A Taylor Series Derivation of Stein Estimators

by

George Casella

and

William Strawderman

BU-1052-MA

December 1991

BU-1052-M

A TAYLOR SERIES DERIVATION OF STEIN ESTIMATORS

George Casella[1]
Cornell University

and

William E. Strawderman[2]
Rutgers University

Part of this research was performed at the Cornell University Workshop on Conditional Inference sponsored by the Army Mathematical Sciences Institute and The Statistics Center, Cornell University, Ithaca, New York, June 3-14, 1991.

Key Words and Phrases: Minimax estimation, spherically symmetric, risk.

AMS 1980 Subject Classification: 62F10, 62C20.

[1]Research supported by National Science Foundation Grant DMS 91-00839 and NSA Grant 90F-073.

[2]Research supported by National Science Foundation Grant DMS 90-23172.

Summary

A number of now-famous conditions for dominance in the subject of Stein estimation have roots in the work of Stein. Unfortunately, the origination of these conditions are not often easily understood. We show here that a straightforward approach, using Taylor series, leads to the derivation of a number of these conditions. While our derivations are not rigorous in terms of giving explicit conditions for domination, they are suggestive of estimators that have been shown to dominate the usual estimators for spherically symmetric distributions. They further suggest that such estimators can be expected to perform well in the general location parameter case.

1. Introduction

It is well known that the sample mean is an inadmissible estimator of the population mean in a wide variety of circumstances. Furthermore, this fact has spawned an incredible amount of research, ranging from theoretical investigations of differential inequalities to applications of empirical Bayes data analysis. The question we ask here goes back to the roots of this subject, to the first paper on Stein estimation, Stein (1956). In that paper, a first form of the now famous James-Stein estimator is given, using arguments (summarized below) that are somewhat esoteric. Our purpose is to see if we could use a more mundane approach, using Taylor series approximations, to arrive at the same result.

Suppose we have one observations, x, on a p-variate random vector X with $E_\theta X = \theta$ and $\mathrm{Cov} X = I$. We want to estimate θ with an estimator $\delta(X)$ using squared error loss

$$L(\theta, \delta(X)) = \|\theta - \delta(X)\|^2 = \sum_{i=1}^{p} (\theta_i - \delta_i(X))^2 \tag{1.1}$$

and associated risk function $R(\theta, \delta) = E_\theta L(\theta, \delta(X))^2$.

In the normal case, it had been established in 1951 (Hodges and Lehmann (1951)), that X is admissible if $p = 1$, but the cases $p \geq 2$ were not answered. Assuming that X was inadmissible, how might an alternative estimator $X + g(X)$, which might dominate X, be derived? The derivation given by Stein (1956), in that first landmark paper, asserts that in the normal case we can write

$$\|X\|^2 = \|X - \theta\|^2 + \|\theta\|^2 + 2\sqrt{\|\theta\|^2}\, Z, \tag{1.2}$$

where $Z = \theta'(X - \theta)/\sqrt{\|\theta\|^2}$ is univariate normal. Then, for large p, we have from (1.2)

$$\|X\|^2 = p + \|\theta\|^2 + O_p(\sqrt{p + \|\theta\|^2}), \tag{1.3}$$

showing that $\|\theta\|^2 \approx \|X\|^2 - p$, and the estimator X should be cut down by a factor of at least $[(\|X\|^2 - p)/\|X\|^2]^{1/2}$, which Stein then modifies to $(\|X\|^2 - p)/\|X\|^2$.

This reasoning seems quite deep to us and, no doubt, there are subtleties that are not explained in Stein's arguments. We wanted to know if, starting from the estimator $X + g(X)$, we could use self-evident arguments to deduce the James-Stein estimator. The self-evident argument is Taylor series approximations, and we will see that this argument leads

to many well-known dominance conditions. While our derivations are only approximations and do not give rigorous conditions for domination of X by $X + g(X)$, they are suggestive of estimators that have been shown to dominate in the spherically symmetric case by Brandwein and Strawderman (1991). The results suggest that such estimators can be expected to perform well in the general case.

There are many derivations of Stein-type estimators. In particular, there is the empirical Bayes derivation of Efron and Morris (1973), and the tail-minimaxity conditions of Berger (1976). Although these derivations are reasonably straightforward, they are not as self-evident as the Taylor series approach.

In Section 2, we describe the approximation in the simplest case ($p = 1$), and then apply it in Section 3 to the multivariate case. Section 4 comments on the relationship of the approximations with known exact results.

2. Univariate Taylor Series Approximations

For simplicity, first look at the (futile) case $p = 1$. An estimator $X + g(X)$ has loss

$$|\theta - (x + g(x))|^2 = |\theta - x|^2 - 2g(x)(\theta - x) + g^2(x). \tag{2.1}$$

Now consider the first-order Taylor expansions of $g(x)$ and $g^2(x)$ around θ. We have

$$\begin{aligned} g(x) &\approx g(\theta) + g'(\theta)(x - \theta), \\ g^2(x) &\approx g^2(\theta) + g^{2'}(\theta)(x - \theta). \end{aligned} \tag{2.2}$$

Substituting (2.2) into (2.1) yields

$$\begin{aligned} |\theta - (x + g(x))|^2 \approx |\theta - x|^2 &+ 2g'(\theta)(x - \theta)^2 + g^2(\theta) \\ &+ [2g(\theta) + g^{2'}(\theta)](x - \theta), \end{aligned} \tag{2.3}$$

and taking expectations shows that the estimator $X + g(X)$ improves on X (to this order of approximation) if $g(\cdot)$ satisfies

$$2g'(\theta) + g^2(\theta) \leq 0 \qquad \text{for all } \theta. \tag{2.4}$$

Here we have used the facts that $E(X - \theta)^2 = 1$ and the expectation of the last term in (2.3) is zero.

The differential inequality of (2.4), and its generalization, as in (3.4), have been the focus of much research (see, for example, Stein (1981) and references in Brandwein and Strawderman (1990)). Continuing in our naive mode, we will search among simple functions for a solution. A simple class of functions are of the form

$$g(t) = at^k, \tag{2.5}$$

for some constant a and integer k. Substituting in (2.4), we need

$$2(kat^{k-1}) + a^2t^{2k} \leq 0 \quad \text{for all } t. \tag{2.6}$$

Inspection of (2.6) shows that the inequality can only be satisfied if $ka < 0$. Furthermore, to satisfy the inequality at both $t = 0$ and $t = \infty$, we need $k - 1 = 2k$, or $k = -1$. Thus, our only chance to satisfy (2.4) with a function of the form (2.5) is to have $k = -1$ and the constant a satisfy $-2a + a^2 \leq 0$. Although this is a dead end for $p = 1$, it suggests that estimators of the form

$$X + g(X) = X + \frac{a}{X} = \left(1 + \frac{a}{X^2}\right)X \tag{2.7}$$

might be reasonable alternatives to the sample mean. Unfortunately, all such estimators have infinite risk, although modifications such as positive part versions of (2.7) have reasonable risk behavior. Of course, none beat X since X is admissible.

3. Multivariate First Order Taylor Series Approximations

The univariate expansions in the previous section can be easily extended to the multivariate case, with constructive results. The loss of estimating the p-vector θ with $X + g(X)$ is

$$\|\theta - (x + g(x))\|^2 = \|\theta - x\|^2 - 2g(x)'(\theta - x) + \|g(x)\|^2. \tag{3.1}$$

We now use the multivariate Taylor expansions of $g(x)$ and $\|g(x)\|^2$ to obtain

$$\begin{aligned} g(x) &\approx g(\theta) + D_g(x - \theta), \\ \|g(x)\|^2 &\approx \|g(\theta)\|^2 + \sum_{i=1}^{p}\left[\frac{\partial}{\partial\theta_i}\|g(\theta)\|^2\right](x_i - \theta_i), \end{aligned} \tag{3.2}$$

where D_g is the Jacobian matrix with (i,j) element $\{\frac{\partial}{\partial\theta_i}g_j(\theta)\}$. Substituting (3.2) into (3.1) and taking expectations yields

$$E_\theta\|\theta-(X+g(X))\|^2 \approx p+2\nabla\cdot g(\theta)+\|g(\theta)\|^2, \tag{3.3}$$

where $\nabla\cdot g(\theta)=\sum_{i=1}^p \frac{\partial}{\partial\theta_i}g_i(\theta)$. Analogous to (2.4), we need

$$2\nabla\cdot g(\theta)+\|g(\theta)\|^2 \le 0 \quad \text{for all } \theta, \tag{3.4}$$

in order for $X+g(X)$ to dominate X (to this order of approximation). Taking the multivariate analog of (2.7), we might try $g(t)=at/\|t\|^2$. This yields

$$\nabla\cdot g(t)=\frac{a(p-2)}{\|t\|^2}, \qquad \|g(t)\|^2=\frac{a^2}{\|t\|^2}, \tag{3.5}$$

and (3.4) is satisfied if $-2(p-2)\le a\le 0$, yielding the James-Stein estimator. Condition (3.4) of course is the now well-known condition for domination of X by $X+g(X)$ in the normal case given in Stein (1981).

Chou and Strawderman (1990) established minimaxity of estimators of the form $X+ag(X)$ where $g(X)$ satisfies (3.4) for distributions which are mixtures of normal distributions of the form $f(\|X-\theta\|^2)=\int N(\theta,\sigma^2 I)dG(\sigma^2)$.

4. Relationship with Exact Conditions

If X has a multivariate normal distribution, that is, $X\sim N_p(\theta,I)$, then the estimator $X+g(X)$ is minimax if condition (3.4) is satisfied. Thus, the first-order Taylor series approximations agree with the exact results under normality. This correspondence might be anticipated, as results of the normal distribution are often related to linearity considerations.

The first-order approximation is not good enough for other distributions, so we might try a second-order approximation for $\|g(x)\|^2$. (Note that the first-order expansion for $g(x)$ in (3.2) yields a second-order expansion on $g(x)'(x-\theta)$ in (3.1).) The second-order expansion for $\|g(x)\|^2$ is

$$\begin{aligned}\|g(x)\|^2 = \|g(\theta)\|^2 &+ \sum_{i=1}^p\left[\frac{\partial}{\partial\theta_i}\|g(\theta)\|^2\right](x_i-\theta_i)\\ &+\frac{1}{2}\sum_{i=1}^p\sum_{j=1}^p\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\|g(\theta)\|^2\right](x_i-\theta_i)(x_j-\theta_j),\end{aligned} \tag{4.1}$$

yielding the risk approximation.

$$(4.2) \qquad E_\theta\|\theta - (X + g(X))\|^2 \approx P + 2\nabla \cdot g(\theta) + \|g(\theta)\|^2 + \tfrac{1}{2}\nabla^2\|g(\theta)\|^2,$$

where $\nabla^2\|g(\theta)\|^2 = \sum_{i=1}^{p} \frac{\partial^2}{\partial\theta_i^2}\|g(\theta)\|^2$. From (4.2), we can see that approximate risk dominance will obtain if (3.4) is satisfied and, in addition,

$$(4.3) \qquad \nabla^2\|g(\theta)\|^2 \leq 0 \quad \text{for all } \theta.$$

Condition (4.3), that $\|g(\theta)\|^2$ is *superharmonic*, is thus seen to be a suggested condition for minimaxity in addition to the Stein differential inequality (3.4).

Stein (1981) showed that superharmonicity of the *prior* distribution in the normal case implied minimaxity of the corresponding generalized Bayes estimator. George (1986) used this condition to obtain minimax multiple shrinkage estimators.

That superharmonicity of $\|g\|^2$ itself is a useful condition in establishing minimaxity of $X + g(X)$ has been shown by Brandwein and Strawderman (1991). They show, for X distributed spherically symmetrically, that a sufficient condition for dominance of X by $X + g(X)$ is that $\|g\|^2$ be superharmonic, $g(X)$ satisfy the Stein differential inequality and $0 < a < \frac{1}{pE_0(1/\|X\|^2)}$ plus a technical condition on the monotonicity of $E(\|X - \theta\|^2 g(X) | \|X - \theta\|^2 = R)$. Note that the ordinary James-Stein estimator, which uses $g(\cdot)$ of (3.5), is superharmonic if and only if $p \geq 4$.

Our approximation indicates that such conditions are likely to be helpful in establishing dominance under milder conditions on the distribution of X.

REFERENCES

Berger, J. (1976). "Tail minimaxity in location vector problems and its applications." *Ann. Statist.* **4**, 33-50.

Brandwein, A. C. and Strawderman, W. E. (1990). "Stein estimation: The spherically symmetric case." *Statistical Science* **5**, 356-369.

Brandwein, A. C. and Strawderman, W. E. (1991). "Generalizations of James-Stein estimators under spherical symmetry." *Ann. Statist.* **19**, 1639-1650.

Chou, J. P. and Strawderman, W. E. (1990). "Minimax estimation of means of multivariate normal mixtures." *J. Mult. Anal.* **35**, 141-150.

Efron, B. and Morris, C. (1973). "Stein's estimation rule and its competitors–an empirical Bayes approach." *J. Amer. Statist. Assoc.* **68**, 117-130.

George, E. I. (1986). "Minimax multiple shrinkage estimation." *Ann. Statist.* **14**, 188-205.

Hodges, J. L. and Lehmann, E. L. (1951). "Some applications of the Cramer-Rao inequality." *Proc. Second Berkeley Symp.*, 13-22.

Stein, C. (1956). "Inadmissibility of the usual estimator of the mean of a multivariate normal distribution." In *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 197-206. University of California Press, Berkeley.

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