A Lower Bound for the Risk of Classes of Shrinkage Estimators in a General Multivariate Estimation Problem and Some Deduced Estimators

by

Christian Robert[1]
Purdue University

Technical Report #87-45

Department of Statistics
Purdue University

October 1987

[1] Research partially supported by l'Institut National de la Statistique et des Etudes Economiques.

◇

A Lower Bound for the Risk of Classes of Shrinkage Estimators in a General Multivariate Estimation Problem and Some Deduced Estimators

Abstract

Given an arbitrary quadratic loss, we propose a lower bound for the associated risk of shrinkage estimators which is of interest for a wide field of estimation models. With respect to the considered class of shrinkage estimators, this bound is optimal.

In the particular case of estimation of the location parameter of an elliptically symmetric distribution, this bound can be used to find the relative improvement brought by a given estimator using the Monte-Carlo method. We also consider a new type of shrinkage estimators whose risk can be as close as one wants of the lower bound near a chosen pole and yet remain bounded.

Key-words

quadratic risk, shrinkage estimators, elliptically symmetric distribution, ε-minimaxity.

A Lower Bound for the Risk of Classes of Shrinkage Estimators in a General Multivariate Estimation Problem and Some Deduced Estimators

1. Introduction

In many multivariate estimation problems, when new estimators are considered — for Bayesian or frequentist reasons —, it is often very difficult to evaluate their interest compared with existing estimators. Even if it is sometimes possible to show that they dominate, w.r.t. a given criterion, an "usual" estimator (e.g. the best invariant one), the improvement they bring and the remaining possible improvement are most of the time unknown.

In particular, this is often the case for the wide field of shrinkage estimation. In most of the cases, shrinkage estimators are shown to dominate the least squares estimator by using an "unbiased estimator of the risk" introduced by Stein (1973) but the comparison with other minimax estimators is not undertaken; Bock (1987) and Brown (1987) have even established that it is impossible to show that some inadmissible estimators, like the positive-part James-Stein estimator, are dominated using these methods.

We give in this paper a method to compute a lower bound of the quadratic risk of a shrinkage estimator, for a wide class of multivariate estimation problems. We also derive from this lower bound, in the spherically symmetric case, some interesting new shrinkage estimators, whose risk near the pole can be arbitrarily near of this bound and yet remain uniformly bounded.

2. A Lower Bound of the Risk in the General Case

The following general framework contains a lot of classical statistical models.

2.1. Model

In a real vector space E of dimension n, we observe a random variable y. We suppose that its distribution belongs to a parametrized set,

$$\mathcal{F} = \{F_{\theta,\delta}; (\theta, \delta) \in \Theta \times \Delta\},$$

where Θ and Δ are contained in some vector spaces. The parameter we are considering is $\mu(\theta)$, where μ is a mapping from Θ into Ξ, vector space of dimension k (note that μ has not to satisfy $\mu(\Theta) = \Xi$). Let us define a symmetric bilinear form q on Ξ. The associated quadratic form is denoted $\overline{q}$ (i.e. $\overline{q}(x) = q(x, x)$, $\forall\ x \in \Xi$). The estimators of $\mu(\theta)$, mappings from E into Ξ, are compared w.r.t. the risk associated with q (see below).

2.2. Shrinkage Estimators

We suppose that we already have an estimator of $\mu(\theta)$, φ^0, such that the risk of φ^0 is finite:

$$\mathsf{E}_{\theta,\delta}[\overline{q}(\varphi^0(y) - \mu(\theta))] < +\infty \qquad \forall\ \theta, \delta,$$

where the expectation is taken w.r.t. the distribution of y, $F_{\theta,\delta}$.

Let c be a linear mapping on Ξ. For every function h from E into $\mathbb{R}_+$, we consider the associated "shrinkage" estimator

$$\varphi_h(y) = \varphi^0(y) - h(y) \cdot c(\varphi^0(y)), \tag{2.1}$$

whose form is inspired by the normal case (see Judge and Bock (1978)). Then

Proposition. Among the estimators φ_h, for a given $\theta_0 \in \Theta$, the estimator associated with the function

$$h_{\theta_0}(y) = \frac{q(\varphi^0(y) - \mu(\theta_0), c(\varphi^0(y)))}{\overline{q}(c(\varphi^0(y)))}$$

minimizes the risk $\mathrm{E}_{\theta,\delta}[\overline{q}(\varphi_h(y) - \mu(\theta))]$ for every θ such as $\mu(\theta) = \mu(\theta_0)$.

Proof.

The risk can be written

$$\mathrm{E}_{\theta,\delta}[\overline{q}(\varphi^0(y) - \mu(\theta))] - 2\mathrm{E}_{\theta,\delta}[h(y)q(\varphi^0(y) - \mu(\theta), c(\varphi^0(y)))] + \mathrm{E}_{\theta,\delta}[h^2(y)\overline{q}(c(\varphi^0(y)))]$$

and, for every $y \in E$,

$$h(y)\left\{\overline{q}(c(\varphi^0(y)))h(y) - 2q(\varphi^0(y) - \mu(\theta), c(\varphi^0(y)))\right\}$$

is minimized by h_θ.

Furthermore the risk of the estimator φ_{h_θ} is finite at θ. It follows from the Cauchy-Schwartz inequality that it is sufficient to show

$$\mathrm{E}_{\theta,\delta}[\overline{q}(c(\varphi^0(y)))h_\theta^2(y)] < +\infty.$$

Note that

$$\mathrm{E}_{\theta,\delta}[\overline{q}(c(\varphi^0(y)))h_\theta^2(y)] = \mathrm{E}_{\theta,\delta}\left[\frac{q^2(\varphi^0(y) - \mu(\theta), c(\varphi^0(y)))}{\overline{q}(c(\varphi^0(y)))}\right] \leq \mathrm{E}_{\theta,\delta}[\overline{q}(\varphi^0(y) - \mu(\theta))] < +\infty. \qquad \square$$

2.3. Remarks

1. This result can be easily generalized to the case when the estimator is shrinking towards an arbitrary pole κ. The estimators we consider are thus

$$\varphi_n(y) = \varphi^0(y) - h(y)c(\varphi^0(y) - \kappa) \tag{2.2}$$

and the function associated to the estimator minimizing the risk at $\mu(\theta)$ is

$$h_{\theta,\kappa}(y) = \frac{q(c(\varphi^0(y) - \kappa), \varphi^0(y) - \mu(\theta))}{\overline{q}(c(\varphi^0(y) - \kappa))}.$$

2. That h_θ depends upon $\varphi^0(y)$ is fairly natural (in particular, if $\varphi^0(y) = \mu(\theta)$, we must have $h(y) = 0$). What is more surprising is that h_θ does not depend upon other functions of y. For example, in the case of the estimation of a normal vector when the variance is known up to a multiplicative factor σ^2, an estimator of σ^2 usually appears into the estimators.

3. When $\mu(\theta) = 0$, h_θ is equal to one if c is a orthogonal projection for the scalar product defined by q. Furthermore, if c is the identity linear mapping, the estimator associated with h_θ is $\varphi_{h_\theta}(y) = 0$. Thus its risk is equal to 0. For $\mu(\theta) \neq 0$ (which is the chosen pole), the risk of φ_{h_θ} is always positive, except if $k = 1$ where our result is of no interest.

4. The estimator minimizing the risk at $\mu(\theta)$ depends obviously upon $\mu(\theta)$. Therefore we find again the well-known result that there cannot exist a uniformly optimal estimator.

5. The bound we can deduce from the proposition by considering, for every θ, $\mathbb{E}_{\theta,\delta}[\overline{q}(\varphi_{h_\theta}(y) - \mu(\theta))]$ is optimal because, for every θ, there exists an estimator in the class (2.1) which reaches this bound. Therefore, if an estimator in this class has a risk equal to the bound for a given $\mu(\theta)$, it will be "pseudo-admissible" in the sense that it will be only dominated by estimators which also reach this bound. This class of pseudo-admissible estimators is interesting only in a neighborhood of $\mu(\theta)$, seen as a "second possible pole". In fact, it seems that these estimators do not have good uniform properties (w.r.t. $\mu(\theta)$) like minimaxity or even finiteness of the risk. (See Figure 1)

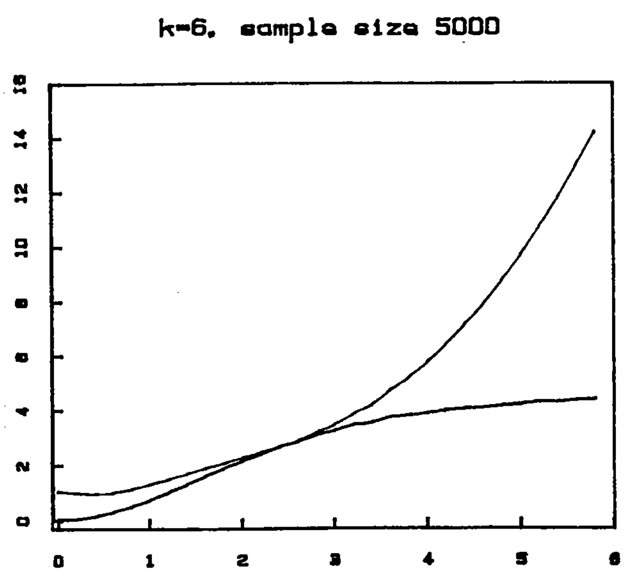


Figure 1 – Risk of φ_θ compared with the lower bound ($\theta = (1, \ldots, 1)^t$).

3. Elliptically Symmetric Distributions

3.1. General Model

The estimation of the location parameter of an elliptically symmetric distribution is a particular case of the previous framework which has been widely considered in the literature, especially in the normal case (see, e.g., Berger (1980), Bock (1985), Cellier, Fourdrinier, Robert (1987)).

We consider a random vector $y \in \mathrm{R}^k$ whose distribution is an elliptically symmetric one with dispersion matrix Σ and location parameter θ. If this random vector is absolutely continuous with respect to the Lebesgue measure, its p.d.f. can be written

$$f((y-\theta)^t \Sigma^{-1} (y-\theta)),$$

where f is a function from R_+ into R_+. We suppose here that Σ is unknown but the lower bound is the same when Σ is partially or totally unknown (see Remark 2).

Most of the shrinkage estimators of θ considered in this problem can be written (see Judge and Bock (1978))

$$\varphi(y) = (I_k - h(y)C)y$$

where C is a (k, k) matrix. Therefore they appear as a particular case of the class (2.1). Let Q be a p.s.d. matrix; if the risk is defined by $\mathrm{E}_{\theta,\Sigma}[(y-\theta)^t Q(y-\theta)]$ and θ_0 is a given vector of R^k, the estimator minimizing this risk at θ_0 is, according to the Proposition, associated with the function

$$h_{\theta_0}(y) = \frac{(y-\theta_0)^t QCy}{y^t c^t QCY}$$

and the lower bound is then

$$\mathrm{tr}(\Sigma^{-1}Q) - \mathrm{E}_{\theta_0,\Sigma}[\frac{((y-\theta_0)^t QCy)^2}{y^t C^t QCy}],$$

which can be easily computed for any θ_0 through simulation. Note that this minimum is always less than $\mathrm{tr}(\Sigma^1 Q)$, the risk of the least squares estimator.

3.2. The Symmetric Case

When $C = \Sigma = Q = I_k$, the estimator associated with h_θ is

$$\varphi_\theta(y) = \frac{y^t\theta}{y^t y}\, y$$

and the lower bound depends only upon $\theta^t\theta$ because of the spherical symmetry of the distribution. This result allows easier computations because the problem becomes essentially unidimensional (see attached figures).

In the symmetric case, for the normal distribution one well-known shrinkage estimator is the positive-part James-Stein estimator,

$$\tilde{\varphi}(y) = (1 - (k-2)\frac{1}{y^t y})^+ \; y,$$

where $(x)^+ = \max(0, x)$. It performs well but it is known to be inadmissible. However, an estimator uniformly dominating this estimator is unknown. As one can see on Figures 2 and 3, the difference between the risk of the positive-part James-Stein estimator and the lower bound is approximately one for every value of $\theta^t\theta$; we have thus refined the result of Spruill (1986) who shows that the risk of a spherically symmetric shrinkage estimator is at least the risk of the "primitive" James-Stein estimator minus two.

Even if the positive-part James-Stein estimator is almost admissible, one can hope to gain over it in some bounded regions of the parameter space. The shrinkage estimators we consider now can be used in that purpose.

3.3. Some Deduced Estimators

As we have already said, for an arbitrary $\theta \in \mathbb{R}^k$, the choice of the estimator φ_{θ_0} is only sensible in a neighborhood of θ_0 (see Figure 1). In order to apply the result of the previous proposition, one can replace θ in h_θ by an estimator $\hat{\theta}$. But, if this estimator $\hat{\theta}$ is a scalar estimator (e.g. the positive-part James-Stein estimator), the resulting estimator is also $\hat{\theta}$.

Another possibility is to consider a convex combination of estimators y_{θ_i}, for a collection of fixed θ_i, $i \in I$,

$$\varphi(y) = \sum_{i \in I} \rho_i(y)\varphi_{\theta_i}(y),$$

where,

$$\forall\, y \in \mathbb{R}^k, \sum_{i \in I} \rho_i(y) = 1 \text{ and } 0 \leq \rho_i(y) \leq 1, \forall\, i \in I.$$

For this "multiple shrinkage estimator" (see George (1986)), it would be interesting to take the weight ρ_i non null only in a neighborhood of $\theta_i (i \in I)$. Consider then a partition of $\mathbb{R}^k$ into "hypercubes" associated with the collection $\{\theta_i, i \in I\}$, the $\theta_i (i \in I)$ being the vertices of these hypercubes. The simplest case is to use a uniform partition where each hypercube has edges of constant length d; d is said to be the *diameter* of the partition. Thus, if θ_0 is a given vertex of the collection $\{\theta_i, i \in I\}$, this collection can be rewritten

$$\{\theta_0 + d\xi, \quad \xi \in \mathbb{Z}^k\}.$$

Now, if y belongs to a given hypercube, it is natural to shrink y toward the vertex of this hypercube which is nearest to 0; in other words, the "neighborhood" of $\theta = (v_1, v_2, \ldots, v_k)$ will be the hypercube

$$\{(v_1 + \frac{v_1}{|v_1|}e_1\; d, v_2 + \frac{v_2}{|v_2|}e_2\; d, \ldots, v_k + \frac{v_k}{|v_k|}e_k\; d); e_i \in [0, 1, \; 1 \leq i \leq k\}$$

with the convention $\frac{0}{|0|} = 1$. And 0 must be obviously shrunken towards itself; this implies that 0 is a vertex of the partition. Therefore, for every $d \in \mathbb{R}_+^*$, the partition $\mathcal{T}_d$ of $\mathbb{R}^k$ is the collection of hypercubes of edges of length d and vertices in $\{\theta_\xi; \xi \in d \cdot \mathbb{Z}^k\}$.

Let $[x]$ denote the integral part of x (with $[-2.3] = -2$); for $d \in \mathbb{R}_+^*$, we define ip_d to be the mapping from $\mathbb{R}^k$ into $d\,\mathbb{Z}^k$ which associates to every $y \in \mathbb{R}^k$

$$ip_d(y) = (d[\frac{y_1}{d}], d[\frac{y_2}{d}], \ldots, d[\frac{y_k}{d}])^t.$$

It is easy to check that, for the partition $\mathcal{T}_d$, y always belongs to the hypercube whose vertex nearest to 0 is $ip_d(y)$.

Consider the following choice of the weighting functions:

$$\rho_\xi^d(y) = \begin{cases} 1 & \text{if } \xi = ip_d(y) \\ 0 & \text{otherwise} \end{cases} \qquad (\xi \in d \cdot \mathbb{Z}^k).$$

This choice can be smoothened to make the weighting functions continuous or even C^∞. With this crude choice, the resulting estimator,

$$\hat{\varphi}_d(y) = \sum_{\xi \in d\mathbb{Z}^k} \rho_\xi^d(y)\varphi_\xi(y),$$

does not appear anymore as a multiple shrinkage estimator but rather as an estimator of φ_θ where θ is estimated by $\hat{\theta}_d(y) = ip_d(y)$. We have then

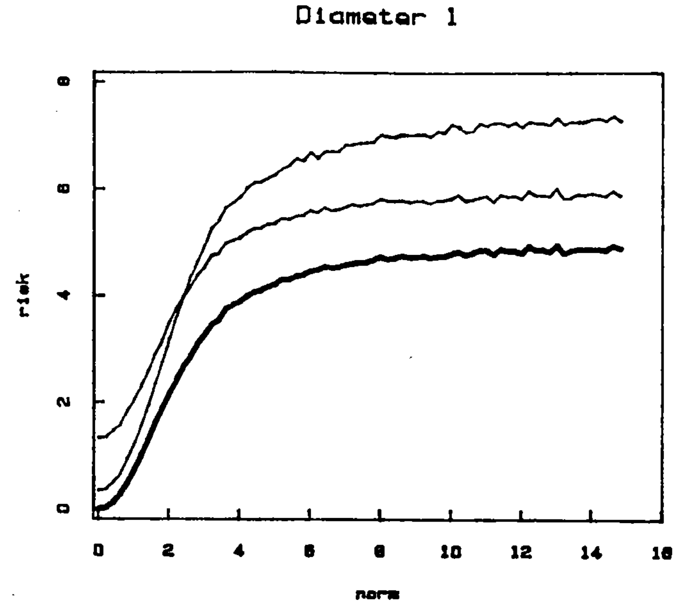
$$\hat{\varphi}_d(y) = \frac{y^t ip_d(y)}{y^t y}\, y.$$

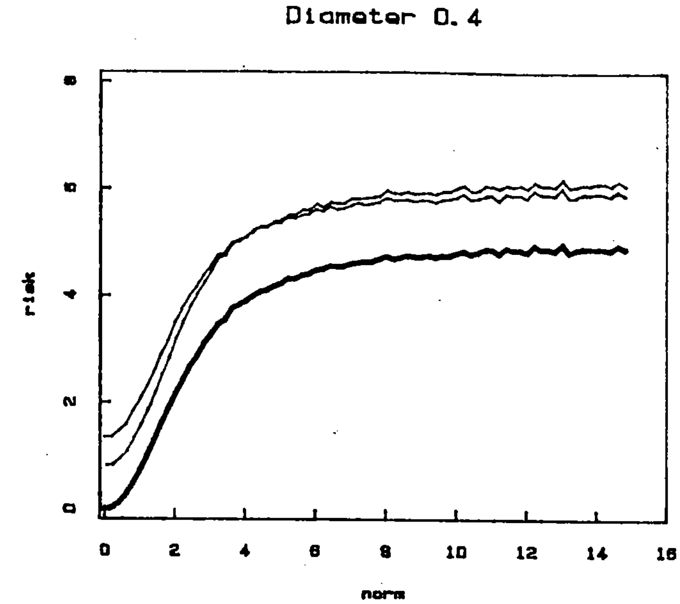
But, as we are estimating θ and $\tilde{\varphi}$ uniformly dominates the least squares estimator, we can logically substitute for $\hat{\theta}_d$ the estimator $\tilde{\theta}_d(y) = ip_d(\tilde{\varphi}(y))$. The simulations we have done show that this choice significantly improves the estimation of θ. Note that, as d goes to 0, $\tilde{\theta}_d$ goes to $\tilde{\varphi}$ and

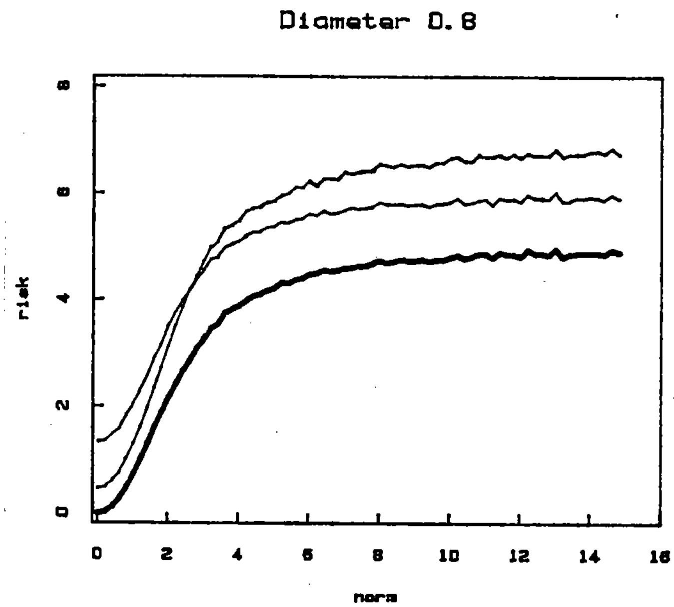
$$\tilde{\varphi}_d(y) = \frac{y^t ip_d(\tilde{\varphi}(y))}{y^t y}\, y$$

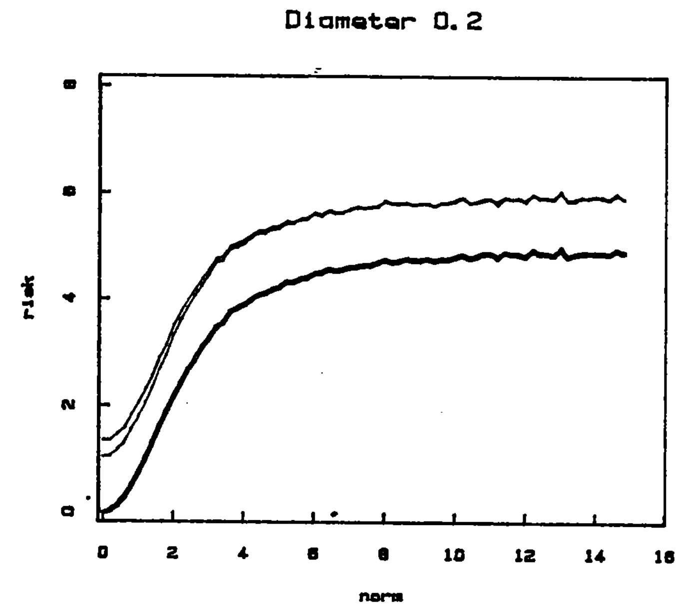
goes also to $\tilde{\varphi}(y)$, for every $y \in \mathbb{R}^k$.

This class of estimators is interesting from the point of view of our lower bound. In fact, as the following figures show, the risk of $\tilde{\varphi}_d$ goes closer to the lower bound near 0 as d grows larger. One drawback is that these estimators are no longer minimax even though the simulations seem to establish that their risk levels off for $\theta^t\theta$ large enough. The maximum risk is an increasing function of d (see Figure 2); the more one gains near 0, the more one loses (w.r.t. the least squares estimator) for large values of θ. Another drawback is that they do not bring any improvement for large values of θ.









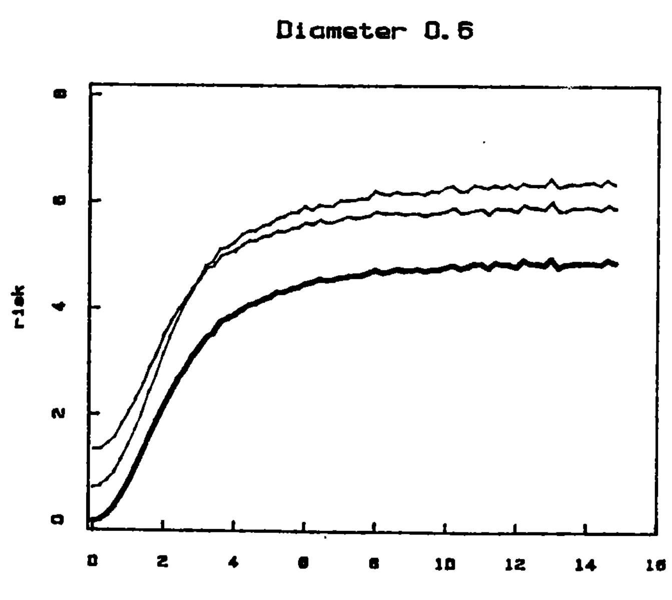


Figure 2 – Risks of estimators $\tilde{\varphi}_d$ compared with the risk of the positive-part James-Stein estimator and the lower bound ($k = 6$, 5000 observations)

We have then got a family of ε-minimax shrinkage estimators which dominate the positive-part James-Stein estimator near 0, even if the rigorous determination of the ε associated with a given $\tilde{\varphi}_d$ seems rather difficult. But it is worthwhile to note that, for this kind of problems, ***restricted risk Bayes estimators*** are nearly optimal (see Berger (1982), Chen (1983), Spruill (1986)). Given a prior, the associate estimator nearly minimizes the corresponding Bayes risk among the ε-minimax shrinkage estimators; in particular, one can choose the prior to be the uniform distribution over a region of interest.

However, even if these estimators $\tilde{\varphi}_d$ are not optimal in this way, they provide, for "good" values of the diameter, performing competitors of the positive-part James-Stein estimator, $\tilde{\varphi}$. To compare them, let us define the two following distance indicators:

(a) the *proportional saving performance* (PSP)

$$PSP(\tilde{\varphi}_d) = \frac{\sup_\theta\{R(\tilde{\varphi},\theta) - R(\tilde{\varphi}_d,\theta)\}}{\sup_\theta\{R(\tilde{\varphi}_d,\theta) - R(\tilde{\varphi},\theta)\}},$$

(b) the *relative proportional saving performance* (RPSP)

$$RPSP(\tilde{\varphi}_d) = \frac{\sup_\theta\{\frac{(R(\tilde{\varphi},\theta)-R(\tilde{\varphi}_d,\theta))}{R(\tilde{\varphi},\theta)}\}}{\sup_\theta\{\frac{(R(\tilde{\varphi}_d,\theta)-R(\tilde{\varphi},\theta))}{R(\tilde{\varphi},\theta)}\}}$$

where $R(\varphi,\theta) = \mathsf{E}_\theta[||\varphi(y) - \theta||^2]$.

Both of them evaluate the ratio of maximal improvement over maximal loss, the interesting estimators being associated with indicators greater than one. RPSP evaluates the performance of an estimator in terms of percentage of improvement (or loss) with respect to $\tilde{\varphi}$ and is, from our point of view, more reliable. Note that the denominators will always be positive, as $\tilde{\varphi}$ is nearly admissible.

Table 1 contains some simulation results about the performances of estimators $\tilde{\varphi}_d$ for both indicators (the numbers between parentheses are the numerators of the corresponding indicators).

k \ d	0.05	0.1	0.2	0.5	0.8	1	Adaptive
6	(0.08) 16.05	(0.16) 8.56	(0.3) 4.32	(0.65) 1.58	(0.88) 0.85	(1.0) 0.62	(0.45) 7.69
	(0.06) 72.46	(0.12) 38.48	(0.22) 19.11	(0.48) 7.05	(0.66) 3.80	(0.74) 2.75	(0.34) 29.25
9	(0.09) 12.51	(0.17) 6.73	(0.33) 3.41	(0.69) 1.16	(0.93) 0.61	(1.03) 0.44	(0.42) 6.13
	(0.07) 88.63	(0.14) 47.58	(0.26) 23.99	(0.54) 8.22	(0.73) 4.29	(0.81) 3.10	(0.33) 37.13
13	(0.10) 11.18	(0.20) 5.53	(0.37) 2.69	(0.76) 0.91	(0.98) 0.46	(1.08) 0.33	(0.41) 6.11
	(0.08) 116.86	(0.16) 57.84	(0.3) 28.09	(0.61) 9.37	(0.79) 4.75	(0.87) 3.36	(0.33) 47.44
16	(0.11) 11.56	(0.21) 5.47	(0.40) 2.50	(0.79) 0.79	(1.01) 0.39	(1.09) 0.27	(0.41) 5.33
	(0.09) 150.64	(0.17) 71.19	(0.33) 32.49	(0.65) 10.24	(0.83) 5.11	(0.90) 3.55	(0.34) 57.79

Table 1 — Values of PSP ($\tilde{\varphi}_d$) and RPSP ($\tilde{\varphi}_d$).

The *'adaptive diameter estimator'* is a modification of the estimator $\tilde{\varphi}_d$ where d is replaced by $\hat{d} = (y^t y)^{-1/2}$, the diameter being inversely proportional to the norm of y. As one can see in the previous table, this estimator performs rather well, giving good indicators *and* significant improvement over the positive-part James-Stein estimator (see also Figure 3).

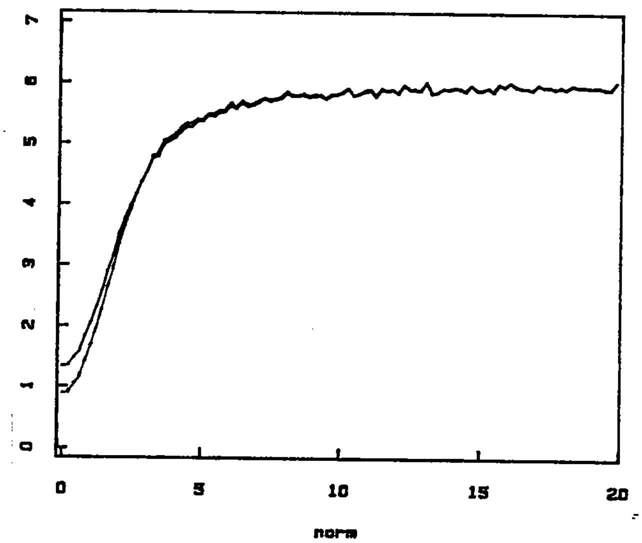


Figure 3 — Risk of the adaptive diameter estimator compared with the risk of $\tilde{\varphi}$ ($k = 6$, 5000 observations)

Acknowledgment

I would like to thank Anirban DasGupta for suggesting significant improvements over an earlier draft of this paper.

References

BERGER, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean, *Ann. Stat.* **8**(**4**), 716–761.

BERGER, J. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results, in *Statistical Decision Theory and Related Topics III* (**1**), 109–141.

BOCK, M. E. (1985). Minimax estimators that shift towards a hypersphere for location vectors of spherically symmetric distributions, *J. Mult. Anal.* **17**, 127–147.

BOCK, M. E. (1987). Shrinkage estimators: Pseudo-Bayes Rules for Normal Mean Vectors. Technical Report Purdue University #87-18.

BROWN, L. (1987). The differential inequality of a statistical estimation problem. *Proc. Fourth. Purdue Symp. Dec. Theo. Rel. Topic* (S. S. Gupta and J. Berger, eds.) Springer Verlag, New York.

CELLIER, D. FOURDRINIER, D., ROBERT C. (1987). Estimateurs à rétrécisseur du paramètre de position d'une loi à symétrie sphérique. Note aux Comptes Rendus de l'Académie des Sciences de Paris **304(14)**, 439–442.

CHEN, S. (1983). Restricted Risk Bayes Estimation, Ph.D. Thesis, Purdue University.

GEORGE, E. (1986). Minimax multiple shrinkage estimation. *Ann. Stat.* **14(1)**, 188–205.

JUDGE, G., BOCK, M. E. (1978). The statistical implications of Pre-test and Stein-rule estimators in Econometrics. North-Holland.

SPRUILL, M. (1986). Some approximate restricted Bayes estimators of a normal mean, *Statistics & Decisions* **4**, 337–351.