IMS Collections
Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková
Vol. 7 (2010) 25–34
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL703

The unbearable transparency of Stein estimation

Rudolf Beran

University of California, Davis

Abstract: Charles Stein [10] discovered that, under quadratic loss, the usual unbiased estimator for the mean vector of a multivariate normal distribution is inadmissible if the dimension n of the mean vector exceeds two. On the way, he constructed shrinkage estimators that dominate the usual estimator asymptotically in n. It has since been claimed that Stein's results and the subsequent James–Stein estimator are counter-intuitive, even paradoxical, and not very useful. In response to such doubts, various authors have presented alternative derivations of Stein shrinkage estimators. Surely Stein himself did not find his results paradoxical. This paper argues that assertions of "paradoxical" or "counter-intuitive" or "not practical" have overlooked essential arguments and remarks in Stein's beautifully written paper [10]. Among these overlooked aspects are the asymptotic geometry of quadratic loss in high dimensions that makes Stein estimation transparent; the asymptotic optimality results that can be associate with Stein estimation; the explicit mention of practical multiple shrinkage estimators; and the foreshadowing of Stein confidence balls. These ideas are fundamental for studies of modern regularization estimators that rely on multiple shrinkage, whether implicitly or overtly.

1. Introduction

In a profoundly prophetic paper that opened a new statistical world to exploration, Charles Stein [10] discovered, among other things, that the usual unbiased estimator for the mean of an n-dimensional multivariate normal distribution is inadmissible under quadratic loss if $n \geq 3$. It has since been claimed that Stein's results are counter-intuitive, even paradoxical. In response, Efron and Morris [6] presented an alternative empirical Bayes approach to Stein estimation. Stigler [13] gave another derivation based on a "Galtonian perspective". Fundamental results such as Stein's clearly merit rederivations that increase our understanding. But surely Stein himself did not find his results paradoxical. Is it not more likely that such claims merely overlook arguments and remarks in his pioneering paper [10]? This article briefly examines some of those arguments in the context of the paper's era and of later developments.

Sections 1 and 3 in Stein [10] presented the first of the paper's brilliant insights. Observed is the random n-vector X, whose distribution about the unknown mean vector ξ is n-dimensional normal with identity covariance matrix. A fuller notation would write X_n and ξ_n to express the dependence on n. We follow Stein [10] in not so doing. The quality of an estimator $\hat{\xi} = \hat{\xi}(X)$ of ξ is measured through its

[1]Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA, e-mail: beran@wald.ucdavis.edu
AMS 2000 subject classifications: Primary 62F12, 62J07; secondary 62-02.
Keywords and phrases: dimensional asymptotics, orthogonal equivariance.

normalized quadratic *loss* $n^{-1}|\hat{\xi}-\xi|^2$ and through the corresponding *risk* $R_n(\hat{\xi},\xi) = n^{-1}\mathrm{E}|\hat{\xi}-\xi|^2$, where $|\cdot|$ is Euclidean norm and E is expectation under the model. The risk of the usual unbiased estimator X is thus 1.

Suppose that $\lim_{n\to\infty}|\xi|^2/n = a < \infty$. By the weak law of large numbers, the following relations are very nearly true with high probability when the dimension n is large:

$$|n^{-1/2}\xi|^2 \approx a, \quad |n^{-1/2}X - n^{-1/2}\xi|^2 \approx 1, \quad |n^{-1/2}X|^2 \approx 1+a. \tag{1.1}$$

Asymptotically in n, we have a right-angled triangle in which, approximately, $n^{-1/2}X$ is the hypotenuse, $n^{-1/2}\xi$ is the base, and $n^{-1/2}X - n^{-1/2}\xi$ is the vector that joins base to hypotenuse. The angle θ between $n^{-1/2}\xi$ and $n^{-1/2}X$ is thus determined approximately by $\cos(\theta) \approx a^{1/2}/(1+a)^{1/2}$.

In seeking estimators of ξ that are admissible or minimax, it suffices to consider estimators equivariant under the orthogonal group on R^n.

This follows from the Hunt–Stein theorem and compactness of the orthogonal group. By Section 3 of Stein [10], every orthogonally equivariant estimator $\hat{\xi}(X)$ has the form

$$\hat{\xi}(X) = h(|X|)X \tag{1.2}$$

for some real-valued function h; it therefore lies along the vector X.

Under the asymptotic geometry of the previous paragraph, the orthogonal projection of $n^{-1/2}\xi$ onto $n^{-1/2}X$ defines an orthogonally equivariant oracle estimator $n^{-1/2}\hat{\xi}_O$ whose loss $|n^{-1/2}\hat{\xi}_O - n^{-1/2}\xi|^2$ is asymptotically minimal. For large n, $\hat{\xi}_O$ satisfies

$$n^{-1/2}\hat{\xi}_O = |n^{-1/2}\xi|\cos(\theta)X/|X| \approx [a/(1+a)]n^{-1/2}X. \tag{1.3}$$

Consider the *asymptotic Stein* estimator

$$\hat{\xi}_{AS} = [(|n^{-1/2}X|^2 - 1)/|n^{-1/2}X|^2]X = [1 - n/|X|^2]X. \tag{1.4}$$

By (1.1) and (1.3), $\hat{\xi}_{AS}$ asymptotically approximates $\hat{\xi}_O$ for *every* positive finite a. Consequently, under the asymptotics of the preceding two paragraphs and for every positive finite a, the estimator $\hat{\xi}_{AS}$ minimizes limiting loss, and hence risk, among all orthogonally equivariant estimators. By the geometry of the situation the minimized loss or risk is, with probability tending to one,

$$n^{-1}|\hat{\xi}_{AS} - \xi|^2 \approx n^{-1}|\hat{\xi}_O - \xi|^2 = |n^{-1/2}\xi|^2\sin^2(\theta) \approx a/(1+a). \tag{1.5}$$

This agrees with the evaluation that follows equation (8) of Stein [10].

In the Introduction to Stein [10], on p. 198, a geometrical rationale for $\hat{\xi}_{AS}$ was stated succinctly (notation adjusted): "It certainly seems more reasonable [in estimating ξ] to cut X down at least by a factor of $[(|X|^2 - n)/|X|^2]^{-1/2}$ to bring the estimate within the sphere. Actually, because of the curvature of the sphere combined with the uncertainty of our knowledge of ξ, the best factor, to within the approximation considered here, turns out to be $(|X|^2 - n)/|X|^2$." The phrasing indicates full awareness of the intuitive asymptotic geometry described above. It seems likely that few contemporaries shared this awareness.

Stein's penetrating asymptotic insights led to extensive later investigations for finite n. For instance, the James–Stein [8] estimator

$$\hat{\xi}_S = [1 - (n-2)/|X|^2]X \tag{1.6}$$

is a refinement of $\hat{\xi}_{AS}$ that is orthogonally equivariant, improves on the risk for $n \geq 2$, and also minimizes limiting loss as $n \to \infty$.

2. Optimality in the fixed length submodel

The preceding section showed heuristically that the James–Stein and asymptotic Stein estimators possess asymptotic optimality properties. These can be refined and proved by studying orthogonally equivariant estimators of ξ in detail, a project begun fruitfully in Section 3 of Stein [10] and continued here.

The orthogonal group is not transitive over the the full parameter space of the $N(\xi, I)$ model but is transitive in the fixed length submodel where $|\xi| = \rho_0$, a fixed known value, and only the direction vector $\mu = \xi/|\xi|$ is unknown. In this submodel, the conditional risk, given $|X|$, of any orthogonally equivariant estimator (1.2) is

$$n^{-1}[h^2(|X|)|X|^2 - 2h(|X|)\mathrm{E}(\xi' X||X|) + \rho_0^2]. \tag{2.1}$$

Let $\hat{\mu} = X/|X|$ denote the direction vector of X. The conditional distribution of $\hat{\mu}$ given $|X|$ is Langevin on the unit sphere in R^n, with mean direction $\mu = \xi/|\xi|$ and dispersion parameter $\kappa = \rho_0|X|$ (cf. Watson [14] for $n \geq 2$). Let $A_n(z) = I_{n/2}(z)/I_{n/2-1}(z)$ for $z \geq 0$, where $I_\nu(\cdot)$ is the modified Bessel function of the first kind and order ν. The choice of h that minimizes (2.1) is

$$h_0(|X|) = |X|^{-2}\mathrm{E}(\xi' X||X|) = \rho_0|X|^{-1}\mathrm{E}(\mu'\hat{\mu}||X|) = \rho_0|X|^{-1}A_n(\rho_0|X|). \tag{2.2}$$

The minimum risk orthogonally equivariant estimator of ξ is therefore

$$\hat{\xi}_E(\rho_0) = \rho_0 A_n(\rho_0|X|)\hat{\mu}, \qquad n \geq 1. \tag{2.3}$$

See Beran [2] for further details and references.

The foregoing considerations, compactness of the orthogonal group, and the Hunt–Stein theorem prove the following result:

- *In the fixed length submodel where $|\xi| = \rho_0$, the minimum risk orthogonally equivariant estimator of ξ is $\hat{\xi}_E(\rho_0)$, defined in (2.3). This estimator is minimax and admissible among all estimators of ξ.*

Another orthogonally equivariant estimator of ξ is

$$\hat{\xi}_{AE}(\rho_0) = (\rho_0^2/|X|)\hat{\mu}. \tag{2.4}$$

This estimator will be seen to approximate $\hat{\xi}_E(\rho_0)$ for large n and to have asymptotically the same risk. Exact calculations using (2.1) ultimately yield the following result:

- *In the fixed length submodel where $|\xi| = \rho_0$,*

$$R_n(\hat{\xi}_E(\rho_0), \xi) = n^{-1}E[\rho_0^2 - \rho_0^2 A_n^2(\rho_0|X|)] \tag{2.5}$$

$$R_n(\hat{\xi}_{AE}(\rho_0), \xi) = n^{-1}E[\rho_0^2 - 2\rho_0^3|X|^{-1}A_n(\rho_0|X|) + \rho^4|X|^{-2}]. \tag{2.6}$$

These exact risk expressions in the fixed length submodel have simple approximations as $n \to \infty$. This is to be expected from the informal asymptotics in the Introduction. For $t \geq 0$, define the function

$$r(t) = t/(1+t). \tag{2.7}$$

Note that $\lim_{n\to\infty} z_n = z \geq 0$ implies $\lim_{n\to\infty} z_n A_n(nz_n) = (z^2 + 1/4)^{1/2} - 1/2$. This limit together with (2.5) and (2.6) yield:

- *In the fixed length submodel where $|\xi| = \rho_0$ and for every finite $c > 0$,*

$$\lim_{n\to\infty} \sup_{\rho_0^2 \le nc} |R_n(\hat{\xi}_E(\rho_0), \xi) - r(\rho_0^2/n)| = 0 \tag{2.8}$$

$$\lim_{n\to\infty} \sup_{\rho_0^2 \le nc} |R_n(\hat{\xi}_{AE}(\rho_0), \xi) - r(\rho_0^2/n)| = 0. \tag{2.9}$$

The estimators $\hat{\xi}_E(\rho_0)$ and $\hat{\xi}_{AE}(\rho_0)$ are asymptotically equivalent in the sense that

$$\lim_{n\to\infty} \sup_{\rho_0^2 \le nc} E|\hat{\xi}_E(\rho_0) - \hat{\xi}_{AE}(\rho_0)|^2 = 0. \tag{2.10}$$

Beran [2] gave the proof details.

3. Asymptotic minimaxity: From Stein to Pinsker

The foregoing results for the fixed length submodel have powerful implications for estimation of ξ in the full $N(\xi, I)$ model. The first of these is an asymptotic lower bound on maximum risk over balls in the parameter space:

- *In the full $N(\xi, I)$ model, for every finite $c > 0$,*

$$\liminf_{n\to\infty} \inf_{\hat{\xi}} \sup_{|\xi|^2 \le nc} R_n(\hat{\xi}, \xi) \ge r(c), \tag{3.1}$$

the infimum being taken over all estimators $\hat{\xi}$.

This result follows easily from preceding considerations. Indeed, as Stein [10] pointed out, the estimation problem is invariant under the orthogonal group, which is compact. By the Hunt–Stein theorem,

$$\inf_{\hat{\xi}} \sup_{|\xi|^2 \le nc} R_n(\hat{\xi}, \xi) = \inf_{\hat{\xi}_I} \sup_{|\xi|^2 \le nc} R_n(\hat{\xi}, \xi), \tag{3.2}$$

the infimum on the right side being taken only over orthogonally equivariant estimators $\hat{\xi}_I$. Using the first bulleted result in the previous subsection on the fixed length model, with $\rho_0 = n^{1/2}c^{1/2}$,

$$\inf_{\hat{\xi}_I} \sup_{|\xi|^2 \le nc} R_n(\hat{\xi}, \xi) \ge \inf_{\hat{\xi}_I} \sup_{|\xi|^2 = nc} R_n(\hat{\xi}, \xi) = \sup_{|\xi|^2 = nc} R_n[\hat{\xi}_E(n^{1/2}c^{1/2}), \xi]. \tag{3.3}$$

Because of (2.8), the right side of (3.3) converges to $r(c)$, thereby establishing (3.1).

This result is actually an instance of Pinkser's [9] theorem on estimation of ξ. See Beran and Dümbgen [5] for a relevant statement of the latter. The argument above pursues ideas broached in Section 3 of Stein [10] rather than ideas in Pinsker's later, more general study of the problem through Bayes estimators.

To construct estimators that achieve the lower bound (3.1) for *every* $c > 0$, it suffices to construct a good estimator $\hat{\rho}$ of $|\xi|$ from X and then form the *adaptive* estimators

$$\hat{\xi}_E(\hat{\rho}) = \hat{\rho} A_n(\hat{\rho}|X|)\hat{\mu}, \qquad \hat{\xi}_{AE}(\hat{\rho}) = (\hat{\rho}^2/|X|)\hat{\mu}. \tag{3.4}$$

The following local asymptotic minimax result governs estimation of $|\xi|^2$:

- *In the full $N(\xi, I)$ model, for every finite $b > 0$,*

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{\hat\rho} \sup_{||\xi|^2/n-b|\le n^{-1/2}c} n^{-1}E(\hat\rho^2 - |\xi|^2)^2 \ge 2 + 4b, \tag{3.5}$$

the infimum being taken over all estimators $\hat\rho$. If $\hat\rho^2 = |X|^2 - n + d$ or $[|X|^2 - n + d]_+$, where d is a constant, then

$$\lim_{n\to\infty} \sup_{||\xi|^2/n-b|\le n^{-1/2}c} n^{-1}E(\hat\rho^2 - |\xi|^2)^2 = 2 + 4b \tag{3.6}$$

for every finite $c > 0$.

For a proof, see Beran [2]. A related treatment for estimators of $|\xi|$ was given by Hasminski and Nussbaum [7].

If $\hat\rho^2$ is $[|X|^2 - n + 2]_+$, then $\hat\xi_{AE}(\hat\rho)$ coincides with the positive-part James–Stein estimator and $\hat\xi_E(\hat\rho)$ is defined. The James–Stein estimator $\hat\xi_S$ is $\hat\xi_{AE}(\hat\rho)$ when $\hat\rho^2 = |X|^2 - n + 2$. This definition works formally even when $|X|^2 - n + 2$ is negative. For such $\hat\rho$, the asymptotic risks of the adaptive estimators in (3.4) are readily found:

- *In the full $N(\xi, I)$ model with $\hat\rho^2 = |X|^2 - n + d$ or $[|X|^2 - n + d]_+$, the following holds for every finite $c > 0$:*

$$\lim_{n\to\infty} \sup_{|\xi|^2\le nc} |R_n(\hat\xi_{AE}(\hat\rho), \xi) - r(|\xi|^2/n)| = 0. \tag{3.7}$$

Consequently,

$$\lim_{n\to\infty} \sup_{|\xi|^2\le nc} R_n(\hat\xi_{AE}(\hat\rho), \xi) = r(c), \tag{3.8}$$

for every finite $c > 0$. Hence, $\hat\xi_{AE}(\hat\rho)$ achieves the asymptotic minimax bound (3.1). The same conclusions hold for $\hat\xi_E(\hat\rho)$ when $\hat\rho^2 = [|X|^2 - n + d]_+$.

This result entails, in particular, that the James–Stein estimator $\hat\xi_S$ and the positive-part James–Stein estimator are both asymptotically minimax for ξ on balls about the origin. Such is not the case for the classical estimator X because

$$\lim_{n\to\infty} \sup_{|\xi|^2\le nc} R_n(X, \xi) = 1 > r(c) \tag{3.9}$$

for every $c > 0$.

4. Stein confidence sets

Remark (viii) on p. 205 of Stein [10] briefly stated: "Nevertheless it seems clear that we shall obtain confidence sets which are appreciably smaller geometrically than the usual disks centered at the sample mean vector." A method for constructing such confidence balls was described in the penultimate paragraph of Stein [12], in connection with a general conjecture. We describe how, asymptotically in n, Stein's method yields geometrically smaller confidence sets for ξ that are centered at the James–Stein estimator $\hat\xi_S$.

Consider confidence balls for ξ centered at estimators $\hat\xi = \hat\xi(X)$,

$$C(\hat\xi, \hat d) = \{x\colon |\hat\xi - x| \le \hat d\}. \tag{4.1}$$

The radius $\hat{d} = \hat{d}(X)$ is such that the coverage probability $\mathrm{P}(C(\hat{\xi}, \hat{d}) \ni \xi)$ under the model is exactly or asymptotically α. The geometrical size of $C(\hat{\xi}, \hat{d})$, viewed as a set-valued estimator of ξ, is measured by the *geometrical risk*

$$G_n(C(\hat{\xi}, \hat{d}), \xi) = n^{-1/2}\mathrm{E} \sup_{x \in C(\hat{\xi}, \hat{d})} |x - \xi| = n^{-1/2}\mathrm{E}|\hat{\xi} - \xi| + n^{-1/2}\mathrm{E}(\hat{d}). \tag{4.2}$$

This geometrical risk extends to confidence sets the quadratic risk criterion that supports Stein point estimation.

The classical confidence ball for ξ is

$$C_C = C(X, \chi_n^{-1}(\alpha)), \tag{4.3}$$

where the square of $\chi_n^{-1}(\alpha)$ is the α-th quantile of the chi-squared distribution with n degrees of freedom. C_C is a ball centered at X whose squared radius for large n is approximately $n + (2n)^{1/2}\Phi^{-1}(\alpha)$. Here Φ^{-1} denotes the quantile function of the standard normal distribution. From this and (4.2):

- *For every $\alpha \in (0, 1)$ and every $c > 0$,*

$$P(C_C \ni \xi) = \alpha \quad \text{for every } \xi. \tag{4.4}$$

$$\lim_{n \to \infty} \sup_{|\xi|^2 \le nc} |G_n(C_C, \xi) - 2| = 0. \tag{4.5}$$

Stein confidence balls for ξ have the form (4.1), with the James–Stein estimator $\hat{\xi}_S$ as center. To construct suitable critical values $\hat{d}$ in this case, consider the root

$$D_n(X, \xi) = n^{-1/2}\{|\hat{\xi}_S - \xi|^2 - [n - (n-2)^2/|X|^2]\}, \tag{4.6}$$

which compares the loss of the James–Stein estimator with an unbiased estimator of its risk. By orthogonal invariance, the distribution of $D_n(X, \xi)$ depends on ξ only through $|\xi|^2$ and can thus be written as $H_n(|\xi|^2)$. Let $\Rightarrow$ designate weak convergence of distributions. The triangular array central limit theorem implies:

- *Suppose that $\lim_{n\to\infty} |\xi|^2/n = a < \infty$. Then*

$$H_n(|\xi|^2) \Rightarrow N(0, \sigma^2(a)), \tag{4.7}$$

where

$$\sigma^2(t) = 2 - 4t/(1+t)^2 \ge 1. \tag{4.8}$$

It follows from (3.6) that $\hat{\rho}^2 = [|X|^2 - n + 2]_+$ is a good estimator of $|\xi|^2$ such that $\lim_{n\to\infty} \sup_{|\xi|^2 \le nc} \mathrm{P}[|\hat{\rho}^2/n - |\xi|^2/n| > \epsilon] = 0$ for every $c > 0$ and $\epsilon > 0$. This and (4.7) motivate approximating $H_n(|\xi|^2)$ by $N(0, \sigma^2(\hat{\rho}^2/n))$. The latter approximation and the definition (4.6) of $D_n(X, \xi)$ suggest the *asymptotic Stein confidence ball*

$$C_{SA} = C(\hat{\xi}_S, \hat{d}_A(\alpha)), \tag{4.9}$$

where

$$\hat{d}_A(\alpha) = [n - (n-2)^2/|X|^2 + n^{1/2}\sigma^2(\hat{\rho}^2/n)\Phi^{-1}(\alpha)]_+^{1/2}. \tag{4.10}$$

Asymptotic analysis establishes

- *For every* $\alpha \in (0,1)$ *and every* $c > 0$,

$$\lim_{n\to\infty} \sup_{|\xi|^2 \le nc} |P(C_{SA} \ni \xi) - \alpha| = 0 \tag{4.11}$$

and

$$\lim_{n\to\infty} \sup_{|\xi|^2 \le nc} |G_n(C_{SA}, \xi) - r_S(|\xi|^2/n)| = 0, \tag{4.12}$$

where

$$r_{SA}(t) = 2[t/(1+t)]^{1/2} < 2. \tag{4.13}$$

Like the classical confidence ball centered at X, the Stein confidence ball C_{SA} entered at $\hat{\xi}_S$ has correct asymptotic coverage probability α, uniformly over large compact balls about the shrinkage point $\xi = 0$. Comparing (4.12) with (4.5), the geometrical risk of C_{SA} is asymptotically smaller than that of C_C, particularly when ξ is near 0.

To obtain valid bootstrap critical values for Stein confidence sets requires care because the naive bootstrap fails. Define the *constrained length* estimator of ξ by

$$\hat{\xi}_{CL} = [1 - (n-2)/|X|^2]_+^{1/2} X. \tag{4.14}$$

The triangular array central limit theorem implies:

- *Suppose that* $\lim_{n\to\infty} |\xi|^2/n = a < \infty$. *Then, for* σ^2 *defined in* (4.8)

$$H_n(|\hat{\xi}_{CL}|^2) \Rightarrow N(0, \sigma^2(a)), \tag{4.15}$$

while

$$H_n(|X|^2) \Rightarrow N(0, \sigma^2(1+a)), \qquad H_n(|\hat{\xi}_S|^2) \Rightarrow N(0, \sigma^2(a^2/(1+a))), \tag{4.16}$$

the weak convergences all being in probability.

See Beran [1] for proof details.

In view of (4.7), the bootstrap distribution estimator $\hat{H}_B = H_n(|\hat{\xi}_{CL}|^2)$ converges weakly in probability to $H_n(|\xi|^2)$, as desired, while the naive bootstrap distribution estimators $H_n(|X|^2)$ and $H_n(|\hat{\xi}_S|)^2$ do not. Let $\hat{d}_B(\alpha)$ be the α-th quantile of $\hat{H}_B$. Conclusions (4.11) and (4.12) continue to hold for the *bootstrap Stein confidence ball* $C_{SB} = C(\hat{\xi}_S, \hat{d}_B(\alpha))$. Further analysis reveals that both the asymptotic and bootstrap forms of the Stein confidence ball have coverage errors of order $O(n^{-1/2})$ and that coverage accuracy of order $O(n^{-1}$ is achieved by a prepivoted bootstrap construction of the confidence ball radius. See Beran [1] for details.

5. Multiple Stein shrinkage

The James–Stein estimator is often viewed as a curiosity of little practical use. The semifinal paragraph on p. 198 of Stein [10] addressed this point and showed how to resolve it: "A simple way to obtain an estimator which is better for most practical purposes is to represent the parameter space ... as an orthogonal direct sum of two or more subspaces, also of large dimension and apply spherically symmetric

estimators separately in each." The geometric asymptotic reasoning in Stein's paper extends readily to multiple shrinkage.

Let $O = [O_1|O_2|\dots|O_s]$ be a specified $n \times n$ orthogonal matrix partitioned into s submatrices $\{O_k : 1 \le k \le s\}$ such that O_k is $n \times n_k$, each $n_k \ge 1$, and $\sum_{k=1}^s n_k = n$. Define $P_k = O_k O_k'$. The $\{P_k : 1 \le k \le s\}$ are orthogonal projections into R^n, are mutually orthogonal, and sum to I_n. The mean vector ξ and the data vector X can then be expressed as sums, $\xi = \sum_{k=1}^s P_k \xi$ and $X = \sum_{k=1}^s P_k X$, the summands in each case being mutually orthogonal.

Consider the *candidate multiple shrinkage estimators*

$$\hat{\xi}(a) = \sum_{k=1}^s a_k P_k X, \qquad a \in [0,1]^s, \tag{5.1}$$

where $a = (a_1, a_2, \dots, a_s)$. These form the closure of the class of candidate penalized least squares estimators

$$\operatorname*{argmin}_{\xi \in R^n} [|X - \xi|^2 + \sum_{k=1}^s \lambda_k |P_k \xi|^2], \qquad \lambda_k \ge 0, \quad 1 \le k \le s. \tag{5.2}$$

Let $\tau_k = n^{-1} \operatorname{tr}(P_k) = n_k/n$ and let $w_k = n^{-1}|P_k \xi|^2$. Then, the normalized quadratic risk $n^{-1}\mathrm{E}|\hat{\xi}(a) - \xi|^2$ is

$$R(\hat{\xi}(a), \xi) = \sum_{k=1}^s r(a_k, \tau_k, w_k), \tag{5.3}$$

where $r(a_k, \tau_k, w_k) = (a_k - \tilde{a}_k)^2(\tau_k + w_k) + \tau_k \tilde{a}_k$, with $\tilde{a}_k = w_k(\tau_k + w_k)^{-1}$. Let $\tilde{a} = (\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_s)$. The *oracle* multiple shrinkage estimator that minimizes risk is clearly $\tilde{\xi}_{MS} = \hat{\xi}(\tilde{a})$ and the oracle risk is

$$R(\tilde{\xi}_{MS}) = \sum_{k=1}^s \tau_k w_k (\tau_k + w_k)^{-1}. \tag{5.4}$$

Unfortunately, $\tilde{\xi}_{MS}$ depends on the unknown $\{w_k\}$.

Let $\hat{w}_k = \breve{w}_+$, where $\breve{w}_k = p^{-1}|P_k X|^2 - \tau_k$, and $\breve{w}_+$ is the positive part of $\breve{w}$. Note that $\hat{w}_k$ is non-negative like w_k and satisfies the inequality $|\hat{w}_k - w_k| \le |\breve{w}_k - w_k|$.

Replacing w_k with $\hat{w}_k$ in the oracle estimator just described yields the *multiple shrinkage estimator*

$$\hat{\xi}_{MS} = \sum_{k=1}^s \hat{w}_k (\tau_k + \hat{w}_k)^{-1} P_k X. \tag{5.5}$$

Plugging $\{\hat{w}_k\}$ into (5.4) also yields an estimator for the risk of $\hat{\xi}_{MS}$,

$$\hat{R}(\hat{\xi}_{MS}) = \sum_{k=1}^s \tau_k \hat{w}_k (\tau_k + \hat{w}_k)^{-1}. \tag{5.6}$$

Asymptotically in n, the following holds:

- *For every finite $c > 0$ and fixed integer s,*

$$\lim_{n \to \infty} \sup_{n^{-1}|\xi|^2 \le c} |R(\hat{\xi}_{MS}, \xi) - R(\tilde{\xi}_{MS}, \xi)| = 0. \tag{5.7}$$

Moreover, for V equal to either the loss $n^{-1}|\hat{\xi}_{MS} - \xi|^2$ or the risk $R(\hat{\xi}_{MS}, \xi)$,

$$\lim_{n\to\infty} \sup_{n^{-1}|\xi|^2 \le c} E|\hat{R}(\hat{\xi}_{MS}) - V| = 0. \tag{5.8}$$

Thus, the risk of the multiple shrinkage estimator $\hat{\xi}_{MS}$ converges to the best risk achievable over the candidate class; and its plug-in risk estimator converges to its actual risk or loss. Stein [11] improved on $\hat{\xi}_{MS}$ through an exact risk analysis for finite n and described an application to estimation of means in ANOVA models. The foregoing development is extended in Beran [4] to multiple affine shrinkage of a data matrix X, with first application to MANOVA models.

A much larger class of candidate estimators is generated by including, for each value of n, every possible selection of the column dimensions $n_1, n_2, \ldots, n_s$. Redefine $\hat{\xi}_{MS}$ and $\tilde{\xi}_{MS}$ to minimize, respectively, estimated risk and risk over this larger class of candidate estimators. Convergences (5.7) and (5.8) continue to hold, by applying the analysis in Beran and Dümbgen ([5], p. 1832) of bounded total variation shrinkage.

6. Adaptive symmetric linear estimators

Larger than the class of candidate multiple shrinkage estimators is the class of *candidate symmetric linear estimators*

$$\hat{\xi}(A(t)) = A(t)X, \qquad t \in \mathcal{T}, \tag{6.1}$$

where $\{A(t) \colon t \in \mathcal{T}\}$ is a family of $n \times n$ positive semidefinite matrices indexed by t. This class of estimators includes penalized least squares estimators with multiple quadratic penalties, running weighted means, nested submodel fits in regression, and more.

Let $\{\lambda_k(t) \colon 1 \le k \le s\}$ denote the distinct eigenvalues of $A(t)$ and let $\{P_k(t) \colon 1 \le k \le s\}$ denote the associated eigenprojections. Here $s \le n$ may depend on n. Then

$$\hat{\xi}(A(t)) = \sum_{k=1}^{s} \lambda_k(t) P_k(t) X, \qquad t \in \mathcal{T} \tag{6.2}$$

represents $\hat{\xi}(A(t))$ as a candidate multiple shrinkage estimator.

If the index set $\mathcal{T}$ is not too large, in the covering number sense of modern empirical process theory, it may be possible to find $\hat{t} = \hat{t}(X) \in \mathcal{T}$ such that the risk of the adaptive estimator $\hat{\xi}(A(\hat{t}))$ converges to the smallest risk achievable over the candidate class (6.2) as n tends to infinity. See Beran and Dümbgen [5] and Beran [3] for instances of such asymptotics. Such results link the profound insights and results in Stein [10] with modern theory for regularized estimators of high-dimensional parameters—estimators that have proved their value in practice.

7. Envoi

Gauss offered two brief justifications for the method of least squares. The first was what we now call the maximum likelihood argument. The second, mentioned years later in a letter to Bessel, was the concept of risk and the start of what we now call the Gauss–Markov theorem.

Stein's prophetic work [10] revealed that neither maximum likelihood estimators nor unbiased estimators necessarily have low risk when the dimension of the parameter space is not small. Despite the wonderfully transparent asymptotic geometry in his paper—geometry that extends readily to useful multiple shrinkage estimators and to the construction of confidence balls around these—many found his insights unbearable and labelled his findings paradoxical. Few contemporaries appear to have read his paper [10] carefully. Modern regularization estimators that reduce risk through beneficial multiple shrinkage have made manifest the fundamental nature of Stein's achievement.

References

[1] Beran, R. (1995). Stein confidence sets and the bootstrap. *Statistica Sinica* **5** 109–127.

[2] Beran, R. (1996). Stein estimation in high dimensions: a retrospective. In *Madan Puri Festschrift* (E. Brunner and M. Denker, eds.) 91–110. VSP, Zeist.

[3] Beran, R. (2007). Adaptation over parametric families of symmetric linear estimators. *Journal of Statistical Planning and Inference* (Special Issue on Nonparametric Statistics and Related Topics) **137** 684–696.

[4] Beran, R. (2008). Estimating a mean matrix: boosting efficiency by multiple affine shrinkage. *Annals of the Institute of Statistical Mathematics* **60** 843–864.

[5] Beran, R. and Dümbgen, L. (1998). Modulation of estimators and confidence sets. *Annals of Statistics* **26** 1826–1856.

[6] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors — an empirical Bayes approach. *Journal of the American Statistical Association* **68** 117–130.

[7] Hasminski, R. Z. and Nussbaum, M. (1984). An asymptotic minimax bound in a regression problem with an increasing number of nuisance parameters. In *Proceedings of the Third Prague Symposium on Asymptotic Statistics* (P. Mandl and M. Hušková, eds.) 275–283. Elsevier, New York.

[8] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.) **1** 361–380. University of California Press.

[9] Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian white noise. *Problems of Information Transmission* **16** 120–133.

[10] 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* (J. Neyman, ed.) **1** 197–206. University of California Press.

[11] Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In *Festschrift for Jerzy Neyman* (F. N. David, ed.) 351–364. Wiley, New York.

[12] Stein, C. (1981) Estimation of the mean of a multivariate normal distribution. *Annals of Statistics.* **9** 1135–1151.

[13] Stigler, S. M. (1990). A Galtonian perspective on shrinkage estimators. *Statistical Science* **5** 147–155.

[14] Watson, G. S. (1983). *Statistics on Spheres.* Wiley-Interscience, New York.