Statistical Science
2012, Vol. 27, No. 1, 24–30
DOI: 10.1214/11-STS382

A Geometrical Explanation of Stein Shrinkage

Lawrence D. Brown and Linda H. Zhao

Abstract. Shrinkage estimation has become a basic tool in the analysis of high-dimensional data. Historically and conceptually a key development toward this was the discovery of the inadmissibility of the usual estimator of a multivariate normal mean.

This article develops a geometrical explanation for this inadmissibility. By exploiting the spherical symmetry of the problem it is possible to effectively conceptualize the multidimensional setting in a two-dimensional framework that can be easily plotted and geometrically analyzed. We begin with the heuristic explanation for inadmissibility that was given by Stein [In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I* (1956) 197–206, Univ. California Press]. Some geometric figures are included to make this reasoning more tangible. It is also explained why Stein's argument falls short of yielding a proof of inadmissibility, even when the dimension, p, is much larger than $p = 3$.

We then extend the geometric idea to yield increasingly persuasive arguments for inadmissibility when $p \geq 3$, albeit at the cost of increased geometric and computational detail.

Key words and phrases: Stein estimation, shrinkage, minimax, empirical Bayes, high-dimensional geometry.

1. INTRODUCTION

More than 50 years ago Stein (1956) published his classic paper, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." The title result is probably the most startling statistical discovery of the past century. Erich Lehmann, who also worked on the admissibility question, more recently described how he was "stunned with disbelief" when Charles first told him of this result (personal communication). Following the initial discovery James and Stein (1961) presented their well-known shrinkage estimator that provides numerically significant improvement of risk relative to that of the usual estimator.

[Hodges and Lehmann (1951) and Girshick and Savage (1951) had earlier provided proofs of admissibility in the unidimensional problem; Lehmann's student Blyth (1951) had published another, more general, argument for this same fact; and Lehmann and Stein (1953) had produced a proof of admissibility in a related one-dimensional hypothesis testing setting.]

Stein (1956) begins by describing the multivariate problem and then gives a heuristic, geometric argument intended to convince that the usual estimator should be inadmissible if the dimension is sufficiently large. The core of this argument will be repeated below, with some additional illustrations that hopefully help to clarify the situation. The argument given by Stein provides insight into why inadmissibility occurs in very high-dimensional problems. But it does not provide a rationale for the fact that 3 is the critical dimension—admissibility holds in dimension 1 and 2 but not in three or more dimensions. [Section 4 of Stein (1956) contains an admissibility proof for two dimensions. See also Brown (1971) and Brown and Fox (1974).]

Lawrence D. Brown is Miers Busch Professor, Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19010-6340, USA (e-mail: lbrown@wharton.upenn.edu). Linda H. Zhao is Professor, Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19010-6340, USA.

The argument in the following note expands Stein's original heuristic idea, clarifies the geometry, and provides justification for the fact that 3 is the critical dimension. The argument is based on plane geometry and some simple "back-of-the-envelope" Taylor series expansions. As with Stein's argument, what is given here is not a proof. It could undoubtedly be expanded into a proof, but without further insight that proof would likely be similar to—and perhaps harder than—the existing inadmissibility proofs in Stein (1956) and Brown (1966). A slightly different geometrically based argument is suggested in Stein (1962) and is additionally expanded in Brandwein and Strawderman (1990). This argument is mentioned in Section 3.

Versions of this argument were presented in the 1960s in oral form independently by L. Brown, by B. Efron, and perhaps by others. But so far as we know the argument here does not appear in print. In addition, we feel it is worthwhile to remind readers of the geometric rationale underpinning Stein shrinkage in a form that displays that 3 is the critical dimension.

2. THE ADMISSIBILITY PROBLEM

Let $\mathbf{X} = (X_1, \ldots, X_p)'$ where X_i, $i = 1, \ldots, p$, are independent normal variables with unknown means $\theta_1, \ldots, \theta_p$ and all with the same known variance, σ^2. Without loss of generality, assume $\sigma^2 = 1$. It is desired to estimate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ with the quality of an estimate being measured through squared error loss, $L(d, \theta) = \|d - \theta\|^2 = \sum(d_i - \theta_i)^2$. Let $\delta = \delta(\mathbf{X})$ denote an estimator. The risk function of δ is denoted by $R(\theta; \delta) = E_\theta(L(\delta(\mathbf{X})))$.

The "usual" estimator of $\boldsymbol{\theta}$ is $\mathbf{X}$ itself, that is, $\delta_0(\mathbf{X}) = \mathbf{X}$. This estimator is intuitive and has several appealing formal properties such as minimaxity, best-invariance, maximum likelihood, etc. [See standard textbooks such as Lehmann and Casella (1998) for discussion of these properties.]

Prior to Stein (1956) it had been firmly conjectured that δ_0 is admissible for any value of p. Admissibility means that there is no other estimator that is better in the sense of risk—formally, that there is no estimator δ' such that $R(\boldsymbol{\theta}; \delta') \le R(\boldsymbol{\theta}; \delta_0)$ with strict inequality at some value of $\boldsymbol{\theta}$. [Actually, though it is not important in the sequel, we note that a well-known supplementary argument shows that δ_0 is inadmissible if and only if there is another estimator that is always strictly better in the sense that $R(\boldsymbol{\theta}; \delta') < R(\boldsymbol{\theta}; \delta_0)$ for all $\boldsymbol{\theta}$.]

What Stein proved in Sections 2–4 of Stein (1956) is:

THEOREM (Stein). *δ_0 is admissible if and only if $p \le 2$.*

Our goal is to explain why δ_0 is inadmissible when $p \ge 3$.

3. SPHERICAL SYMMETRY

A spherically symmetric estimator is one that satisfies

$$\delta(\mathbf{X}) = \tau(\|\mathbf{X}\|)\mathbf{X} \tag{1}$$

for some scalar function, τ. Of course, δ_0 is spherically symmetric. We confine the search for alternatives to δ_0 to the collection of spherically symmetric estimators. Geometrically, these are estimators that lie on the line through $\mathbf{X}$, and whose distance from the origin depends on $\|\mathbf{X}\|$. Such an estimator is given as in (1)–(3).

The restriction to spherically symmetric alternatives is intuitively plausible. To support this intuition, Stein (1956), Section 3, contains a formal proof that δ_0 is inadmissible if and only if there is a spherically symmetric estimator which is better.

Once one has decided to restrict consideration only to spherically symmetric estimators it is possible to correctly plot and study the multivariate problem in a two- dimensional coordinate framework for the sample space. One coordinate measures the sample in the direction of the true parameter, θ; the other coordinate is the length of the orthogonal residual from this direction. This leads to the geometric picture developed in the following section.

4. GEOMETRY FOR SPHERICALLY SYMMETRIC ESTIMATORS

Only spherically symmetric estimators need to be considered. For such estimators relevant distributions depend only on the magnitude of $\boldsymbol{\theta}$; the direction of the vector $\boldsymbol{\theta}$ does not matter. Formally, this means that after the constraint to spherically symmetric estimators it suffices to consider the situation when $\boldsymbol{\theta}$ lies on the θ_1-axis. So, assume $\boldsymbol{\theta} = (\vartheta, 0, \ldots, 0)'$. Let $X = (X_1, X_{(2)}')'$ where $X_{(2)} \in \Re^{p-1}$. Geometrically, $X_{(2)}$ is the residual of $\mathbf{X}$ after projection on the direction determined by $\boldsymbol{\theta}$. Again, only the length of $X_{(2)}$ matters, not its direction in the hyperplane perpendicular to $\boldsymbol{\theta}$. Hence, let $R = \|X_{(2)}\|$. The relevant statistics for the observed sample can thus be rewritten as

$$\mathbf{Z} = (X_1, R) \quad \text{with } X_1 \sim N(\vartheta, 1), R^2 \sim \chi^2_{p-1} \quad \text{and } X_1, R \text{ are independent.} \tag{2}$$

Spherically symmetric estimators as in (1) are expressed similarly in the $\mathbf{Z}$ coordinate system as

$$\delta(\mathbf{Z}) = \tau(\|\mathbf{Z}\|)\mathbf{Z}. \tag{3}$$

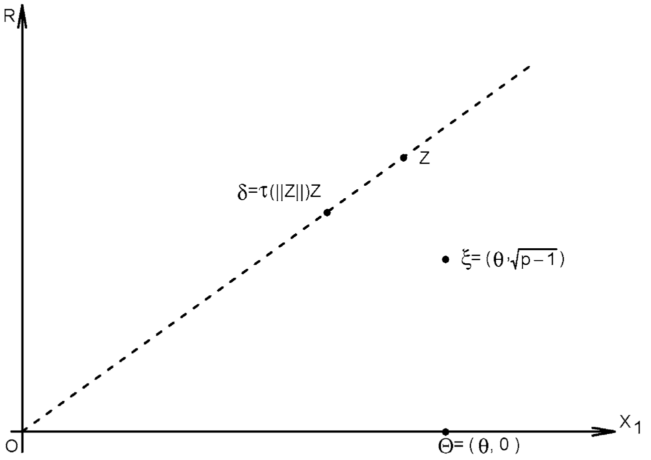

FIG. 1. *A typical observation in the* $\mathbf{Z} = (X_1, R)$ *coordinate system.*

The $\mathbf{Z}$ coordinate system is two-dimensional. Hence it can be conveniently visualized geometrically. A key feature of the transformation leading from the original, $\mathbf{X}$, system to the $\mathbf{Z}$ system is that distances are preserved. In particular, for spherically symmetric estimators

$$\|\delta(\mathbf{X}) - \theta\| = \|\delta(\mathbf{Z}) - (\vartheta, 0)\|.$$

Thus the squared error risks are the same in the two problems.

Pictorially this can be plotted in standard planar coordinates, as pictured in Figure 1. Figure 1 shows a typical observation of $\mathbf{Z}$ in the (X_1, R) coordinate system. It also represents a spherically symmetric estimate corresponding to $\mathbf{Z}$, as given by formulas (1)–(3). Pay special attention to the fact that this estimator is on the line through $\mathbf{Z}$. Figure 1 also shows an additional point $\xi = (\xi_1, \xi_2) = (\vartheta, \sqrt{p-1})$. This represents the intuitive "center" of the distribution of $\mathbf{Z}$.

In terms of Figure 1 the statistical situation can be summarized as follows: You observe $\mathbf{Z}$ with distribution as specified above. You are constrained to use only spherically symmetric estimators that lie on the line from the origin through $\mathbf{Z}$, as shown in the plot. You want to find an estimator that is close to Θ in terms of squared distance. For the point shown on the plot it is fairly clear that there are spherically symmetric estimates that are better than just $\mathbf{Z}$ alone. The point δ shown on the plot is one such better estimate. The goal of the remainder of the paper is to substantiate that situations like that in the figure are *on average* sufficiently typical (at least when $p \geq 3$), and hence that appropriate shrinkage estimators are better than $\mathbf{Z}$ itself.

[Note that $p - 1 = E(R^2)$. Hence it makes sense to think of $\sqrt{p-1} = \xi_2$ as the center of the distribution of R. This is not exactly either the mean or median of R, but it is sufficiently close and is convenient for the following discussion. The exact mean of R is $E(R) = \sqrt{2}\Gamma(p/2)/\Gamma((p-1)/2)$. For $p = 5, 10, 17, 26$, respectively, this takes the values $E(R) = 1.850, 2.918, 3.938, 4.950$ as compared to the values $\xi_2 = \sqrt{p-1} = 2, 3, 4, 5$. Asymptotically, $E(R) = \sqrt{p-1} - 1/(4\sqrt{p-1}) + O((p-1)^{-3/2})$.]

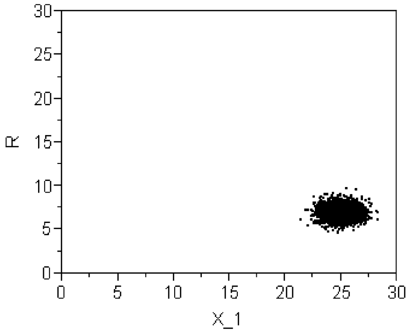

FIG. 2. 2000 *observations of* $\mathbf{Z}$ *in the case* $p = 20$ *and* $\vartheta = 25$.

Figure 2 shows a typical sample of 2000 observations of $\mathbf{Z}$ in the case $p = 20$ and $\vartheta = 25$. The dominant feature is that the sample points are moderately tightly clustered about $\xi = (25, \sqrt{19})$ and hence are much closer to ξ than they are to the parameter point $\boldsymbol{\theta} = (\vartheta, 0)$.

5. STEIN'S HEURISTIC ARGUMENT

It is fairly clear from pictures like Figure 2 that shrinking the observations somewhat toward the origin will often bring the estimator closer to the true mean $\boldsymbol{\theta} = (\vartheta, 0)$. Even more striking—consider what happens in a plot like Figure 2 as $p \to \infty$ for fixed $\boldsymbol{\theta} = (\vartheta, 0)$. Then the cloud of points moves vertically upward. Eventually, virtually the entire cloud lies outside the circle of radius $\|\boldsymbol{\theta}\|$. To be more precise

$$\|X\|^2 = \|\boldsymbol{\theta}\|^2 + p + O_P(\sqrt{p}) \tag{4}$$

as $p \to \infty$ for any fixed $\boldsymbol{\theta}$. This asymptotic fact can be derived from the non-central chi-squared distribution of $\|X\|^2$ or from a simple Taylor approximation as is done in Stein's heuristic argument. Viewed another way, (4) says that

$$\begin{aligned} \|\boldsymbol{\theta}\| &= \sqrt{\|X\|^2 - p - O_P(\sqrt{p})} \\ &= \|X\| - \frac{p + O_P(\sqrt{p})}{2\|X\|}. \end{aligned} \tag{5}$$

Any observation that lies outside of the sphere of radius $\|\boldsymbol{\theta}\|$ can be brought closer to $\boldsymbol{\theta}$ by shrinking it toward the origin so as to lie on the sphere. (Actually,

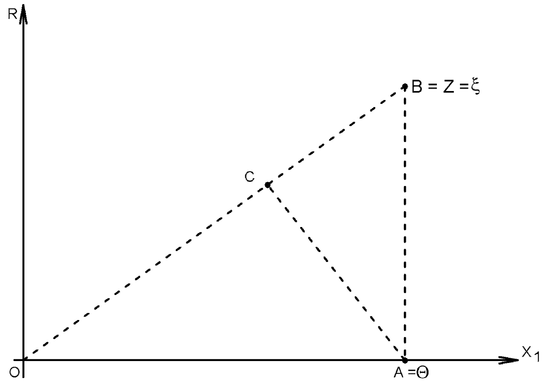

FIG. 3. *Geometry of the naïve optimal estimator: it shows the origin*, O, *the points* A $= (\vartheta, 0)$, B $= (\vartheta, \sqrt{p-1})$ *and* C, *the projection of* A *on the line* $\overline{\text{OB}}$.

somewhat more shrinkage is desirable as will be clear from the discussion of Figure 3, below.) This suggests that shrinkage by a factor $\frac{p+O_P(\sqrt{p})}{\|X\|}$ should be desirable as $p \to \infty$. The argument in Stein (1956) elaborates a little further and shows with a Taylor expansion that shrinkage by a factor $\frac{p+O(1)}{\|X\|}$ is still advantageous as $p \to \infty$ for any fixed $\boldsymbol{\theta}$. This motivates the use of the estimator $\delta_p(X) \triangleq (1 - \frac{p}{\|X\|^2})X$. This is related to what is used in Stein (1956) to prove inadmissibility of the usual estimator. The James–Stein (1961) estimator δ_{p-2} is better than the usual one when $p \geq 3$, as proved in that paper and later in a more efficient manner through Stein's unbiased estimate of risk in Stein (1973, 1981).

[Since $p \to \infty$ the difference between the factor $p/\|X\|^2$ in this argument and the factor $(p-2)/\|X\|^2$ in James and Stein (1961) is irrelevant. For fixed p it can be shown by the arguments mentioned above that δ_p dominates δ_0 whenever $p \geq 4$.]

Stein (1956) writes that "With some additional precision this [heuristic argument] could be made ...[in]to... a proof that for sufficiently large [p] the usual estimator is inadmissible." This is the type of exaggeration that may be excused by the above being only meant as a heuristic argument. In fact much more than "some additional precision" is needed to prove the usual estimator is inadmissible for sufficiently large p. The reason that the above does not easily yield a proof of inadmissibility is that it only holds for any fixed $\boldsymbol{\theta}$ as $p \to \infty$. It does not hold uniformly in $\boldsymbol{\theta}$, but a uniform argument is needed in order to prove inadmissibility.

To be more precise, for any p no matter how large, $\inf_{\boldsymbol{\theta}}\{P_{\boldsymbol{\theta}}(\|X\| \geq \|\boldsymbol{\theta}\|)\} = 1/2$, rather than approaching 1 as is implicitly suggested within the heuristic argument, and as would be needed to easily convert the heuristic argument into a proof.

Hence a more elaborate argument is needed to prove that the usual estimator is inadmissible. The following discussion presents a heuristic argument for inadmissibility that is consistent with the geometric insight in Stein's motivation.

6. DESIRED AMOUNT OF SHRINKAGE; TYPICAL OBSERVATION

Figures 1 and 2 show that the observations are close to $\xi = (\vartheta, \sqrt{p-1})$, whereas the estimate should be as close as possible to $\boldsymbol{\theta} = (\vartheta, 0)$. Figure 3 illustrates the geometry of this situation when $\mathbf{Z} = (\vartheta, \sqrt{p-1})$. It shows the origin (O), the point A $= \boldsymbol{\theta} = (\vartheta, 0)$ which is the desired target of the estimate, and the point B $= \xi = (\vartheta, \sqrt{p-1})$ which is a typical observation. For such an observation any spherically symmetric estimator must be on the line $\overline{\text{OB}}$. The point C in Figure 3 is the point on that line which is closest to the desired target, A. A similar triangles yield that

$$\frac{|\overline{\text{AB}}|}{|\overline{\text{OB}}|} = \frac{|\overline{\text{BC}}|}{|\overline{\text{AB}}|},$$

where $|\overline{\text{AB}}|$ denotes the length of the segment $\overline{\text{AB}}$, etc. Simplifying yields

$$|\overline{\text{BC}}| = \frac{|\overline{\text{AB}}|^2}{|\overline{\text{OB}}|} = \frac{p-1}{\|\xi\|}. \tag{6}$$

The point C is the best estimate based on an observation at B $= \xi$. By (6) it can be written as

$$\text{B} = \left(1 - \frac{p-1}{\|\xi\|^2}\right)\xi.$$

By comparison with (1), this suggests that the optimal spherically symmetric estimator will be the Naïve Geometrically Optimal estimator

$$\delta_{\text{NGO}}(\mathbf{Z}) = \left(1 - \frac{p-1}{\|\mathbf{Z}\|^2}\right)\mathbf{Z}. \tag{7}$$

The discussion leading to (7) suggests that

$$\delta_{p-1}(X) = \left(1 - \frac{p-1}{\|X\|^2}\right)X$$

should dominate δ_0. The above motivation and construction of δ_{NGO} does not suffer from the defect noted above in Stein's original heuristics—it does not require $p \to \infty$ for each fixed ϑ. However, it suggests that δ_0 is inadmissible even for $p = 2$. This suggestion is not correct; and so a more careful heuristic argument is needed to get a better description of the relevant geometry.

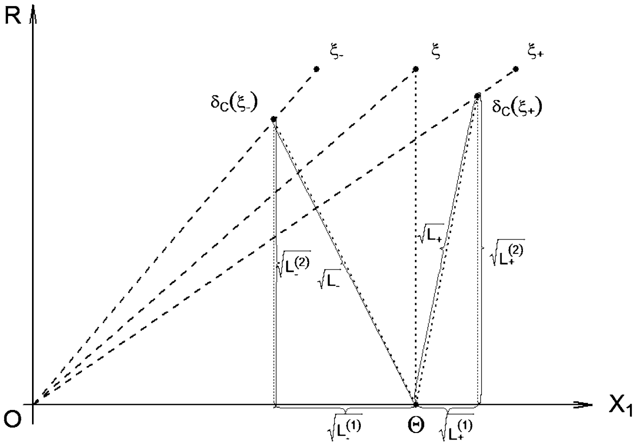

FIG. 4. *The values of $\xi_\pm$ and their respective estimates.*

7. STOCHASTIC VARIATION

The estimator δ_{NGO} in (7) is only optimal at $\xi = (\vartheta, \sqrt{p-1})$, the central point of the distribution of $\mathbf{Z}$. Of course, $\mathbf{Z}$ is not identically ξ, but is only stochastically close to ξ. The calculation leading to (7) is only approximate, not exact. There is a small price in accuracy to be paid in order to accommodate the stochastic variation of $\mathbf{Z}$. In order to better understand the composition of this price consider a particular pair of equally likely possible points for $\mathbf{Z}$. These points are labeled ξ_+, ξ_- in Figure 4. They are defined as

$$\xi_\pm = \big(\vartheta \pm 1, \sqrt{p-1}\big).$$

These points exhibit typical stochastic variation in the direction of $\theta = (\vartheta, 0)$ since their mean and mean squared distance in that direction match those of the full distribution. While they do not accurately model the stochastic variation in the direction orthogonal to $\theta = (\vartheta, 0)$, it turns out that this additional variability is only of secondary importance. Thus, we will ignore the effect of this orthogonal variation for now. It becomes clear from the exact expression discussed later at (14)–(15) that the orthogonal variation is indeed of secondary importance in calculation of the difference in risks.

Note that $L_+ < \|\xi_+ - \vartheta\|^2 = \|\xi_- - \vartheta\|^2$ but L_- can be $> \|\xi_- - \vartheta\|^2$. Calculations in the test show that $\frac{1}{2}(L_+ + L_-) < \|\xi_\pm - \vartheta\|^2$when $0 < C < 2(p-2)$.

In order to allow for additional discussion consider the general form

$$\delta_C(\mathbf{Z}) = \Big(1 - \frac{C}{\|\mathbf{Z}\|^2}\Big)\mathbf{Z}. \tag{8}$$

The case $C = p - 1$ is motivated by the preceding geometric argument. But the following calculations suggest that because of the stochastic variation modeled through ξ_+, ξ_- a preferable choice is $C = p - 2$, as in the ordinary James–Stein estimator.

Break down the risk into two components corresponding to the directions determined by the coordinates $\mathbf{Z} = (X_1, R)$. This is similar to the suggestion in Stein (1956), remark (vii). Related calculations are described in Efron and Morris (1971). Let $L_\pm$ denote the squared error from an observation at one of the two equally likely points ξ_+, ξ_- , respectively,

$$\begin{aligned} L_\pm &= \bigg[\Big(1 - \frac{C}{\|\xi_\pm\|^2}\Big)(\vartheta \pm 1) - \vartheta\bigg]^2 \\ &\quad + \bigg[\Big(1 - \frac{C}{\|\xi_\pm\|^2}\Big)\sqrt{p-1} - 0\bigg]^2 \\ &= \bigg[\pm 1 - \frac{C}{\|\xi_\pm\|^2}(\vartheta \pm 1)\bigg]^2 \\ &\quad + \Big(1 - \frac{C}{\|\xi_\pm\|^2}\Big)^2 (p-1) \\ &\triangleq L_\pm^{(1)} + L_\pm^{(2)}, \quad \text{say.} \end{aligned}$$

Let $R_{|\xi_\pm}(\boldsymbol{\theta}, \delta_C)$ denote the conditional risk given that $Z = \xi_+$ or ξ_-. Then

$$\begin{aligned} R_{|\xi_\pm} &= \tfrac{1}{2}\big(L_+^{(1)} + L_-^{(1)}\big) + \tfrac{1}{2}\big(L_+^{(2)} + L_-^{(2)}\big) \\ &\triangleq R_{|\xi_\pm.}^{(1)} + R_{|\xi_\pm.}^{(2)}, \quad \text{say.} \end{aligned}$$

Is δ_C better than δ_0 for this conditional problem? To examine this we look at the coordinate-wise difference in conditional risks. For δ_0 the coordinate-wise risks are 1 and $p - 1$, respectively. Hence the coordinate-wise differences are

$$\begin{aligned} 1 - R_{|\xi_\pm}^{(1)} = \frac{1}{2}\bigg(&\frac{2C(\vartheta + 1)}{\|\xi_+\|^2} - \frac{C^2(\vartheta + 1)^2}{\|\xi_+\|^4} \\ &- \frac{2C(\vartheta - 1)}{\|\xi_-\|^2} - \frac{C^2(\vartheta - 1)^2}{\|\xi_-\|^4}\bigg) \end{aligned}$$

and

$$\begin{aligned} &(p-1) - R_{|\xi_\pm}^{(2)} \\ &\quad = \frac{1}{2}(p-1)\bigg(\Big(\frac{2C}{\|\xi_+\|^2} + \frac{2C}{\|\xi_-\|^2}\Big) \\ &\qquad - \Big(\frac{C^2}{\|\xi_+\|^4} + \frac{C^2}{\|\xi_-\|^4}\Big)\bigg). \end{aligned}$$

In order to better interpret this expression rearrange terms so as to write the improvement of δ_C over δ_0 in this conditional problem as

$$
\begin{aligned}
\Delta_{|\xi_\pm} &\overset{\Delta}{=} p - R_{|\xi_\pm} \\
&= C\vartheta\Big(\frac{1}{\|\xi_+\|^2} - \frac{1}{\|\xi_-\|^2}\Big) \\
&\quad + Cp\Big(\frac{1}{\|\xi_+\|^2} + \frac{1}{\|\xi_-\|^2}\Big) \\
&\quad - \frac{1}{2}C^2\Big(\frac{(\vartheta+1)^2 + p - 1}{\|\xi_+\|^4} \\
&\qquad + \frac{(\vartheta-1)^2 + p - 1}{\|\xi_-\|^4}\Big) \\
&= C\vartheta\Big(\frac{1}{\|\xi_+\|^2} - \frac{1}{\|\xi_-\|^2}\Big) \\
&\quad + \Big(Cp - \frac{1}{2}C^2\Big)\Big(\frac{1}{\|\xi_+\|^2} + \frac{1}{\|\xi_-\|^2}\Big)
\end{aligned}
\tag{9}
$$

since $\xi_\pm = (\vartheta \pm 1)^2 + p - 1$.

If it were so that $\|\xi_+\|^2 = \|\xi_-\|^2$ *then* the first major term on the right of (9) would be =0, and the difference in (9) would be positive for any $0 < C < 2p$. In particular, for any $p \geq 2$ it *would be* positive for $C = p - 1$. (It could even be positive for $p = 1$!) This of course makes no sense as a statistical solution and only confirms that it provides an incorrect insight to ignore that $\|\xi_+\|^2 > \|\xi_-\|^2$.

Now, look at (9), and take into account that $\|\xi_+\|^2 > \|\xi_-\|^2$. Then, $\frac{1}{\|\xi_+\|^2} - \frac{1}{\|\xi_-\|^2} < 0$, and the first term on the right of (3.9) is negative and partially compensates for the remaining term which is positive when $C = p - 1$. In more detail,

$$
\begin{aligned}
\frac{1}{\|\xi_+\|^2} - \frac{1}{\|\xi_-\|^2} &= \frac{\|\xi_-\|^2 - \|\xi_+\|^2}{\|\xi_+\|^2\|\xi_-\|^2} \\
&= -4\frac{\vartheta}{\|\xi_+\|^2\|\xi_-\|^2},
\end{aligned}
$$

$$
\frac{1}{\|\xi_+\|^2} + \frac{1}{\|\xi_-\|^2} = \frac{\|\xi_-\|^2 + \|\xi_+\|^2}{\|\xi_+\|^2\|\xi_-\|^2} = 2\frac{\vartheta^2 + p}{\|\xi_+\|^2\|\xi_-\|^2}.
$$

Hence the difference in conditional risks for $p \geq 2$ is

$$
\begin{aligned}
\Delta_{|\xi_\pm} &= p - R_{|\xi_\pm} \\
&= \frac{2}{\|\xi_+\|^2\|\xi_-\|^2} \\
&\quad \cdot \Big(\Big(C(p-2) - \frac{C^2}{2}\Big)\vartheta^2 + \Big(Cp - \frac{C^2}{2}\Big)p\Big) \\
&> \frac{2(\vartheta^2 + p)}{\|\xi_+\|^2\|\xi_-\|^2}\Big(C(p-2) - \frac{C^2}{2}\Big).
\end{aligned}
\tag{10}
$$

It follows that the difference in conditional risks is positive so long as $0 < C < 2(p - 2)$. In particular the difference is positive for $p \geq 3$ and $C = p - 1$, the value motivated by the geometric argument centered on Figure 3. On the other hand, the best choice of constant in (10) is the slightly smaller value $C = p - 2$. The improvement in risks is not as great as that suggested in the argument around Figure 3, and this can be considered as a necessary penalty due to the randomness in **X**. In summary, the result in (10) provides a heuristic motivation for inadmissibility to hold whenever$p \geq 3$.

8. WHAT CAN BE PROVED

Note in (10) that the three terms in the leading fraction are all approximately equal; that is, $\vartheta^2 + p \approx \|\xi_+\|^2 \approx \|\xi_+\|^2$. Hence the argument leading to (10) suggests that the unconditional difference in risks, $\Delta = R(\boldsymbol{\theta}, \delta_0) - R(\boldsymbol{\theta}, \delta_C)$, will be well approximated as

$$
\begin{aligned}
\Delta &= R(\boldsymbol{\theta}, \delta_0) - R(\boldsymbol{\theta}, \delta_C) \\
&\approx \frac{2}{\|\boldsymbol{\theta}\|^2 + p}\Big(C(p-2) - \frac{C^2}{2}\Big).
\end{aligned}
\tag{11}
$$

The quality of this approximation improves as $\|\boldsymbol{\theta}\| \to \infty$ in the sense that

$$
\Delta \sim \frac{2}{\|\boldsymbol{\theta}\|^2 + p}\Big(C(p-2) - \frac{C^2}{2}\Big) \quad \text{as } \|\boldsymbol{\theta}\| \to \infty.
\tag{12}
$$

The preceding arguments can be refined to prove the assertion in (12). This is essentially the path followed by Stein in his original argument in Stein (1956). In order to allow calculations accurate only for large $\|\boldsymbol{\theta}\|$, Stein replaced δ_C with the estimator

$$
\delta_{C;a} = \Big(1 - \frac{C}{a + \|\mathbf{X}\|^2}\Big)\mathbf{X}.
$$

Then an exact Taylor expansion that can be considered as an elaboration of the above calculations yields

$$
\begin{aligned}
&R(\boldsymbol{\theta}, \delta_0) - R(\boldsymbol{\theta}, \delta_{C;a}) \\
&\quad = \frac{2}{a + \|\boldsymbol{\theta}\|^2}\Big(C(p-2) - \frac{C^2}{2}\Big) \\
&\qquad + o\Big(\frac{1}{a + \|\boldsymbol{\theta}\|^2}\Big)
\end{aligned}
\tag{13}
$$

uniformly in $\|\boldsymbol{\theta}\|$. It follows that δ_0 is inadmissible.

The argument in Stein (1956) for (13) involves only low-order moments of $\mathbf{X} - \boldsymbol{\theta}$. Hence it can be generalized from the normal distribution setting to apply to more general location parameter problems. It can also be adapted to apply (with modifications) to problems in which the loss function is not squared error. Such generalizations appear in Brown (1966).

When one considers only the normal distribution setting, then

$$\begin{aligned}\Delta &= R(\boldsymbol{\theta}, \delta_0) - R(\boldsymbol{\theta}, \delta_C) \\ &= E_\theta\left(\frac{1}{\|\mathbf{X}\|^2}\right)\left(C(p-2) - \frac{C^2}{2}\right).\end{aligned} \tag{14}$$

This result is proved but not explicitly stated in James and Stein (1961). It is explicitly stated and proved using the unbiased estimate of the risk in Stein (1973, 1981).

Note that

$$E_{\boldsymbol{\theta}}\left(\frac{1}{\|\mathbf{X}\|^2}\right) \approx \frac{1}{E_{\boldsymbol{\theta}}(\|\mathbf{X}\|^2)} = \frac{1}{\|\boldsymbol{\theta}\|^2 + p}, \tag{15}$$

with the approximation being quite close except when $\|\boldsymbol{\theta}\|$ is small. Hence the heuristic approximation in (10) and (11) is quite close to the truth. This validates the heuristic idea to approximate the unconditional difference in risks by the conditional difference given $\boldsymbol{\theta} = \xi_+, \xi_-$.

ACKNOWLEDGMENT

Research supported in part by NSF Grant DMS-10-07657.

REFERENCES

BLYTH, C. R. (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Statist.* **22** 22–42. MR0039966

BRANDWEIN, A. C. and STRAWDERMAN, W. E. (1990). Stein estimation: The spherically symmetric case. *Statist. Sci.* **5** 356–369.

BROWN, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.* **37** 1087–1136. MR0216647

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903. MR0286209 [Correction. *Ann. Statist.* **1** (1973) 594–596. MR0362592]

BROWN, L. D. and FOX, M. (1974). Admissibility in statistical problems involving a location or scale parameter. *Ann. Statist.* **2** 807–814. MR0370850

EFRON, B. and MORRIS, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators. I. The Bayes case. *J. Amer. Statist. Assoc.* **66** 807–815. MR0323014

GIRSHICK, M. A. and SAVAGE, L. J. (1951). Bayes and minimax estimates for quadratic loss functions. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 53–73. Univ. California Press, Berkeley. MR0045365

HODGES, J. L., JR. and LEHMANN, E. L. (1951). Some applications of the Cramér-Rao inequality. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 13–22. Univ. California Press, Berkeley. MR0044795

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.*, *Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

LEHMANN, E. L. and STEIN, C. M. (1953). The admissibility of certain invariant statistical tests involving a translation parameter. *Ann. Math. Statist.* **24** 473–479. MR0056249

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley. MR0084922

STEIN, C. (1973). Estimation of the mean of a multivariate normal distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics* (*Charles Univ.*, *Prague*, 1973), *Vol. II* 345–381. Charles Univ., Prague. MR0381062

STEIN, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. R. Statist. Soc. Ser. B Stat. Methodol.* **24** 265–296.

STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098