_n >^4 \}$. Now by (8.3.1) and (8.3.2) the last expression does not exceed $8\{c'(\mathrm{tr}(A'A)^2)^2 + 16m||A'A\underset{\sim}{\mu}_n||^4\}$ which in turn does not exceed $8[\lambda(A'A)]^2\{c'(\mathrm{tr}\, A'A)^2 + 16m\, ||A\underset{\sim}{\mu}_n||^2\}$. Finally, since $\lambda(A'A) = (\lambda'(A))^2$, we may take $c'' = 8 \max\{c'\, \sigma^{-4}, 16m\}$ to conclude (8.3.3).

8.4. Proof of Lemma 4.1.

Observe that $||M_n(h)\underset{\sim}{y}_n||^2 = \sum_{i=1}^{n} (\sum_{j=1}^{h} w_{n,h}(j) y_{i(j)})^2 \leq \sum_{i=1}^{n} \sum_{j=1}^{h} w_{n,h}(j) y_{i(j)}^2$

$= \sum_{j=1}^{h} w_{n,h}(j) \sum_{j=1}^{n} y_{i(j)}^2 = \sum_{j=1}^{h} (w_{n,h}(j) - w_{n,h}(j+1)) (\sum_{k=1}^{j} \sum_{i=1}^{n} y_{i(k)}^2)$, where

$w_{n,h}(h+1)$ is set to be 0. For $1 \leq \ell \leq n$, $2 \leq j \leq n$, let $\eta(\ell,j)$ denote the cardinal number of the set $\bigcup_{k=2}^{j} \{i: i(k) = \ell\}$. It is clear that

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

Now Lemma 2.2 of Li (1982) showed that there exists a universal constant λ_5 (depending only on the dimension p) such that $\eta(\ell,j) \leq \lambda_5(j-1)$ for any n, ℓ, j. Therefore we see that $||M_n(h)\underset{\sim}{y}_n||^2 \leq \sum_{j=1}^{h} (w_{n,h}(j) - w_{n,h}(j+1)) \lambda_5 j$.

$(\sum_{\ell=1}^{n} y_\ell^2) = \lambda_5 \sum_{i=1}^{n} y_\ell^2$. This means that $\lambda(M_n'(h)M_n(h)) \leq \lambda_5$. Hence $\lambda'(M_n(h)) \leq \lambda_5^{1/2}$. Taking $\Lambda = \lambda_5^{1/2}$, the proof of Lemma 4.1 is complete.

□

8.5. Proof of Theorem 4.5.

First, applying Theorem 4.2, we see that

$$\lim_{n\to\infty} \sup_{\underset{\sim}{\mu}_n \in R^n} p\{ \sup_{h \in H_n'} |SURE_n(h) - \frac{1}{n}||\tilde{\underset{\sim}{\mu}}_n(h) - \underset{\sim}{\mu}_n||^2| \geq 2\delta \} = 0.$$

Next, (4.6) implies that

$$\lim_{n\to\infty} \sup_{\underset{\sim}{\mu}_n \in R^n} p\{ \sup_{h \in H_n - H_n'} | \mathrm{SURE}_n(h) - \frac{1}{n} ||\tilde{\underset{\sim}{\mu}}_n(h) - \underset{\sim}{\mu}_n||^2 | \geq 2\delta \} \leq \varepsilon.$$

Combining these two statements, we have

$$\lim_{n\to\infty} \sup_{\underset{\sim}{\mu}_n \in R^n} p\{ \sup_{h \in H_n} | \mathrm{SURE}_n(h) - \frac{1}{n} ||\tilde{\underset{\sim}{\mu}}_n(h) - \underset{\sim}{\mu}_n||^2 | \geq 2\delta \} \leq \varepsilon.$$

Finally, we may take $\varepsilon \to 0$ to complete the proof of Theorem 4.5. □

8.6 Proof of Lemma 4.2.

Proceeding as in the beginning of the proof of Theorem 3.1, it suffices to show that for any $\delta_1, \delta_2 > 0$, there exists $a_n \to 0$, such that for n large,

$$(8.6.1) \qquad p\left\{ \inf_{h \geq 0} \frac{\sum_{i=1}^n (\mu_i + \varepsilon_i)^2 (h + \lambda_i)^{-2}}{\sum_{i=1}^n (\mu_i^2 + \sigma^2)(h + \lambda_i)^{-2}} \leq a_n \right\} \leq \frac{\delta_2}{2},$$

$$(8.6.2) \qquad p\left\{ \sup_{h \geq 0} \frac{\sum_{i=1}^n (h + \lambda_i)^{-1} \cdot \left| \sum_{i=1}^n (h + \lambda_i)^{-1} (\varepsilon_i^2 - \sigma^2) \right|}{n \sum_{i=1}^n (h + \lambda_i)^{-2} (\mu_i^2 + \sigma^2)} \geq a_n \delta_1 \right\} \leq \frac{\delta_2}{2},$$

$$(8.6.3) \qquad p\left\{ \sup_{h \geq 0} \frac{\sum_{i=1}^n (h + \lambda_i)^{-1} \cdot \left| \sum_{i=1}^n (h + \lambda_i)^{-1} \mu_i \varepsilon_i \right|}{n \sum_{i=1}^n (h + \lambda_i)^{-2} (\mu_i^2 + \sigma^2)} \geq a_n \delta_1 \right\} \leq \frac{\delta_2}{2}.$$

Note that (8.6.1) ∿ (8.6.3) correspond to (8.2.3) ∿ (8.2.5) respectively.

Proof of (8.6.1). First observe that as a special case of Lemma 8.2.1,

(8.6.4) $\quad \mathrm{var}(\mu_i + \varepsilon_i)^2 \le c(\mu_i^2 + \sigma^2)$

for $c = \max\{8\sigma^2, 2m/\sigma^2\}$.

Define $G_n(h) = \sum_{i=1}^{n} (\mu_i^2 + \sigma^2)(h + \lambda_n)^2 / (h + \lambda_i)^2$. Obviously, $G_n(h)$ is a continuous increasing function of h. Let L be a large number to be chosen later. Let ℓ_n be the largest integer such that $2^{\ell_n} L < G_n(0)$, and k_n be the largest integer such that $2^{k_n} L \le G_n(\infty)$. For i, $\ell_n + 1 \le i \le k_n$, define $h_i^{(n)} = G_n^{-1}(2^i L)$. Take $h_{\ell_n}^{(n)} = 0$ and $h_{k_n+1}^{(n)} = \infty$.

First, consider the case that $\ell_n \ge 1$. Write the left hand side of (8.6.1) as

$$p\left\{ \inf_{h\ge 0} \frac{\sum_{i=1}^{n} (\mu_i + \varepsilon_i)^2 (h + \lambda_n)^2 / (h + \lambda_i)^2}{G_n(h)} \le a_n \right\},$$

which does not exceed

$$\sum_{j=\ell_n}^{k_n} p\left\{ \inf_{h_j^{(n)} \le h \le h_{j+1}^{(n)}} \frac{\sum_{i=1}^{n} (\mu_i+\varepsilon_i)^2 (h+\lambda_n)^2/(h+\lambda_i)^2}{G_n(h)} \le a_n \right\}$$

$$\le \sum_{j=\ell_n}^{k_n} p\left\{ \frac{G_n(h_j^{(n)})}{G_n(h_{j+1}^{(n)})} \cdot \frac{\sum_{i=1}^{n} (\mu_i+\varepsilon_i)^2 (h_j^{(n)}+\lambda_n)^2/(h_j^{(n)}+\lambda_i)^2}{G_n(h_j^{(n)})} \le a_n \right\}$$

$$\le \sum_{j=\ell_n}^{k_n} p\left\{ \left| \frac{\sum_{i=1}^{n} (\mu_i+\varepsilon_i)^2 (h_j^{(n)}+\lambda_n)^2 / (h_j^{(n)}+\lambda_i)^2}{G_n(h_j^{(n)})} - 1 \right| \ge |1 - 2a_n| \right\}.$$

Here by the fact that $G_n(h_j^{(n)})/G_n(h_{j+1}^{(n)}) \le 2^{-1}$, the last inequality holds for n large so that $2a_n < 1$. Now by the definition of $G_n(h)$ and the Chebychev

inequality, the last expression does not exceed

$$(1-2a_n)^{-2} \sum_{j=\ell_n}^{k_n} \text{var} \left\{ \frac{\sum_{i=1}^{n} (\mu_i+\varepsilon_i)^2 (h_j^{(n)}+\lambda_n)^2/(h_j^{(n)}+\lambda_i)^2}{G_n(h_j^{(n)})} \right\}$$

$$\leq (1-2a_n)^{-2}\, c \sum_{j=\ell_n}^{k_n} \left\{ \frac{\sum_{i=1}^{n} (\mu_i^2+\sigma^2)(h_j^{(n)}+\lambda_n)^4/(h_j^{(n)}+\lambda_i)^4}{G_n(h_j^{(n)})^2} \right\} \text{ by (8.6.4))}$$

$$\leq (1-2a_n)^{-2}\, c \sum_{j=\ell_n}^{k_n} G_n(h_j^{(n)})^{-1} \qquad (\text{since } (h_j^{(n)}+\lambda_n)(h_j^{(n)}+\lambda_i)^{-1} \leq 1)$$

$$= (1-2a_n)^{-2} \cdot c\, L^{-1} \sum_{j=\ell_n}^{k_n} 2^{-j}$$

$$\leq (1-2a_n)^{-2} \cdot c\, L^{-1}.$$

Now, since $a_n \to 0$, one can easily set L suitably (e.g. $L = 4c\, \delta_2^{-1}$) so that the last expression does not exceed $\delta_2/2$ for n large enough. Hence (8.6.1) follows.

We turn to the case that $\ell_n \leq 0$. The left hand side of (8.6.1) is no greater than

$$p \left\{ \inf_{0 \leq h \leq h_1^{(n)}} \frac{\sum_{i=1}^{n} (\mu_i+\varepsilon_i)^2 (h+\lambda_n)^2/(h+\lambda_i)^2}{L} \leq a_n \right\}$$

$$+ \sum_{j=1}^{k_n} p \left\{ \inf_{h_j^{(n)} \leq h \leq h_{j+1}^{(n)}} \frac{\sum_{i=1}^{n} (\mu_i+\varepsilon_i)^2 (h+\lambda_n)^2/(h+\lambda_i)^2}{G_n(h)} \leq a_n \right\}$$

We have shown that the second term is no greater than $(1-2a)^{-2}\, cL^{-1}$.

On the other hand clearly the first term does not exceed

$$p\{ (\mu_n+\varepsilon_n)^2 L^{-1} \le a_n\} = p\{ |\mu_n+\varepsilon_n| \le a_n^{1/2} L^{1/2}\} \le kL^{1/2} a_n^{1/2}.$$

Here the last inequality is due to (A.2). Now set L suitably so that for large n, $kL^{1/2}a_n^{1/2} + 2cL^{-1} \le \delta_2/2$ (e.g. take $L = 8c\delta_2^{-1}$). This establishes (8.6.1). □

Before turning to the proof of (8.6.2), we state a useful lemma.

Lemma 8.6.1. Assume that $w_i, i=1,\ldots,n$ are independent random variables having means 0 and finite second moments. Then for any $\delta > 0$,

$$(8.6.5) \qquad p\{ \sup_{0\le c_1\le c_2\le\ldots\le c_n\le 1} | \sum_{i=1}^n c_i w_i| \ge \delta\} \le \delta^{-2} \sum_{i=1}^n Ew_i^2.$$

If in addition w_i's have finite fourth moments, then for any $\delta > 0$,

$$(8.6.6) \qquad p\{ \sup_{0\le c_1\le c_2\le\ldots\le c_n\le 1} | \sum_{i=1}^n c_i w_i | \ge \delta\} \le \delta^{-4} E(\sum_{i=1}^n w_i)^4.$$

This lemma follows immediately from Kolomogorov's inequality and its extension (see, e.g. Chung (1974), page), after observing that $\sup_{0\le c_1\le\ldots\le c_n\le 1} |\sum_{i=1}^n c_i w_i| = \sup_{1\le j\le n} |\sum_{i=j}^n w_i|$. (8.6.5) had been used in Speckman (1981a,1982).

Proof of (8.6.2).

Using the inequality that $\sum_{i=1}^n (h+\lambda_i)^{-1} \le \sqrt{n}\,(\sum_{i=1}^n (h+\lambda_i)^{-2})^{1/2}$ (this follows from Cauchy-Schwartz inequality), we bound the left hand side of (8.6.2) by

$$p\left\{\sup_{h\geq 0}\ \frac{\left|\sum_{i=1}^{n}(\varepsilon_i^2-\sigma^2)/(h+\lambda_i)\right|}{n^{1/2}\left(\sum_{i=1}^{n}(h+\lambda_i)^{-2}\right)^{1/2}}\geq a_n\delta_1\sigma^2\right\}$$

which equals

$$(8.6.7)\quad p\left\{\sup_{h\geq 0}\ \frac{\left|\sum_{i=1}^{n}(\varepsilon_i^2-\sigma^2)(h+\lambda_n)/(h+\lambda_i)\right|}{n^{1/2}\left(\sum_{i=1}^{n}(h+\lambda_n)^2/(h+\lambda_i)^2\right)^{1/2}}\geq a_n\delta_1\sigma^2\right\}.$$

Define $F_n(h)=\sum_{i=1}^{n}(h+\lambda_n)^2/(h+\lambda_i)^2$. Then $F_n(h)$ is increasing with $F_n(\infty)=n$, and $F_n(0)\geq 1$. Let b_n be a positive number to be suitably chosen later. Now (8.6.7) is no greater than

$$(8.6.8)\quad p\left\{\sup_{h\geq b_n}\ \frac{\left|\sum_{i=1}^{n}(\varepsilon_i^2-\sigma^2)(h+\lambda_n)/(h+\lambda_i)\right|}{n^{1/2}(F_n(h))^{1/2}}\geq a_n\delta_1\sigma^2\right\}$$

$$+\ p\left\{\sup_{0\leq h\leq b_n}\ \frac{\left|\sum_{i=1}^{n}(\varepsilon_i^2-\sigma^2)(h+\lambda_n)/(h+\lambda_i)\right|}{n^{1/2}(F_n(h))^{1/2}}\geq a_n\delta_1\sigma^2\right\}.$$

Observe that $(h+\lambda_n)/(h+\lambda_i)\leq 1$ and $(h+\lambda_n)/(h+\lambda_i)\leq(h+\lambda_n)/(h+\lambda_{i+1})$.

Hence the first term of (8.6.8) does not exceed

$$p\left\{\sup_{0\leq c_1\leq c_2\leq\ldots\leq c_n}\left|\sum_{i=1}^{n}c_i(\varepsilon_i^2-\sigma^2)\right|\geq n^{1/2}F_n(b_n)^{1/2}a_n\delta_1\sigma^2\right\},$$

which by (8.6.5) is no greater than

(8.6.9) $$\frac{m}{F_n(b_n)\, a_n^2\, \delta_1^2\, \sigma^4} \quad .$$

The second term of (8.6.8) is evaluated as follows.

Let Δ_n be a small positive number to be suitably chosen later. Define k_n to be the largest integer such that $1 \le k_n \le n$ and $(b_n+\lambda_n)/(b_n+\lambda_{k_n}) \le \Delta_n$. It follows that $F_n(b_n) \ge \sum_{i=k_n+1}^{n} (b_n+\lambda_n)^2/(b_n+\lambda_i)^2 \ge (n-k_n)\,\Delta_n^2$. Hence we have

(8.6.10) $$n - k_n \le F_n(b_n)\Delta_n^{-2} \quad .$$

Now returning to (8.6.8), since $F_n(h) \ge 1$, the second term does not exceed

$$p\left\{ \sup_{0\le h\le b_n} \left| \sum_{i=1}^{k_n} (\varepsilon_i^2 - \sigma^2)(h+\lambda_n)/(h+\lambda_i)\Delta_n \right| \ge n^{1/2} a_n \delta_1 \sigma^2/2\Delta_n \right\}$$

$$+ p\left\{ \sup_{0\le h\le b_n} \left| \sum_{i=k_n+1}^{n} (\varepsilon_i^2 - \sigma^2)(h+\lambda_n)/(h+\lambda_i) \right| \ge n^{1/2} a_n \delta_1 \sigma^2 \right\}$$

Taking $c_i = (h+\lambda_n)/(h+\lambda_i)\Delta_n$ and applying (8.6.5), the first term of the above expression does not exceed

(8.6.11) $$\frac{k_n \cdot m \cdot 4\Delta_n^2}{n a_n^2 \delta_1^2 \sigma^4} \le \frac{4\Delta_n^2 m}{a_n^2 \delta_1^2 \sigma^4} \quad .$$

Similarly, the second term does not exceed

$$\frac{(n - k_n)m}{n\, a_n^2 \delta_1^2 \sigma^4}$$

which by (8.6.10) is no greater than

$$(8.6.12) \qquad \frac{F_n(b_n)m}{\Delta_n^2 \, n a_n^2 \delta_1^2 \sigma^4} \quad .$$

Combining (8.6.9), (8.6.11), (8.6.12), we see that the left hand side of (8.6.2) does not exceed

$$(8.6.13) \qquad (F_n(b_n)^{-1} + 4\Delta_n^2 + F_n(b_n)\Delta_n^{-2} n^{-1})m \, a_n^{-2}\delta_1^{-2}\sigma^4 \quad .$$

The minimum of the above expression over $\Delta_n \geq 0$, $F_n(b_n) \geq 0$ is achieved at

$$(8.6.14) \qquad \Delta_n = \frac{1}{4} n^{-\frac{1}{6}} \quad ,$$

and

$$(8.6.15) \qquad F_n(b_n) = \frac{1}{4} n^{\frac{1}{3}} \quad ,$$

Suppose that $F_n(0) \leq \frac{1}{4} n^{\frac{1}{3}}$. Then b_n is well-defined. With (8.6.14) and (8.6.15), the value of (8.6.13) equals $12\delta_1^{-2} \sigma^4 m \, a_n^{-2} \, n^{-\frac{1}{3}}$ which tend to 0 if

$$(8.6.16) \qquad n \, a_n^6 \to \infty \, .$$

On the other hand if $F_n(0) > \frac{1}{4} n^{\frac{1}{3}}$. Then, we may take $b_n = 0$. In this case, the second term of (8.6.8) vanishes. Hence the left hand side of (8.6.2) does not exceed (8.6.9) which does not exceed $4m\delta_1^{-2} \sigma^{-4} n^{-\frac{1}{3}} a_n^{-2}$. This quantity also tends to 0 under (8.6.16). Therefore, we have proven (8.6.2) under the assumption of (8.6.16).

Proof of (8.6.3)

Write $\lambda_0 = \infty$ and $\lambda_{n+1} = 0$. Proceeding as in the beginning of the proof of (8.6.2), the left hand side of (8.6.3) does not exceed

$$(8.6.17) \qquad \sum_{j=0}^{n} p\left\{ \sup_{\lambda_{j+1} \le h \le \lambda_j} \frac{\left| \sum_{i=1}^{n} (h+\lambda_i)^{-1} \mu_i \varepsilon_i \right|}{n^{1/2} \left(\sum_{i=1}^{n} (h+\lambda_i)^{-2} (\mu_i^2+\sigma^2) \right)^{1/2}} \ge a_n \delta_1 \sigma \right\}.$$

The numerator can be split into the sum over $1 \le i \le j$ and the sum over $j+1 \le i \le n$. Precisely (8.6.17) does not exceed

$$(8.6.18) \qquad \sum_{j=1}^{n} p\left\{ \sup_{\lambda_{j+1} \le h \le \lambda_j} \frac{\left| \sum_{i=1}^{j} (h+\lambda_i)^{-1} \mu_i \varepsilon_i \right|}{n^{1/2} \left(\sum_{i=1}^{n} (h+\lambda_i)^{-2} (\mu_i^2+\sigma^2) \right)^{1/2}} \ge \frac{1}{2} a_n \delta_1 \sigma \right\}$$

$$+ \sum_{j=0}^{n} p\left\{ \sup_{\lambda_{j+1} \le h \le \lambda_j} \frac{\left| \sum_{i=j+1}^{n} h(h+\lambda_i)^{-1} \mu_i \varepsilon_i \right|}{n^{1/2} \left(\sum_{i=1}^{n} h^2 (h+\lambda_i)^{-2} (\mu_i^2+\sigma^2) \right)^{1/2}} \ge \frac{1}{2} a_n \delta_1 \sigma \right\}.$$

Since $(h+\lambda_i)^{-2}$ is nonincreasing in h, the first term of (8.6.18) does not exceed

$$\sum_{j=1}^{n} p\left\{ \sup_{\lambda_{j+1} \le h \le \lambda_j} \frac{\left| \sum_{i=1}^{j} (h+\lambda_i)^{-1} \mu_i \varepsilon_i \right|}{n^{1/2} \left(\sum_{i=1}^{n} (\lambda_j+\lambda_i)^{-2} (\mu_i^2+\sigma^2) \right)^{1/2}} \ge \frac{1}{2} a_n \delta_1 \sigma \right\}$$

which in turn is no greater than

$$(8.6.19)\qquad \sum_{j=1}^{n} p\left\{ \sup_{\lambda_{j+1}\le h\le\lambda_j} \frac{\left|\sum_{i=1}^{j}[(h+\lambda_i)^{-1}-(\lambda_j+\lambda_i)^{-1}]\mu_i\varepsilon_i\right|}{n^{1/2}\left(\sum_{i=1}^{n}(\lambda_j+\lambda_i)^{-2}(\mu_i^2+\sigma^2)\right)^{1/2}} \ge \frac{1}{4}a_n\delta_1\sigma \right\}$$

$$+\sum_{j=1}^{n} p\left\{ \frac{\left|\sum_{i=1}^{n}(\lambda_j+\lambda_i)^{-1}\mu_i\varepsilon_i\right|}{n^{1/2}\left(\sum_{i=1}^{n}(\lambda_j+\lambda_i)^{-2}(\mu_i^2+\sigma^2)\right)^{1/2}} \ge \frac{1}{4}a_n\delta_1\sigma \right\} .$$

By Cheybychev inequality, the second term of (8.6.19) does not exceed

$$(8.6.20)\qquad \left(\frac{1}{4}a_n\delta_1\sigma\right)^{-4}\sum_{j=1}^{n}\frac{E\left(\sum_{i=1}^{n}(\lambda_j+\lambda_i)^{-1}\mu_i\varepsilon_i\right)^4}{n^2\left(\sum_{i=1}^{n}(\lambda_j+\lambda_i)^{-2}(\mu_i^2+\sigma^2)\right)^2}$$

$$\le \left(\frac{1}{4}a_n\delta_1\sigma\right)^{-4}\sum_{j=1}^{n}\frac{m\left(\sum_{i=1}^{n}(\lambda_j+\lambda_i)^{-2}\mu_i^2\right)^2}{n^2\left(\sum_{i=1}^{n}(\lambda_j+\lambda_i)^{-2}(\mu_i^2+\sigma^2)\right)^2} \quad \text{(by (8.3.2))}$$

which is no greater than $256\,\delta_1^{-4}\sigma^{-4}\,m\,a_n^{-4}n^{-1}$.

Now the first term of (8.6.19) can be rewritten as

$$(8.6.21)\qquad \sum_{j=1}^{n} p\left\{ \sup_{\lambda_{j+1}\le h\le\lambda_j} \frac{\left|\sum_{i=1}^{n}(\lambda_j-h)(h+\lambda_i)^{-1}(\lambda_j+\lambda_i)^{-1}\mu_i\varepsilon_i\right|}{n^{1/2}\left(\sum_{i=1}^{n}(\lambda_j+\lambda_i)^{-2}(\mu_i^2+\sigma^2)\right)^{1/2}} \ge \frac{1}{4}a_n\delta_1\sigma \right\}$$

Since $(\lambda_j-h)(h+\lambda_i)^{-1}$ is nondecreasing in i and is no greater than 1 for $\lambda_{j+1} \le h \le \lambda_j$, we may take $c_i = (\lambda_j-h)(h+\lambda_i)^{-1}$ and $w_i = (\lambda_j+\lambda_i)^{-1}\mu_i\varepsilon_i$ and apply (8.6.6). Thus (8.6.21) does not exceed

$$(\tfrac{1}{4} a_n\delta_1\sigma)^{-4} \sum_{j=1}^{n} \frac{E\,(\sum_{i=1}^{n} (\lambda_j+\lambda_i)^{-1}\mu_i\varepsilon_i)^4}{n^2(\sum_{j=1}^{n} (\lambda_j+\lambda_i)^{-2}\,(\mu_i^2+\sigma^2))^2} = (8.6.20) \quad .$$

Therefore we conclude that the first term of (8.6.18) does not exceed $512\ \delta_1^{-4}\ \sigma^{-4}\ m\ a_n^{-4}\ n^{-1}$.

The second term of (8.6.18) can be evaluated in a similar way. First it does not exceed

$$\sum_{j=0}^{n} p\left\{ \sup_{\lambda_{j+1}\le h\le\lambda_j} \frac{|\sum_{i=j+1}^{n} h(h+\lambda_i)^{-1}\ \mu_i\varepsilon_i|}{n^{1/2}\,(\sum_{i=1}^{n} \lambda_{j+1}^2(\lambda_{j+1}+\lambda_i)^{-2}(\mu_i^2+\sigma^2))^{1/2}} \ge \tfrac{1}{2} a_n\delta_1\sigma \right\},$$

because $h(h + \lambda_i)^{-1}$ is nondecreasing in h. This expression clearly is no greater

$$\text{than } \sum_{j=o}^{h} p\left\{ \sup_{\lambda_{j+1}\le h\le\lambda_j} \frac{|\sum_{i=j+1}^{n} [h\,(h+\lambda_i)^{-1} - \lambda_{j+1}(\lambda_{j+1}+\lambda_i)^{-1}]\mu_i\varepsilon_i|}{n^{1/2}(\sum_{i=1}^{n} \lambda_{j+1}^2(\lambda_{j+1}+\lambda_i)^{-2}(\mu_i^2+\sigma^2))^{1/2}} \ge \tfrac{1}{4} a_n\delta_1\sigma \right\}$$

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

Now by Chebychev inequality and (8.3.2), the second term does not exceed

$$(\tfrac{1}{4} a_n\delta_1\sigma)^{-4} \sum_{j=0}^{n} \frac{m\,(\sum_{i=j+1}^{n} \lambda_{j+1}^2(\lambda_{j+1}+\lambda_i)^{-2}\ \mu_i^2)^2}{n^2\,(\sum_{i=1}^{n} \lambda_{j+1}^2(\lambda_{j+1}+\lambda_i)^{-2}(\mu_i^2+\sigma^2))^2}$$

$$\leq 256\ \delta_1^{-4}\ \sigma_1^{-4}\ \ m\ a_n^{-4}\ n^{-1}\quad ,$$

while by taking $c_i = (h-\lambda_{j+1})(h+\lambda_i)^{-1}$ and $w_i = \lambda_i(\lambda_{j+1}+\lambda_i)^{-1}\ \mu_i\varepsilon_i$ and applying (8.6.6), the first term does not exceed

$$(\tfrac{1}{4}\ a_n\delta_1\sigma)^{-4}\ \sum_{j=0}^{n}\ \frac{E(\sum_{i=j+1}^{n}\ \lambda_i(\lambda_{j+1}+\lambda_i)^{-1}\ \mu_i\varepsilon_i)^4}{n^2\ (\sum_{i=1}^{n}\ \lambda_{j+1}^2(\lambda_{j+1}+\lambda_i)^{-2}(\mu_i^2+\sigma^2))^2}$$

which by (8.3.2) again, is no greater than

$$(\tfrac{1}{4}\ a_n\delta_1\sigma)^{-4}\ \sum_{j=0}^{n}\ \frac{m(\sum_{i=j+1}^{n}\ \lambda_i^2\ (\lambda_{j+1}+\lambda_i)^{-2}\ \mu_i^2)^2}{n(\sum_{i=1}^{n}\ \lambda_{j+1}^2\ (\lambda_{j+1}+\lambda_i)^{-2}\ (\mu_i^2+\sigma^2))^2}.$$

Now observing that $\lambda_i^2 \leq \lambda_{j+1}^2$ for $i \geq j+1$, the last expression is again no greater than $256\ \delta_1^{-4}\ \sigma_1^{-4}\ \ m\ a_n^{-4}\ n^{-1}$.

Combining the results we have obtained, the left hand side of (8.6.3) does not exceed $1024\ \delta_1^{-4}\ \sigma^{-4}\ m\ a_n^{-4}\ n^{-1}$. Hence (8.6.3) holds for n large enough if $a_n^4\ n \to \infty$, which is implied by "$a_n \to 0$" and (8.6.16). This completes the proof of (8.6.3) and hence Lemma 4.2.

8.7. Proof of Theorem 5.1

We shall use the following lemmas.

Lemma 8.7.1. Suppose $b_0 = 0 < b_1 < \ldots < b_k$, and $x_1,\ldots,x_k$ are

independent random variables with means 0 and finite variances. Then for any

$$\varepsilon > 0, \quad p\{ \sup_{1\le h\le k} | b_h^{-1} \sum_{i=1}^{h} x_i | \ge \varepsilon \} \le 4\varepsilon^{-2} \sum_{i=1}^{k} b_i^{-2} Ex_i^2 .$$

Proof. Define $s_0 = 0$ and for $1 \le h \le k$, $s_h = \sum_{i=1}^{h} b_i^{-1} x_i$. Then using the method of Abel's partial sum, it is easy to verify that

$$b_h^{-1} \sum_{i=1}^{h} x_i = b_h^{-1} \sum_{i=1}^{h} (s_i - s_{i-1}) b_i = s_h - b_h^{-1} \sum_{i=1}^{h-1} (b_{i+1} - b_i) s_i .$$

Now, $$\sup_{1\le h\le k} |b_h^{-1} \sum_{i=1}^{h} x_i | \le \sup_{1\le h\le k} |s_h| + (b_h - b_1)^{-1} \sum_{i=1}^{h-1} (b_{i+1} - b_i) |s_i|$$

$$\le 2 \sup_{1\le h\le k} |s_h| .$$

Therefore $$p\{ \sup_{1\le h\le k} |b_h^{-1} \sum_{i=1}^{h} x_i | \ge \varepsilon \}$$

$$\le p\{ \sup_{1\le h\le k} |s_h| \ge \varepsilon/2 \} .$$

Now by Kolomogorov's theorem, the proof of Lemma 8.7.1 is complete. □

Lemma 8.7.2. Suppose $0 < b_1 < \ldots < b_k$. Then

$$\sum_{i=2}^{k} b_i^{-1} - \sum_{i=2}^{k} b_i^{-2} b_{i-1} < b_1^{-1} .$$

Proof. Define $a_0 = 0$ and $a_i = (a_{i-1} + 1)^2/4$ for $i \ge 1$.

Clearly $a_i < 1$. We shall use mathematical induction to show that

$$\sum_{i=2}^{k} b_i^{-1} - \sum_{i=2}^{k} b_i^{-2} b_{i-1} \leq a_{k-1} b_1^{-1} .$$

Suppose the above statement holds for $k \leq n$. Consider the case $k = n + 1$. Observe that $\sum_{i=2}^{n+1} b_i^{-1} - \sum_{i=2}^{n+1} b_i^{-2} b_{i-1} = (\sum_{j=2}^{n} b_{j+1}^{-1} - \sum_{j=2}^{n} b_{j+1}^{-2} b_j) +$ $(b_2^{-1} - b_2^{-2} b_1) \leq a_{n-1} b_2^{-1} + b_2^{-1} - b_2^{-2} b_1$. Here we have used the induction hypothesis. Now, it is easy to check that $(a_{n-1}+1) b_2^{-1} - b_2^{-2} b_1 \leq (a_{n-1}+1)^2/4b_1 =$ $a_n b_1^{-1}$. This completes the proof. □

We begin to prove Theorem 5.1. As we have argued several times, the left hand side of (5.1) does not exceed

$$(8.7.1) \qquad p\{|n^{-1}||\varepsilon_n||^2 - \sigma^2| \geq \frac{1}{2}\delta\} + p\left\{\sup_{h\in H_n} \left| \frac{2\sigma^2(n-h)\sum_{i=h+1}^{n}(\varepsilon_i y_i - \sigma^2)}{n(\sum_{i=h+1}^{n}\mu_i^2 + (n-h)\sigma^2)} \right| \right.$$

$$\geq \frac{1}{2}(1-\Delta)\delta\} + p\{\sup_{h\in H_n} |(\sum_{i=h+1}^{n} y_i^2)/(\sum_{i=h+1}^{n}\mu_i^2 + (n-h)\sigma^2) - 1| \geq \Delta\},$$

for $0 < \Delta < 1$. Note that here we essentially use (3.4) with the last two terms on the right hand side combined together before taking absolute values.

To evaluate the second term of (8.7.1) by Lemma 8.7.1, we first delete the factor "2(n-h)" in the numerator and the factor "n" in the denominator, and then put $x_1 = \sum_{j=p_n+1}^{n} (\varepsilon_j y_j - \sigma^2)$, $x_i = \varepsilon_{p_n-i+2} y_{p_n-i+2} - \sigma^2$ for $2 \leq i \leq p_n$,

$b_i = \sum_{j=p_n+2-i}^{n} (\mu_j^2 + \sigma^2)$ for $1 \le i \le p_n$, $k = p_n$, and $\varepsilon = \sigma^{-2}(1-\Delta_1)(1-\Delta_2)\delta$.

The bound obtained is $16\sigma^4(1-\Delta)^{-2}\delta^{-2} \sum_{i=1}^{p_n} b_i^{-2} Ex_i^2$. Using the normality of ε_i's, we see that $Ex_i^2 = \sigma^2\mu_{p_n-i+2}^2 + 2\sigma^4 \le 2\sigma^2 (\mu_{p_n-i+2}^2 + \sigma^2) = 2\sigma^2(b_i - b_{i-1})$ for $2 \le i \le p_n$ and similarly $Ex_1^2 \le 2\sigma^2 b_1$. Therefore by Lemma 8.7.2, $\sum_{i=1}^{p_n} b_i^{-2} Ex_i^2 \le 4\sigma^2 b_1^{-1} \le 4(n-p)^{-1}$. Thus we conclude that the second term of (8.7.1) does not exceed $64\sigma^4(1-\Delta)^{-2}\delta^{-2}/(n-p)$.

The third term of (8.7.1) can be evaluated in a similar way. Put $x_1 = \sum_{j=p_n+1}^{n} (y_j^2 - \mu_j^2 - \sigma^2)$, $x_i = y_{p_n-i+2}^2 - \mu_{p_n-i+2}^2 - \sigma^2$, for $2 \le i \le p_n$, $b_i = \sum_{j=p_n+2-i}^{n} (\mu_j^2 + \sigma^2)$ for $1 \le i \le p_n$, $k = p_n$, and $\varepsilon = \Delta$. Again $Ex_i^2 = 2\sigma^4 + 4\sigma^2\mu_{p_n-i+2}^2 \le 4\sigma^2(b_i - b_{i-1})$. Therefore the third term of (8.7.1) does not exceed $32\Delta^{-2}(n-p)^{-1}$.

Using Chebychev inequality, the first term of (8.7.1) is bounded by $8\sigma^4 n^{-1}\delta^{-2}$. Therefore (8.7.1) does not exceed

$$8\sigma^4 n^{-1}\delta^{-2} + 32(2\sigma^4(1-\Delta)^{-2}\delta^{-2} + \Delta^{-2})(n-p_n)^{-1} .$$

Minimizing this expression over Δ, $0 < \Delta < |$, we obtain the right hand side of (5.1).

8.8 Proof of Theorem 6.1.

To establish the consistency of $\tilde{\mu}_n(\hat{h},\hat{\sigma}_n)$, it suffies to show that $\frac{1}{n}||\tilde{\mu}_n(\hat{h},\hat{\sigma}_n) - \tilde{\mu}_n(\hat{h})||^2 \to 0$, or equivalently, $(\frac{\hat{\sigma}_n^2}{\sigma^2} - 1) \cdot \frac{1}{n}||\tilde{\mu}_n(\hat{h}) - y_n||^2 \to 0$. Since $\hat{\sigma}_n^2$ is consistent, we need only to check that $\frac{1}{n}||\tilde{\mu}_n(\hat{h}) - y_n||^2$ is bounded in probability. This is done by observing the inequality that $\frac{1}{n}||\tilde{\mu}_n(\hat{h}) - y_n||^2 \le 2(n^{-1}||\tilde{\mu}_n(\hat{h}) - \mu_n||^2 + n^{-1}||\varepsilon_n||^2)$; the first term tends to 0 because of the consistency of $\tilde{\mu}_n(\hat{h})$ and the second term tends to σ^2 by law of large number.

To establish the consistency of $SURE_n(\hat{h},\hat{\sigma}_n)$, it suffices to verify the following two convergence statements:

(i) $|SURE_n(\hat{h},\hat{\sigma}_n) - SURE_n(\hat{h})| \to 0$; (ii) $\frac{1}{n}||\tilde{\mu}_n(\hat{h},\hat{\sigma}_n) - \tilde{\mu}_n(\hat{h})||^2 \to 0$.

(i) follows from the observations that $|SURE_n(\hat{h},\hat{\sigma}_n) - SURE_n(\hat{h})| \le |\hat{\sigma}_n^2 - \sigma^2| + |\hat{\sigma}_n^4 / \sigma^4 - 1| \cdot (\sigma^2 - SURE_n(\hat{h}))$ and that $(\sigma^2 - SURE_n(\hat{h}))$ is bounded above in probability. Here the latter observation holds simply because the consistency of $SURE_n(\hat{h})$ implies that $SURE_n(\hat{h})-\sigma^2$ is bounded below by δ, for any $\delta < 0$, in probability. On the other hand, to establish (ii), we may proceed as in the first paragraph of this subsection. Since we do not assume that $\tilde{\mu}_n(\hat{h})$ is consistent, we have to demonstrate that $n^{-1}||\tilde{\mu}_n(h) - \mu_n||^2$ is bounded in probability by a different argument: the consistency of $SURE_n(\hat{h})$ implies that $n^{-1}||\tilde{\mu}_n(h) - \mu_n||^2$ is bounded by $\delta + SURE_n(\hat{h}) \le \delta + \sigma^2$ in probability for any $\delta > 0$.

Finally, the uniform consistency of $\mathrm{SURE}_n(\hat{h},\hat{\sigma}_n)$ can be established in a similar way. This completes the proof of Theorem 6.1.

8.9 Proof of Lemma 6.1.

The consistency of $\tilde{\underset{\sim}{\mu}}_n(\hat{h})$ and $\mathrm{SURE}_n(\hat{h})$ implies $\mathrm{SURE}_n(\hat{h}) \to 0$, or equivalently,

$$\frac{n||A_n(\hat{h})\underset{\sim}{y}_n||^2}{(\mathrm{tr}\, A_n(\hat{h}))^2} \to \sigma^2 . \tag{8.9.1}$$

Suppose $\tilde{\underset{\sim}{\mu}}_n(\hat{h})$ is consistent. Then $\frac{1}{n}||A_n(\hat{h})\underset{\sim}{y}_n-\underset{\sim}{\varepsilon}_n||^2 = \frac{1}{n}||\hat{\underset{\sim}{\mu}}_n(\hat{h}) - \underset{\sim}{\mu}_n||^2 \to 0$. Hence $\frac{1}{n}||A_n\underset{\sim}{y}_n||^2 \to \sigma^2$. This together with (8.9.1) implies (6.2.1). Conversely, suppose (6.2.1) holds. Since $\frac{1}{n}||\hat{\underset{\sim}{\mu}}_n(\hat{h}) - \underset{\sim}{\mu}_n||^2 \le 2(\frac{1}{n}||\hat{\underset{\sim}{\mu}}_n(\hat{h})-\tilde{\underset{\sim}{\mu}}_n(\hat{h})||^2+\frac{1}{n}||\tilde{\underset{\sim}{\mu}}_n(\hat{h}) -\underset{\sim}{\mu}_n||^2)$, it suffices to show that $\frac{1}{n}||\hat{\underset{\sim}{\mu}}_n(\hat{h})-\tilde{\underset{\sim}{\mu}}_n(\hat{h})||^2 \to 0$, or equivalently,

$$\left(1 - \frac{\sigma^2\, \mathrm{tr}\, A_n(\hat{h})}{||A_n(\hat{h})\underset{\sim}{y}_n||^2}\right)^2 \cdot \frac{1}{n}||A_n(\hat{h})\underset{\sim}{y}_n||^2 \to 0. \tag{8.9.2}$$

Now by (8.9.1) and (6.2.1), we see that the quantity in the parentheses of (8.9.2) tends to 0 and that $\frac{1}{n}||A_n(\hat{h})\underset{\sim}{y}_n||^2 \to \sigma^2$. Therefore (8.9.2) holds. The proof is now complete.

8.10 Proof of Lemma 6.2.

Let Λ_n denote a random variable uniformly distributed on $\{\lambda_{1,n},\cdots, \lambda_{n,n}\}$. We shall write $\lambda_{q,n}$ for $\lambda_{[q\,n],n}$ and $\lambda_{p,n}$ for $\lambda_{[p\,n],n}$. Clearly,

$E\hat{h}(\hat{h}+\Lambda_n)^{-1} = \frac{1}{n}\,\mathrm{tr}\,A_n(\hat{h})$ and $\mathrm{Var}\;\hat{h}(\hat{h}+\Lambda_n)^{-1} = \frac{1}{n}\,\mathrm{tr}\,A_n^2(\hat{h}) - (\frac{1}{n}\,\mathrm{tr}\,A_n(\hat{h}))^2$, where E and Var are with taken with respect to Λ_n only (conditional on h). (6.2.4) and (6.2.2) correspond to

$$(8.10.1)\qquad [\mathrm{Var}\;\hat{h}(\hat{h}+\Lambda_n)^{-1}] \;/\; [E\hat{h}(\hat{h}+\Lambda_n)^{-1}]^2 \to 0$$

and

$$(8.10.2)\qquad E\hat{h}(\hat{h}+\Lambda_n)^{-1} \to 1.$$

Now, fixing $\lambda_{p,n}$ and $\lambda_{q,n}$ for $0 < p < q < 1$, it is easy to see that

$$(8.10.3)\qquad E\hat{h}(\hat{h}+\Lambda_n)^{-1} \le p\,\hat{h}(\hat{h}+\lambda_{p,n})^{-1} + (q-p)\hat{h}(\hat{h}+\lambda_{q,n})^{-1}$$

and

$$(8.10.4)\qquad \mathrm{Var}\;\hat{h}(\hat{h}+\Lambda_n)^{-1} \ge p(1-q)(p+1-q)^{-1}(\hat{h}(\hat{h}+\lambda_{p,n})^{-1} - \hat{h}(\hat{h}+\lambda_{q,n})^{-1})^2.$$

Here the equality of (8.10.3) holds when Λ_n takes values $\lambda_{p,n}$, $\lambda_{q,n}$ and 0 with probabilities p, q-p, and 1-q respectively, while the equality of (8.10.4) holds when Λ_n takes values $\lambda_{p,n}$, $\lambda_{q,n}$ and a suitable number between $\lambda_{p,n}$ and $\lambda_{q,n}$ with probabilities, p, 1-q and q-p respectively. Hence (8.10.1) implies that

$$((\hat{h}+\lambda_{p,n})^{-1} - (\hat{h}+\lambda_{q,n})^{-1})^2 \;/\; (p(\hat{h}+\lambda_{p,n})^{-1} + (q-p)(\hat{h}+\lambda_{q,n})^{-1})^2 \to 0.$$

Multiplying the denominator and the numerator by $(\hat{h}+\lambda_{q,n})^2$ and noting that $\lambda_{p,n} \ge \lambda_{q,n}$, we see that

$$((\hat{h}+\lambda_{q,n})(\hat{h}+\lambda_{p,n})^{-1} - 1)^2 \;/\; [p+(q-p)] \;\to 0.$$

Therefore, we have

$$(\hat{h}\,\lambda_{p,n}^{-1} + \lambda_{q,n}\,\lambda_{p,n}^{-1}) \;/\; (\hat{h}\lambda_{p,n}^{-1} + 1) \;\to 1.$$

Now for p, q such that (6.2.6) holds, we then obtain

$$\hat{h}\,\lambda_{p,n}^{-1} \to \infty \quad \text{and} \quad \hat{h}\,\lambda_{q,n}^{-1} \to \infty \;. \tag{8.10.5}$$

On the other hand, it is clear that

$$1 \geq E\hat{h}(\hat{h}+\Lambda_n)^{-1} \geq \hat{h}(\hat{h}+\lambda_p)^{-1}(q-p) + \hat{h}(\hat{h}+\lambda_q)^{-1}(1-q), \tag{8.10.6}$$

where the second equality holds when Λ_n takes values λ_p, λ_q and ∞ with probabilities q-p, 1-q, and p respectively. Now, (8.10.5) and (8.10.6) imply that

$$\operatorname*{plim}_{n\to\infty} E\,\hat{h}(\hat{h}+\Lambda_n)^{-1} \geq 1-p \;.$$

Finally, taking $p \to 0$, we have established (8.10.2). The proof is now complete.

8.11 Proof of Theorem 6.2.

As discussed in Section 6.2, it suffies to prove (6.2.3), which is in turn implied by the following

$$p\left\{ \inf_{h\in H_n} \frac{||A_n(h)\underset{\sim}{y}_n||^2}{||A_n(h)\underset{\sim}{\mu}_n||^2 + \sigma^2 \operatorname{tr} A_n(h)A_n'(h)} \leq 1 - \delta \right\} \to 0. \tag{8.11.1}$$

Similar to the evaluation of the third term of (8.3.5), the left hand

side of (8.11.1) does not exceed $(1 - \delta)^{-4}\, c''\sigma^{-4} \sum_{h\in H_n} (\lambda'(A_n(h)))^4 \cdot (\mathrm{tr}\, A_n'(h)A_n(h))^{-2}$, which tends to 0 as was shown in the paragraph following Lemma 4.1.

8.12. Proof of Theorem 6.3.

Again, we shall prove (6.2.3). For $\varepsilon > 0$, we have

$$p\left\{\frac{||A_n(\hat{h})y_n||^2}{||A_n(\hat{h})\mu_n||^2 + \sigma^2 \mathrm{tr}\, A_n(\hat{h})} \le 1 - \delta\right\}$$

$$\le p\left\{\inf_{h\in H_n'} \frac{||A_n(h)y_n||^2}{||A_n(h)\mu_n||^2 + \sigma^2 \mathrm{tr}\, A_n(h)} \le 1 - \delta\right\} + \sum_{h\in H_n'}$$

$$\min\left(p\{\hat{h} = h\},\right.$$

$$\left. p\left\{\frac{||A_n(h)y_n||^2}{||A_n(h)\mu_n||^2 + \sigma^2 \mathrm{tr}\, A_n(h)} \le 1 - \delta\right\}\right).$$

The first term of the last expression does not exceed $(1 - \delta)^{-4}\, c''\sigma^{-4} \sum_{h\in H_n'} (n-d(h))^{-2} \le (1 - \delta)^{-4}\, c''\, \sigma^{-4}\, \varepsilon$. Thus it remains to show that the second term tends to 0. Since $\#H_n' \le K$ for any n, we need only to prove that for any sequence $\{h_n \in H_n\}$

$$\min\left(p\{\hat{h} = h_n\},\ (1 - \delta)^{-4}\, c''\, (n-d(h_n))^{-2}\right) \to 0;$$

or equivalently, for any positive number k' and any sequence $\{h_n \in H_n\}$ such that $n - d(h_n) \le k'$, we shall prove that

(8.12.1) $\quad p\{\hat{h} = h_n\} \to 0.$

First observe that by Theorem 4.5, (6.2.1) holds. This implies that $p\{d(\hat{h}) = n\} \to 0$ (since $\text{SURE}(h) = \sigma^2$ if $d(h) = n$) and that (8.9.1) holds. Hence we may further assume that $0 < n - d(h_n) \le k'$. Now, (8.9.1) implies that for $\delta_1 > 0$.

$$(8.12.2) \qquad 1 = \lim_{n\to\infty} p\{||A_n(\hat{h})\underset{\sim}{y}_n||^2 \le (\sigma^2 + \delta_1)n^{-1}(n - d(\hat{h}))^2\}$$

$$\le \lim_{n\to\infty} p\{\hat{h} \neq h_n\} + \lim_{n\to\infty} p\{||A_n(h_n)\underset{\sim}{y}_n||^2 \le (\sigma^2+\delta_1)n^{-1}(k')^2\}.$$

Since $n - d(\hat{h}) > 0$, there exists $\underset{\sim}{c}_n \in R^n$ with $||\underset{\sim}{c}_n|| = 1$ such that $||A_n(h_n)\underset{\sim}{y}_n||^2 \ge (\underset{\sim}{c}_n'\underset{\sim}{y}_n)^2$. Now by Lemma 8.8.2, the second term of the last expression of (8.12.2) tends to 0. Hence, we have $\lim_{n\to\infty} p\{\hat{h} \neq h_n\} = 1$, which implies (8.12.1) as desired.

8.13 Proof of Theorem 6.4.

We shall prove (8.11.1) by using arguments similar to those in the proof of (8.6.1). Let L, Δ be positive numbers to be chosen later. With the definition of $G_n(h)$ given there, let ℓ_n be the largest integer such that $(1 + \Delta)^{\ell_n} L < G_n(0)$ and k_n be the largest integer such that $(1+\Delta)^{k_n} L \le G_n(\infty)$. For i, $\ell_n + 1 \le i \le k_n$, define $h_i^{(n)} = G_n^{-1}((1+\Delta)^i L)$. Take $h_{\ell_n}^{(n)} = 0$ and $h_{k_n+1}^{(n)} = \infty$. Now (6.2.7) implies that for n large, $\ell_n \ge 1$. Evaluating the left hand of (8.11.1) in a manner similar to that of the

evaluation of the left hand side of (8.6.1), we get an upper bound:

$(1-(1+\Delta)(1-\delta))^{-2} \cdot c \cdot L^{-1} \sum_{i=\ell_n}^{k_n} (1+\Delta)^{-j} \leq (\delta-\Delta)^{-2} \cdot c \cdot L^{-1} \Delta^{-1}$. Here we require $\Delta < \delta$. To make the last expression expression less than ε, we may take $L > c(\delta-\Delta)^{-2} \Delta^{-1} \varepsilon^{-1}$. This proves (8.11.1) and hence Theorem 6.4.

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.

Casella, G. (1980). Minimax ridge regression estimation. Ann. Statist. 8, 1036~1056.

Chung, K. L. (1974). A course in probability theory. Second edition. 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.

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.

Knaff, G., Sacks, J. and Ylvisaker, D. (1982). Model robust confidence intervals II. In Statistical Decision Theory and Related Topics III (Ed. Gupta, S. and Berger, J.), 87-102.

Li, K. C. (1982a). Regression models with infinitely many parameters: consistency of bounded linear functionals. Mimeograph Series #82-10, Department of Statistics, Purdue University.

Li, K. C. (1982b). Consistency for cross-validated nearest neighbor estimates in nonparametric regression. Mimeograph Series #82-41, Department of Statistics, Purdue University.

Li, K. C., and Hwang, J. (1982). The data smoothing aspect of Stein estimates. Mimeograph Series #82-38, Department of Statistics, Purdue University.

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. (1983). Bandwidth choice for nonparametric kernel regression. Manuscript.

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. (1981a). Spline smoothing and optimal rates of convergence in nonparametric regression models. Ann. Statist., to appear.

Speckman, P. (1981b). 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. Manuscript.

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. (1980). Optimal rates of convergence for nonparametric estimators. Ann. Statist. 8, 1348~1360.

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.

Uteras, F. (1978). Cross validation techniques for smoothing in one or two dimensions. In Smoothing Techniques for Curve Estimation (T. Gasser and M. Rosenblatt, ed.). Lecture Notes in Mathematics No. 757. Springer, New York.

Uteras, 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. R. Statist. Soc. 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 Vol. 2 (Ed. Gupta, S. and Berger, J.), 383~418.

Wong, W. H. (1982). On the consistency of cross-validation in kernel nonparametric regression. To appear in Ann. Statist.