THE DATA-SMOOTHING ASPECT OF STEIN ESTIMATES

by

Ker-Chau Li*
Department of Statistics
Purdue University

Jiunn Tzon Hwang**
Department of Mathematics
Cornell University

Technical Report #82-38

Department of Statistics
Purdue University

November 1982

*This work is supported in part by National Science Foundation under grant No. MCS-82-00631

**This work is supported in part by National Science Foundation under grant No. MCS-83-01139.

THE DATA-SMOOTHING ASPECT OF STEIN ESTIMATES

by

Ker-Chau Li*
Department of Statistics
Purdue University

Jiunn Tzon Hwang**
Department of Mathematics
Cornell University

Abstract

The data smoothing aspect of Stein estimates is explored in the nonparametric regression settings. We show that appropriately shrinking the raw data towards any linear smoother will provide a robust "smoother" (which dominates the raw data and hence has a bounded maximum risk when the average squared error loss is concerned).

Keywords and Phrases: consistency, kernel estimates, nearest neighbor estimates, nonparametric regression, smoothing splines, Stein effect.

AMS 1980 Subject Classification: Primary 62C20, 62G99
Secondary 62F35, 62J99

1. Introduction

In the estimation of the mean $\underset{\sim}{\theta}=(\theta_1,\ldots,\theta_n)'$ of an n-dimensional normal random vector $\underset{\sim}{y}=(y_1,\ldots,y_n)'$ with the squared length of the error vector as loss when the covariance matrix is an identity, it has been well-known that the James-Stein estimate $\hat{\underset{\sim}{\theta}}=(1-\frac{n-2}{||\underset{\sim}{y}||^2})\,\underset{\sim}{y}$ (James and Stein 1961) improves the

* Work supported by National Science Foundation under grant No. MCS-82-00631.

** Work supported by National Science Foundation under grant No. MCS-83-01139.

trivial estimate $\underset{\sim}{y}$ when $n \geq 3$. $\hat{\underset{\sim}{\theta}}$ shrinks $\underset{\sim}{y}$ toward the origin $\underset{\sim}{0}$. Practically, the shrinking center need not be $\underset{\sim}{0}$. For instance, if we feel that all θ_i are close to each other, it would be appropriate to shrink $\underset{\sim}{y}$ toward $\bar{\underset{\sim}{y}} = (\frac{1}{n} \sum_{i=1}^{n} y_i, \ldots, \frac{1}{n} \sum_{i=1}^{n} y_i)'$. In this case, the estimate would be

$\hat{\underset{\sim}{\theta}} = \underset{\sim}{y} - (\frac{n-3}{||\underset{\sim}{y}-\bar{\underset{\sim}{y}}||^2})(\underset{\sim}{y}-\bar{\underset{\sim}{y}})$. Certainly, under the null case that all θ_i are equal to each other, we should use $\bar{\underset{\sim}{y}}$. But on the other hand, using $\bar{\underset{\sim}{y}}$ may result a huge bias when the null case turns out false (the supremum of the risks is ∞); $\hat{\underset{\sim}{\theta}}$ is safer than $\bar{\underset{\sim}{y}}$. This robustness viewpoint can also be formulated in the Bayes terminologies (Berger 1980). Plotting $\hat{\underset{\sim}{\theta}}$ and $\underset{\sim}{y}$ separately against the coordinate indices, we see that $\hat{\underset{\sim}{\theta}}$ is "smoother" than $\underset{\sim}{y}$, in the sense that the data points in the plot for $\hat{\underset{\sim}{\theta}}$ are closer to a straight line than those in the plot for $\underset{\sim}{y}$.

This data smoothing aspect of Stein estimates will become clearer when the observations $y_1, y_2, \ldots, y_n$ are made at the levels $x_1, x_2, \ldots, x_n \in [0,1]$, with

$$y_i \equiv \theta_i + \varepsilon_i = f(x_i) + \varepsilon_i$$

where f is an unknown smooth function from a class $\mathfrak{F}$. Some appropriate definitions of $\mathfrak{F}$ will be given in the examples of Section 3. Here we only require that $\mathfrak{F} \subset \mathfrak{F}_0 = \{f | f$ is a real function on $[0,1]$ such that $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 < \infty\}$. For the settings of the parametric regression with p regressors, $\mathfrak{F}$ is a finite dimensional space and $\underset{\sim}{\theta}$ often lies in a p-dimensional subspace of R^n; for the nonparametric regression settings, $\mathfrak{F}$

has infinite dimensions and the range of $\underset{\sim}{\theta}$ is often the whole R^n. While the arguments to be used will also apply to the parametric case, we shall focus our attentions to the nonparametric settings hereafter.

Many nonparametric procedures have been proposed for estimating f, including the kernel estimates, the nearest neighbor estimates, and the spline estimates. The asymptotic properties for these estimates such as consistency or convergent rates have already been widely studied. The readers may find a number of references from Stone (1977); see also Agarwal and Studden (1980), Craven and Wahba (1979), Spiegelman and Sacks (1980), Stone (1980, 1982) and Rice and Rosenblatt (1981). Basically, these estimates are linear in the y_i's. Thus for such an estimate $\hat{f}(\cdot)$, the maximum mean average squared error

$$(1.1) \qquad \sup_{f \in \mathfrak{F}} E\left\{\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - \hat{f}(x_i))^2\right\}$$

is infinite except for the trivial case $\hat{f}(x_i)=y_i$, which is of course not consistent asymptotically. But in view of the Stein effect we may still hope that there may exist an estimate $\underset{\sim}{\hat{f}}_n=(\hat{f}(x_1),\ldots,\hat{f}(x_n))'$ which not only dominates $\underset{\sim}{y}$ (hence has bounded risks) but also is consistent in the sense that for any $f \in \mathfrak{F}$,

$$(1.2) \qquad \frac{1}{n}||\underset{\sim}{f}_n-\underset{\sim}{\hat{f}}_n||^2 \to 0, \text{ in probability,}$$

as $n\to\infty$, where $\underset{\sim}{f}_n=(f(x_1),\ldots,f(x_n))'$ and $||\cdot||$ is the Euclidian norm (Typically, we shall assume that the x sequence is dense in [0,1]). If such an $\underset{\sim}{\hat{f}}_n$ can be constructed, then it can be viewed as a robust data "smoother" because it "smooths" the noisy data $\underset{\sim}{y}$ (at least when n is large) and does this in a totally safe manner (the risks are always less than those for the raw data $\underset{\sim}{y}$). The general framework for the construction of $\hat{\underset{\sim}{f}}_n$ will be given in Section 2. Under some conditions, $\hat{\underset{\sim}{f}}_n$ will perform asymptotically at least as well as the usual linear estimates. Section 3 provides some examples. The case of the unknown variances, and some other remarks are discussed in Section 4.

2. Main results.

The main tool used here is due to Stein (1981); namely the estimate of the form

$$(2.1)\qquad \hat{\underset{\sim}{\theta}}=\underset{\sim}{y} - \frac{1}{\underset{\sim}{y}'B\underset{\sim}{y}} \cdot A\underset{\sim}{y}$$

where

$$(2.2)\qquad B=\{(\text{trace } A)\cdot I-2A\}^{-1}A^2$$

and A is a symmetric matrix with

$$(2.3)\qquad 2A < (\text{trace } A)\cdot I$$

in the sense that $\lambda(A)$, the largest characteristic root of A, is less than half of the trace of A. Stein showed that this estimate dominates $\underset{\sim}{y}$ and applied it to the case of three-term moving averages for a suitable A. In this section, we shall demonstrate that the desired robust smoother $\hat{\underset{\sim}{f}}_n$ of Section 1 can be constructed exactly in the same manner.

Consider a sequence of symmetric matrices $\{M_n\}_{n=1}^{\infty}$ such that for any $f \in \mathfrak{F}$,

$$(2.4)\qquad E\frac{1}{n}||\underset{\sim}{f}_n-M_n\underset{\sim}{y}||^2 \to 0\ , \quad \text{as } n \to \infty.$$

Theorem 1. Assume that ε_i's are i.i.d. with mean 0 and variance 1. Let $\hat{\underset{\sim}{f}}_n$ be the $\hat{\underset{\sim}{\theta}}$ of (2.1) with $A = I-M_n$ and B determined by (2.2). Then the following results, (i) $\sim$ (iv), hold:

(i) Assuming the normality of ε_i's, there exists an N such that for $n \geq N$, $\hat{\underset{\sim}{f}}_n$ dominates $\underset{\sim}{y}$.

(ii) $\{\hat{\underset{\sim}{f}}_n\}$ is consistent in the sense that (1.2) holds for any $f \in \mathfrak{F}$.

(iii) Suppose that the convergent rate of (2.4) is no faster than n^{-1}; that is

$$\varliminf_{n\to\infty} E||\underset{\sim}{f}_n - M_n\underset{\sim}{y}||^2 > 0. \tag{2.5}$$

Then the convergent rate of (1.2) is no slower than that of (2.4) in the sense that for any sequence $\{\gamma_n\}$ of positive numbers such that

$$\gamma_n E\frac{1}{n}||\underset{\sim}{f}_n - M_n\underset{\sim}{y}||^2 \to 0, \text{ as } n \to \infty, \tag{2.6}$$

we have

$$\gamma_n n^{-1}||\underset{\sim}{f}_n - \hat{\underset{\sim}{f}}_n||^2 \to 0, \text{ in probability as } n \to \infty. \tag{2.7}$$

(iv) Suppose that the 4th moment of ε_i is finite, and

$$\text{tr } M_n^2 \to \infty \text{ and } (n^{-1}\text{tr } M_n)^2/n^{-1}\text{tr } M_n^2 \to 0, \text{ as } n \to \infty. \tag{2.8}$$

Then we have

$$n^{-1}||\hat{\underset{\sim}{f}}_n - \underset{\sim}{f}_n||^2 = (1 + o_p(1)n^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2 + o_p(En^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2). \tag{2.9}$$

If in addition we assume $\lambda(M_n^2) / \text{tr } M_n^2 \to 0$, then

$$n^{-1}||\hat{\underset{\sim}{f}}_n - \underset{\sim}{f}_n||^2 = (1 + o_p(1))n^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2. \tag{2.9'}$$

Proof. (i) Write $A_n = I - M_n$ and $\underset{\sim}{\varepsilon}_n = (\varepsilon_1,\ldots,\varepsilon_n)'$. Since $n^{-1}\text{tr } M_n^2 \leq En^{-1}||\underset{\sim}{f}_n - M_n\underset{\sim}{y}||^2$, by (2.4) we have

$$(n^{-1}\text{tr } M_n)^2 \leq n^{-1} \text{ tr } M_n^2 \to 0. \tag{2.10}$$

Next observe that

$$-(\text{tr } M_n^2)^{1/2}I \leq M_n \leq (\text{tr } M_n^2)^{1/2}I \tag{2.11}$$

in the sense of nonnegative definiteness. From (2.11), it follows that $n^{-1}|\lambda(A_n)| \leq n^{-1} + n^{-1/2}(n^{-1}\text{tr } M_n^2)^{1/2}$. On the other hand $n^{-1}\text{tr } A_n = 1 - n^{-1}\text{tr } M_n$. Thus by (2.10) we see that (2.3) holds for n large enough where $A = A_n$. Now applying Stein's result, the proof for (i) is complete.

(ii) Due to (2.4) and the inequality that

$$\begin{aligned} n^{-1}||\hat{\underset{\sim}{f}}_n-\underset{\sim}{f}_n||^2 &\leq (1-(\underset{\sim}{y}'B_n\underset{\sim}{y})^{-1})^2n^{-1}||\underset{\sim}{\varepsilon}_n||^2 \\ &+ 2|1-(\underset{\sim}{y}'B_n\underset{\sim}{y})^{-1}|(\underset{\sim}{y}'B_n\underset{\sim}{y})^{-1}n^{-1}||\underset{\sim}{\varepsilon}_n||\cdot||M_n\underset{\sim}{y} - \underset{\sim}{f}_n|| \\ &+ (\underset{\sim}{y}'B_n\underset{\sim}{y})^{-2}n^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2 , \end{aligned} \tag{2.12}$$

it suffices to show that

$$\underset{\sim}{y}'B_n\underset{\sim}{y} \to 1 \quad \text{in probability.} \tag{2.13}$$

Now (2.10), (2.11) and the inequalities $(1-x)^{-1} \leq 1+2x$ for small positive x and $(1+x)^{-1} \geq 1 - x$ for $x > 0$,

$$n^{-1}(1+2n^{-1}(2+|\text{tr } M_n|+ 2(\text{tr } M_n^2)^{1/2}))A_n^2 \geq B_n \geq n^{-1}(1-n^{-1}|\text{tr } M_n|-2n^{-1}(\text{tr } M_n^2)^{1/2})A_n^2,$$

from which it follows that

$$\begin{aligned} |\underset{\sim}{y}'B_n\underset{\sim}{y}-1| &\leq |n^{-1}||A_n\underset{\sim}{y}||^2-1| + 2n^{-1}(2+|\text{tr } M_n|+ 2(\text{tr } M_n^2)^{1/2})n^{-1}||A_n\underset{\sim}{y}||^2 \\ &\leq |n^{-1}||A_n\underset{\sim}{y}||^2-1| + 2(2n^{-1} + 3(n^{-1}\text{tr } M_n)^{1/2})n^{-1}||A_n\underset{\sim}{y}||^2. \end{aligned} \tag{2.14}$$

Now in view of (2.10), it remains to show that

(2.15) $$n^{-1}||A_n\underset{\sim}{y}||^2 - 1 \to 0 \quad \text{in probability.}$$

Finally, since

(2.16) $$|1-n^{-1}||A_n\underset{\sim}{y}||^2| \le n^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2 + 2|n^{-1} < M_n\underset{\sim}{y} - \underset{\sim}{f}_n, \varepsilon_n >| + |n^{-1}||\underset{\sim}{\varepsilon}||^2 - 1|$$

(2.15) follows from (2.4), Cauchy-Schwartz inequality and the fact that $n^{-1}||\underset{\sim}{\varepsilon}||^2 \to 1$.

(iii) From (2.12) and (2.13), it is clear that (2.7) is implied by $\gamma_n(1-\underset{\sim}{y}B_n\underset{\sim}{y})^2 \to 0$ in probability, which in view of (2.14) will follow from

(2.17) $$\gamma_n(1 - n^{-1}||A_n\underset{\sim}{y}||^2)^2 \to 0 \quad \text{in probability,}$$

(2.18) $$\gamma_n n^{-2} \to 0$$

and

(2.19) $$\gamma_n n^{-1} \text{tr } M_n^2 \to 0.$$

Now (2.5) and (2.6) implies that

(2.20) $$\gamma_n n^{-1} \to 0 ,$$

which implies (2.18). (2.19) obviously follows from (2.6). Finally (2.17) follows from (2.16), (2.6) and (2.20). This complete the proof of (iii).

(iv) First we shall prove (2.9). By (2.12) and (2.13), it suffices to show that

$$(1-\underset{\sim}{y}'B_n\underset{\sim}{y})^2 = o_p(n^{-1}||M_n\underset{\sim}{y}-\underset{\sim}{f}_n||^2 + En^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2) ,$$

which in view of the first inequality of (2.14), will hold if

(2.21) $$(n^{-1}||A_n\underset{\sim}{y} - 1||^2 - 1)^2 = o_p(n^{-1}\text{tr } M_n^2 + n^{-1}||A_n\underset{\sim}{f}_n||^2 + n^{-1}||M_n\underset{\sim}{y}-\underset{\sim}{f}_n||^2)$$

and

(2.22) $$n^{-2}(2 + |\text{tr } M_n| + 2(\text{tr } M_n^2)^{1/2})^2 = o(n^{-1}\text{tr } M_n^2) .$$

Now (2.22) clearly follows from (2.8) while in view of (2.16), (2.21) will hold if we have

(2.23) $$(n^{-1}||\underset{\sim}{\varepsilon}_n||^2 - 1)^2 = o_p(n^{-1}\text{tr } M_n^2) ,$$

(2.24) $$(n^{-1} < M_n\underset{\sim}{\varepsilon}_n, \underset{\sim}{\varepsilon}_n >)^2 = o_p(n^{-1}\text{tr } M_n^2) ,$$

and

(2.25) $$(n^{-1} < A_n\underset{\sim}{f}_n, \underset{\sim}{\varepsilon}_n >)^2 = o_p(n^{-1}||A_n\underset{\sim}{f}_n||^2) .$$

Finally (2.23) follows from (2.8); (2.25) holds because $E(n^{-1} < A_n\underset{\sim}{f}_n, \underset{\sim}{\varepsilon}_n >)^2 = n^{-2}||A_n\underset{\sim}{f}_n||^2$; (2.24) holds because

$$E(n^{-1} < M_n\underset{\sim}{\varepsilon}_n, \underset{\sim}{\varepsilon}_n >)^2 = (En^{-1} < M_n\underset{\sim}{\varepsilon}_n, \underset{\sim}{\varepsilon}_n >)^2 + \text{Var } n^{-1} < M_n\underset{\sim}{\varepsilon}_n, \underset{\sim}{\varepsilon}_n >$$

$$\leq (n^{-1}\text{tr } M_n)^2 + m\, n^{-2}\text{tr } M_n^2$$

where m denotes the 4th moment of ε_i. Therefore (2.9) is established. To show (2.9'), we need to prove that $o(En^{-1}||M_n\underset{\sim}{y}-\underset{\sim}{f}_n||^2) = o_p(n^{-1}||M_n\underset{\sim}{y}-\underset{\sim}{f}_n||^2)$, which in turn will hold if

(2.26) $$(\text{Var } n^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2) / (En^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2)^2 \to 0.$$

A straightforward computation shows that $\text{Var } n^{-1}||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2 \leq 2n^{-2}(\text{Var}||M_n\underset{\sim}{\varepsilon}||^2 + \text{Var } 2 < M_n\underset{\sim}{\varepsilon}, A_n\underset{\sim}{f}_n >) \leq 2n^{-2}(m \text{ tr } M_n^4 + 4||M_nA_n\underset{\sim}{f}_n||^2)$
$\leq 2\lambda(M_n^2)n^{-2}(m \text{ tr } M_n^2 + 4||A_n\underset{\sim}{f}_n||^2) \leq 2(m+2)\lambda(M_n^2)(\text{tr } M_n^2)^{-1} \cdot (n^{-1}\text{tr } M_n^2 + n^{-1}||A_n\underset{\sim}{f}_n||^2)^2$

$= 2(m+2)\lambda(M_n^2)(\text{tr } M_n^2)^{-1} (E||M_n\underset{\sim}{y} - \underset{\sim}{f}_n||^2)^2$. Therefore (2.26) holds, completing the proof. □

Note that (2.8) implies (2.5) and in nonparametric regression (2.8) holds typically; see Section 3 for examples. Moreover, $\lambda(M_n^2)$ is usually bounded. In fact if $\lambda(M_n^2) > 1$ then the linear estimate $M_n\underset{\sim}{y}_n$ is obviously inadmissible and can be improved upon by other linear estimates; for example, writing $M_n = \sum_{i=1}^{n} \lambda_i \underset{\sim}{e}_i \underset{\sim}{e}_i'$ with $\underset{\sim}{e}_i$'s being eigenvectors with eigenvalues λ_i's, and putting $\lambda_i' = \min\{1, \max\{\lambda_i, 0\}\}$, it is clear that $(\sum_{i=1}^{n} \lambda_i' \underset{\sim}{e}_i \underset{\sim}{e}_i')\underset{\sim}{y}$ improves $M_n\underset{\sim}{y}$.

Quite often M_n may be asymmetric. Thus in what follows, we shall construct a reasonable Stein estimate of the form similar to (2.1) for the asymmetric A.

For any n×n matrix A, let $\lambda\{A\}$ denote the maximum eigenvalue of $\{\frac{A+A'}{2}\}$; suppose that

(2.27) $$\text{trace } A > 2\lambda\{A\} .$$

Define $\hat{\underset{\sim}{\theta}}$ by (2.1) with B determined by

(2.28) $$B = r^{-1}A'A$$

where r is a positive number such that

(2.29) $$0 < r < 2\,[\text{trace } A - 2\lambda\{A\}\,].$$

<u>Proposition</u> 1. Assume that (2.1),(2.27) ∿ (2.29) hold. Then $\hat{\underset{\sim}{\theta}}$ dominates $\underset{\sim}{y}$.

The proof of this proposition is given in the Appendix. A good choice of r seems to be $r = \text{trace } A - 2\lambda\{A\}$. Using this r and defining $\hat{\underset{\sim}{f}}_n$ to be the $\hat{\underset{\sim}{\theta}}$

of (2.1) with $A = I - M_n$ and B determined by (2.28), Theorem 1 holds (where M_n^2 should be replaced by $M_n'M_n$). We omit the proof because it is similar to the symmetric case.

Before closing this section we introduce the following lemma which will be used in Section 3. The proof is given in the Appendix.

Lemma 1. For any $n\times n$ matrix A, the maximum singular value of A(i.e., the square root of $\lambda(A'A)$) is no less than $\lambda\{A\}$.

3. Examples.

Example 1. Periodical f and the symmetrized nearest neighbor method.

Take $\mathfrak{F}$ = {f|f is continuously differentiable on [0,1] such that $f(0)=f(1)$ and $f'(0)=f'(1)$}. Suppose $x_1\le\ldots\le x_n \in [0,1]$ satisfy the condition that as $n\to\infty$,

$$(3.1)\qquad \max\{x_{i+1}-x_i \mid i=0,1,\ldots,n-1\} \to 0$$

where $x_0\equiv x_n-1$. Consider the following simple variant of the nearest neighbor estimate of $f(x_i)$ defined by

$$(3.2)\qquad \sum_{j=0}^{k} w_{jn}(y_{i+j}+y_{i-j})$$

where w_{jn}, $j=0,\ldots,k$ are nonnegative numbers such that

$$(3.3)\qquad \sum_{j=0}^{k} w_{jn}=\frac{1}{2},$$

and we write $y_{n+j}=y_j$ and $y_{-j}=y_{n-j}$. Choose k suitably such that as $n\to\infty$, (3.4) $k \to \infty$,

(3.5) $\max \{x_{i+k}-x_i \mid i=-k+1,-k+2,\ldots,n-k\} \to 0$,

and

(3.6) $\sup_{0 \le j \le k} w_{jn} \to 0$,

where we write $x_i = x_{n+i} - 1$ for $i \le 0$.

Denote the estimate of $\underset{\sim}{f}_n$ defined by (3.2) by $M_n \underset{\sim}{y}$ for a symmetric matrix M_n. Then under (3.1) ∿ (3.6), it can be shown that (2.4) holds; see Priestley and Chao (1972) for the related results. Thus using Theorem 1 we obtain a robust smoother by shrinking $\underset{\sim}{y}$ toward the symmetrized nearest neighbor estimate $M_n \underset{\sim}{y}$ suitably. To see how large N will be in (i) of Theorem 1, we need to compare the trace and the maximum eigenvalue of $A=I-M_n$. The following lemma is helpful. The proof is given in the Appendix.

<u>Lemma 2</u>. For the M_n defined by the symmetrized nearest neighbor estimates (3.2) and (3.3) the maximum eigenvalue of M_n^2 is no greater than 1.

Using this lemma, (2.3) holds if $n(1-2w_{on}) > 4$. Moreover, since tr $M_n^2 = 2n \sum_{j=o}^{k} w_{jo}^2 \ge 2^{-1} n(k+1)^{-1}$, (3.1) and (3.5) imply that tr $M_n^2 \to \infty$. To ensure $(n^{-1} \operatorname{tr} M_n)^2 / n^{-1} \operatorname{tr} M_n^2 \to 0$, we may impose the condition that $k\, w_{on}^2 \to 0$, which can be easily satisfied, for example by $w_{on} = k^{-1}$.

<u>Example 2</u>. Nearest neighbor and kernel estimates.

Take $\mathfrak{F} = \{f \mid f$ is continuously differentiable on $[0,1]\}$. Consider the k -

nearest neighbor estimate first. Denote the jth closest point to point x_i among $x_1, \ldots, x_n$ by $x_{j(i)}$ (ties are broken in a systematic way). Given a sequence of positive numbers $w_{1n}, \ldots, w_{kn}$ such that $\sum_{i=1}^{k} w_{in} = 1$, the k - nearest neighbor estimate for $f(x_i)$ is defined by

$$(3.7) \qquad \sum_{j=1}^{k} w_{jn} y_{j(i)}$$

where j(i) is the index such that $x_{j(i)}$ is the jth nearest neighbor to x_i. Assume that as $n \to \infty$, $\{x_1, \ldots, x_n\}$ gets dense in [0,1]. Choose k and $\{w_{in}\}_{i=1}^{k}$ such that as $n \to \infty$, we have $k \to \infty$, $\sup_{1 \le i \le n} | x_i - x_{k(i)} | \to 0$, and $\sup \{w_{in} \mid i=1,\ldots,k\} \to 0$. Let $M_n \underset{\sim}{y}$ denote the estimate for $\underset{\sim}{f}_n$ defined by (3.7). It can be shown that $M_n \underset{\sim}{y}$ is consistent in the sense that (2.4) holds. To evaluate $\lambda\{I-M_n\}$, we find the following lemma helpful, whose proof is given in the Appendix.

<u>Lemma 3</u>. Assume that $w_{1n} \ge w_{2n} \ge \cdots \ge W_{kn}$. Then the maximum singular value of the matrix M_n defined by the estimate (3.7) is no greater than $\sqrt{2}$.

Using this lemma and Lemma 1, (2.27) will hold if $n(1-w_{1n}) > 2(1+ \sqrt{2})$. However, if the weight sequence $\{w_{in}\}$ is not decreasing, then it seems hard to find a useful bound for $\lambda\{M_n\}$. As in Example 1, to ensure $(n^{-1}\text{tr } M_n)^2/ n^{-1}\text{tr } M_n^2 \to 0$, it suffices to have $k^{-1}w_{in}^2 \to 0$. Similar results apply to the kernel estimates. We omit the details.

<u>Example 3</u>. Smoothing splines.

Consider the case that $\mathfrak{F} = W_2^k[0,1] = \{f | f$ has absolutely continuous derivatives $f, f', \ldots, f^{(k-1)}$ and $\int_0^1 f^{(k)}(x)^2 dx < \infty\}$. The smoothing spline estimate for f is the solution solving

$$(3.8)\qquad \underset{f \in \mathfrak{F}}{\text{Min}}\ \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + h_n \int_0^1 f^{(k)}(x)^2 dx$$

where the smoothing parameter h_n is a positive number. Let M_n be the nxn matrix such that $M_n \underset{\sim}{y}$ is the above smoothing spline solution evaluating at $x_1, x_2, \ldots, x_n$. It can be shown that M_n is symmetric and for an appropriately chosen sequence of h_n (e.g., $h_n = 0(n^{-2k/2k+1})$), $M_n \underset{\sim}{y}$ will be consistent (see, for example, Wahba 1978). Thus Theorem 1 is applicable. To see when (2.3) holds, we need to compute trace $(I-M_n)$ and $\lambda\{I-M_n\}$. Some results from Demmler and Reinsch (1975) (see also Reinsch 1967 or Speckman 1981 a, b, 1982) will be useful. Introduce the space of natural polynomial splines S_n^k defined by $S_n^k = \{f\colon f \in C^{2k-2}[0,1]$, f is a polynomial of degree $2k-1$ on (x_i, x_{i+1}) $i=1,\ldots,n-1$, and $f^{(k)} \equiv 0$ on $[0,x_1]$ and $[x_n,1]\}$.

Let $\{\phi_{jn}\}_{j=1}^n$ be the eigenfunctions with eigenvalues $\{\rho_{jn}\}_{j=1}^n$ satisfying

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

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

for $j, j' = 1,\ldots,n$, with

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

Here $\delta_{jj'}$ is the Kronecker delta. Note that $\{\phi_{jn}\}_{j=1}^n$ is a basis of S_n^k, and the smoothing spline solution of (3.8) can be written as a linear combination of ϕ_{jn}'s. Moreover, it was shown that trace $M_n = \sum_{i=1}^{n} (1+h_n \rho_{in})^{-1}$ and $\lambda\{I-M_n\} = \frac{h_n \rho_{nn}}{1+h_n \rho_{nn}}$. Now we can obtain the following.

Lemma 4. Suppose $\rho_{nn} < \sum_{i=1}^{n-1} \rho_{in}$. Then for any $h_n > 0$, trace $\{I-M_n\} > 2\lambda\{I-M_n\}$.

The proof of this lemma will be given in the Appendix. Note that $\lambda\{M_n\} = 1$ and under a mild condition on the sequence $\{x_i\}_{i=1}^{n}$, (2.8) will hold if h_n is chosen appropriately, so that $h_n \to 0$ and $n\, h_n^{1/2k} \to \infty$ (see Craven and Wahba (1979)).

4. Remarks.

Remark 1. If the common variance σ^2 of ε's is not known but we also observe a real random variable S, distributed independently of $\underset{\sim}{y}$ as $\sigma^2 \chi^2_k$. Then Stein (1981) showed that instead of (2.1), the estimate

$$(4.1) \qquad \hat{\underset{\sim}{\theta}} = y - \frac{S}{k+2} \cdot \frac{1}{y'By} \cdot Ay$$

dominates $\underset{\sim}{y}$. Similarly, for asymmetric A, Proposition 1 holds if (2.1) is replaced by (4.1). If as $n \to \infty$, $k \to \infty$, the consistency result of (ii) of Theorem 1 holds. Moreover, if $\underline{\lim}\, kEn^{-1}||M_n \underset{\sim}{y}_n - \underset{\sim}{f}_n||^2 > 0$ $(= \infty)$, then (iii) ((iv) respectively,) of Theorem 1 also holds. This can be easily seen by observing that $n^{-1}||\hat{\underset{\sim}{f}}_n - \hat{\underset{\sim}{f}}_n(\sigma^2)||^2 = O_p(k^{-1})$, where $\hat{\underset{\sim}{f}}_n$ is the estimate $\hat{\underset{\sim}{\theta}}$ of (4.1) and $\hat{\underset{\sim}{f}}_n(\sigma^2)$ is $\hat{\underset{\sim}{\theta}}$ of (4.1) with S/(k+2) being replaced by σ^2.

Remark 3. It is well-known that Stein effect occurs for distributions other than the normal one; for example, see Shinozaki (1984). But even if Stein estimate (2.1), does not dominate $\underset{\sim}{y}$, it still has a bounded maximum risk (because $(\frac{1}{\underset{\sim}{y}'B\underset{\sim}{y}})^2 \underset{\sim}{y}'A'A\underset{\sim}{y} \le \frac{\text{trace } A}{\underset{\sim}{y}'A'A\underset{\sim}{y}}$) provided that the distribution has a bounded density. Thus the advantage of $\hat{\underset{\sim}{f}}_n$ over linear smoother $M_n\underset{\sim}{y}$ does not depend on the normality assumption.

Remark 4. The average squared error loss in (1.1) is reasonable in the case that we are interested in predicting the values of f at $x_1,\ldots,x_n$. Suppose we are also interested in interpolating to other x values. Then the loss function would be different; e.g., it may be $\int_0^1 (f(t)-\hat{f}(t))^2 w(t)dt$ with a chosen weight function $w(t) \ge 0$. It is clear that with such a loss function, any estimate (since it is based on only finitely many observations) would have infinite maximum risk unless the $\mathfrak{F}$ is either finite-dimensional or bounded in certain sense (e.g., the second derivative of f is less than a fixed number). Thus it is difficult to discuss Stein effect for such loss functions. However, since for large n the average squared error loss would be approximately equal to the integrated squared loss with $w(\cdot)$ being the density function of x_i's. Thus one can expect that an estimator performing well under the average squared error loss would also do well under the integrated squared error loss with the correct $w(\cdot)$. For the Stein estimate constructed in this paper, we can easily interpolate to other x values by using any spline interpolation method (e.g., connecting by line segments), just like one can use spline interpolation in any kernel (window) estimates, nearest neighbor estimates, or spline estimates. Our restriction to the average squared error loss is mainly to avoid the more complicated numerical analysis involved in doing the interpolation.

Remark 5. To select a good smoothing parameter (k in Examples 1 and 2, or h_n in Example 3), one may want to choose the one which minimizes the unbiased estimate of the risk of $\hat{\underset{\sim}{f}}_n$. This turns out to be related with the generalized cross-validation method; for details, see Li (1983).

Appendix

Proof of Proposition. As in Stein (1981), we have

$$E||\hat{\theta} - \theta||^2 = E||y - \frac{1}{y'By} Ay - \theta||$$

$$= n + E_{\theta}\left\{\frac{y'A'Ay}{(y'By)^2} - \frac{2 \text{ trace } A}{y'By} + \frac{4y'A'By}{(y'By)^2}\right\}$$

$$= n + E_{\theta}\left\{\frac{r^2}{y'A'Ay} - \frac{2r \text{ trace } A}{y'A'Ay} + \frac{4r(y'A'A'Ay)}{(y'A'Ay)^2}\right\}.$$

In view of (2.10), it suffices to show that

$$\frac{y'A'A'Ay}{y'A'Ay} \leq \lambda(A).$$

To establish this inequality, observe that

$$\max_{y} \frac{y'A'A'Ay}{y'A'Ay} \leq \max_{Z} \frac{Z'A'Z}{Z'Z} = \lambda\left\{\frac{A'+A}{2}\right\}.$$ □

Proof of Lemma 1. Since the maximum eigenvalue of A'A is no less than the maximum eigenvalue of (A'A+AA')/2, the desired result follows from the fact that $\frac{A'A+AA'}{2} - \left(\frac{A'+A}{2}\right)^2$ is nonnegative definite.

Proof of Lemma 2. For any $\underset{\sim}{y} = (y_1,\ldots,y_n)' \in R^n$, define $\bar{y}_i = \sum_{j=0}^{k} w_{jn}(y_{i-j} + y_{i+j})$. Clearly, $\bar{y}_i^2 \leq \sum_{j=0}^{k} w_{jn}(y_{i-j}^2 + y_{i+j}^2)$. Thus $||M_n \underset{\sim}{y}||^2 = \sum_{i=1}^{n} \bar{y}_i^2 \leq \sum_{i=1}^{n} \sum_{j=o}^{k} w_{jn}(y_{i-j}^2 + y_{i+j}^2) = \sum_{i=1}^{n} y_i^2$. This implies that the maximum eigenvalue of M_n^2 is no greater than 1. □

Proof of Lemma 3. This will be similar to the proof of Lemma 1. Defining $\bar{y}_i$ by (3.7), we have $||M_n \underset{\sim}{y}||^2 = \sum_{i=1}^{n} \bar{y}_i^2 \leq \sum_{i=1}^{n} \sum_{j=1}^{k} w_{jn} y_{i(j)}^2 \leq 2 \sum_{i=1}^{n} y_i^2$. This implies that the maximum eigenvalue of $M_n'M_n$ is no greater than 2. □

Proof of Lemma 4. Let $\rho = \sum_{i=1}^{n-1} \rho_{in}$. Then trace $(I-M_n)-\lambda\{I-M_n\}$ =

$$h_n \cdot \sum_{i=1}^{n-1} \frac{\rho_{in}}{1+h_n\rho_{in}} \geq h_n \cdot \sum_{i=1}^{n-1} \frac{\rho_{in}}{1+h_n \cdot \rho} = h_n \cdot \frac{\rho}{1+h_n \cdot \rho} > h_n \cdot \frac{\rho_{nn}}{1+h_n\rho_{nn}} .$$ □

Acknowledgements. We thank Professor James Berger for very helpful discussions on this work and a referee for useful suggestions that lead to a better presentation.

References

Agarwal, G.G. and Studden, W.J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. Ann. Statist. 8, 1307~1325.

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. Ann. Statist. 8, 716~761.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Numer. Math. 31, 377-404.

Demmler, A. and Reinsch, C. (1975). Oscillation matrices with spline smoothing. Numer. Math. 24, 375-382.

James, W. and Stein, C. (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Probab. 1, 361-380. Univ. California Press.

Li, K.C. (1983). From Stein's unbiased risk estimate to the mathod of generalized cross-validation. Technical report. Department of Statistics, Purdue University.

Priestley, M.B. and Chao, M.T. (1972). Non-parametric function fitting. J. Roy. Statist. Soc. Ser. B 34, 385-392.

Reinsch, C. (1967). Smoothing by Spline functions. Numer. Math. 10, 177-183.

Rice, J. and Rosenblatt, M. (1981). Integrated mean square error of a smoothing spline. J. Approx. Th., 33, 353~369.

Shinozaki, N. (1984). Simultaneous estimation of location parameters under quadratic loss. To appear in Ann. Statist.

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. Unpublished Manuscript.

Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. Ann. Statist. 8, 240-246.

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

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

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. Ann. Statist. 8, 1348∼1360.

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

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.