Contributions to the Encyclopedia of Biostatistics

George Casella *
Cornell University

July 22, 1996

Abstract

Three short expository articles on shrinkage and Stein estimation for the Encyclopedia of Biostatistics.

*This research was supported by NSF Grant No. DMS-9625440. This is paper BU-1347-M in the Biometrics Unit, Cornell University, Ithaca, NY 14853. File three-1.tex

Shrinkage

The problem of estimating the mean of a normal distribution is central to the practice of statistics. This simple problem is at the heart of many of the most common procedures used today, such as the analysis of variance or regression. If we have a random sample $X_1, \cdots, X_n$ from a normal population with mean μ and variance σ^2, the natural estimator of μ is the sample mean $\bar{X} = (1/n)\sum_i X_i$. A question of interest is whether this estimator is the *best* estimator of the parameter μ.

When assessing the performance of an estimator, in particular whether it is best, it is necessary to have a criterion with which to measure it against. A most popular measure is *squared error loss*, where we measure the performance of an estimator d of a parameter θ by the function

$$L(\theta, d) = (\theta - d)^2, \tag{1}$$

which is called a *loss function.*

Under the loss function (1), $\bar{X}$ has many optimality properties. For example, it is a *minimax estimator* of μ (see Minimax Estimation), meaning that of all estimators of μ, its loss has the smallest maximum value. There are other properties that $\bar{X}$ enjoys, including the property of *admissibility.* An estimator d of a parameter θ is an *admissible estimator* of θ under the loss $L(\theta, d)$ if there is no other estimator d' that satisfies

$$E_\theta[L(\theta, d)] \geq E_\theta[L(\theta, d')], \quad \text{for all } \theta,$$

with strict inequality for some values of θ.

Is $\bar{X}$ an admissible estimator of θ? Hodges and Lehmann (1951) and Blyth (1951) showed that it was. That is, there is no estimator that is uniformly better. However, if the problem is made slightly more complex, an interesting result unfolds. Suppose that instead of estimating the mean of one normal population, we are interested in estimating the mean of many normal populations, that is, we observe $\bar{X}_k$, $k = 1, \cdots, p$, where $\bar{X}_k$ is the mean of n observations from a normal population with mean μ_k and variance σ^2, and we want to estimate $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)$. The loss of an estimator $\mathbf{d} = (d_1, \cdots, d_p)$ is measured by the sum of squared errors, that is

$$L(\boldsymbol{\mu}, \mathbf{d}) = \sum_{k=1}^{p} (\mu_k - d_k)^2, \tag{2}$$

and we ask if $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)$ is still an admissible estimator of $\boldsymbol{\mu}$. If $p = 2$, Stein (1956) showed that the answer is yes, but he also showed that if $p > 2$, the answer is no. Using arguments based on the idea that, for estimating more than 2 means, $\bar{\mathbf{X}}$ tends to be "too long", Stein demonstrated the existence of a better estimator, a *shrinkage* estimator. Such an estimator shrinks the vector $(\bar{X}_1, \cdots, \bar{X}_p)$ toward a specific point in the parameter space. In James and Stein (1961) it was shown that the estimator

$$\mathbf{d}^{JS}(\bar{\mathbf{x}}) = \left(1 - \frac{(p-2)\sigma^2}{|\bar{\mathbf{x}}|^2}\right)\bar{\mathbf{x}},$$

which shrinks $\bar{\mathbf{X}}$ toward 0, uniformly dominates $\bar{\mathbf{X}}$ as an estimator of $\boldsymbol{\mu}$ under the loss (2), so $\bar{\mathbf{X}}$ is not an admissible estimator.

This extremely surprising result has resulted in an enormous amount of research in areas such as decision theory and empirical Bayes analysis. Many superior procedures have been since derived. See the review article by Brandwein and Strawderman (1990), or the book by Lehmann and Casella (1997).

References

1. Blyth C. R (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Statist.* **22**, 22-42.

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

3. Hodges, J. L., Jr., and Lehmann, E. L. (1951). Some applications of the Cramér-Rao inequality. *Proc. Second Berkeley Symp. Math. Statist. Prob.* **1**, University of California Press, 13-22.

4. James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1**, University of California Press, 311-319.

5. Lehmann, E. L. and Casella, G. (1997). *Theory of Point Estimation, Second Edition.* New York: Springer-Verlag.

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

The James-Stein Estimator

The discovery of Stein (1956) that the sample mean of a normal population is inadmissible in three or more dimensions was based on an argument using the estimator

$$\mathbf{d}^1(\mathbf{x}) = \left(1 - \frac{b}{a + |\mathbf{x}|^2}\right)\mathbf{x}.$$

where we observe $\mathbf{X} = \mathbf{x}$, with $\mathbf{X} \sim N(\boldsymbol{\theta}, I)$, a p-dimensional normal random variable. If $p \geq 3$, Stein showed that for sufficiently small b and sufficiently large a,

$$E_\theta \left|\mathbf{d}^1(\mathbf{X}) - \boldsymbol{\theta}\right|^2 < E_\theta \left|\mathbf{X} - \boldsymbol{\theta}\right|^2 \quad \text{for all } \boldsymbol{\theta}, \tag{1}$$

demonstrating the inadmissibility of $\mathbf{X}$ under squared error loss. This result only demonstrated the existence of a better estimator, as Stein did not give specific values of a and b that would satisfy (1). This was remedied in James and Stein (1961), where it was shown that the estimator

$$\mathbf{d}^{JS}(\mathbf{x}) = \left(1 - \frac{c}{|\mathbf{x}|^2}\right)\mathbf{x} \tag{2}$$

dominates $\mathbf{X}$ as long as $0 \leq c \leq p - 2$. In fact, James and Stein (1961) show that the optimal value of c is $c = p - 2$, and using this value (2) is usually referred to as the *James-Stein estimator*. Starting from (2), entire families of improved estimators of $\boldsymbol{\theta}$ have been derived. Note, in particular, that since $\mathbf{X}$ is a minimax estimator of $\boldsymbol{\theta}$, any estimator that dominates it is also a minimax estimator. Thus, research began into finding better families of minimax estimators of a multivariate normal mean.

One of the most important developments was due to Baranchik (1970), who proved that estimators of the form

$$\mathbf{d}^B(\mathbf{x}) = \left(1 - \frac{r(|\mathbf{x}|)}{|\mathbf{x}|^2}\right)\mathbf{x}$$

are minimax provided

(i). $0 \leq r(\cdot) \leq 2(p-2)$ and

(ii). the function r is nondecreasing.

An immediate consequence of Baranchik's result was the minimaxity of (and the dominance of $\mathbf{X}$ by) the *positive- part Stein estimator*

$$\mathbf{d}^+(\mathbf{x}) = \left(1 - \frac{p-2}{|\mathbf{x}|^2}\right)^+ \mathbf{x} \tag{3}$$

where $(\cdot)^+$ indicates that the quantity in parentheses is replaced by 0 whenever it is negative. This represents a big improvement over (2), as it does not suffer from aberrant behavior when $\mathbf{x}$ is near 0. (There, the James-Stein estimator can actually get infinitely large.) In fact, the positive-part estimator (3) is so good that even though it is known to be inadmissible, it took over twenty-five years to exhibit an estimator that dominates it. (The inadmissibility of (3) follows from Brown 1971, who showed that the admissible estimators must be generalized Bayes estimators. Because of the "point" at $|\mathbf{x}|^2 = p - 2$, (3) is not smooth enough to be generalized Bayes. The work of Efron and Morris 1973, Section 5, showed that (3) was close to being a Bayes rule, hence close to admissible, so it was suspected that

it would difficult to dominate. Finally, Shao and Strawderman (1994) exhibited a dominating estimator.)

References

1. Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.* **41**, 642-645.

2. Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42**, 855-903. (Corr: *Ann. Statist.* **1**, 594-596.)

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

4. James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1**, University of California Press, 311-319.

5. Shao, P. Y-S. and Strawderman, W. E. (1994). Improving on the James-Stein positive-part estimator. Technical Report, Department of Statistics, Rutgers University. To appear in *Ann. Statist.*

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

Shrinkage Estimation

Starting from the work of Stein (1956) (see the James-Stein Estimator), the topic of shrinkage estimation received an enormous amount of attention in the statistical literature. The original shrinkage estimators were developed for the case of estimating the mean of a multivariate normal distribution under squared error loss, based on observing $\mathbf{X} = \mathbf{x}$, with $\mathbf{X} \sim N(\boldsymbol{\theta}, I)$, a p-dimensional normal random variable. However, results on shrinkage have been generalized to the extent that these estimators can now be routinely applied to actual problems.

In terms of practical applicability, the direction pointed out by Lindley (1962) has proved quite fruitful. Lindley showed that one could shrink toward a point chosen by the data and demonstrated, for $p \geq 4$, the minimaxity of the estimator

$$\mathbf{d}^L(\mathbf{x}) = \bar{x}\mathbf{1} + \left(1 - \frac{p-3}{|\mathbf{x} - \bar{x}1|^2}\right)(\mathbf{x} - \bar{x}1),$$

where $\mathbf{1}$ is a column vector of 1s and $|\mathbf{x} - \bar{x}\mathbf{1}|^2 = \sum(x_i - \bar{x})^2$. Dimension 4 is needed here, rather than the 3 dimensions needed for the minimaxity of the James-Stein estimator because we are now shrinking to a 1- dimensional subspace, rather than the 0-dimensional point toward which the James- Stein estimator shrinks. The idea of shrinking toward a subspace has enhanced the applicability of shrinkage estimators, and has tied them in with empirical Bayes estimation. Much of this topic was developed in the sequence of papers by Efron and Morris (1973ab, 1975), where the connection with minimax estimation is thoroughly explored. A comprehensive treatment of theory and applications of empirical Bayes methods is given by Morris (1983), and less technical introductions are given by Casella (1985, 1992).

On the more theoretical side, in the normal case, Strawderman(1971) was the first to exhibit *proper Bayes minimax* estimators, estimators that not only dominated $\mathbf{X}$, but were themselves proper Bayes and admissible. These estimators have the form of Baranchik's estimators (see The James-Stein Estimator) and a particular one is given by

$$\mathbf{d}^S(\mathbf{x}) = \left(1 - \frac{c(|\mathbf{x}|)}{|\mathbf{x}|^2}\right)\mathbf{x} \tag{1}$$

where

$$c(|\mathbf{x}|) = p + 2 - \frac{2\exp(-\frac{1}{2}|\mathbf{x}|^2)}{\int_0^1 \lambda^{p/2}\exp(-\lambda|\mathbf{x}|^2/2)\,d\lambda}.$$

The estimator (1) can be derived from the Bayes model

$$\begin{aligned} \mathbf{X} &\sim N_p(\boldsymbol{\theta}, I) \\ \boldsymbol{\theta} &\sim N_p(0, \lambda^{-1}(1-\lambda)I) \\ \lambda &\sim \text{Uniform } (0,1), \end{aligned}$$

which is a proper Bayes model if $p \geq 5$.

Thus far we have discussed only the normal distribution, however, domination of the usual estimator by shrinkage estimator in many other situations, even in discrete families. For example, if $X_i \sim Poisson(\lambda_i), i = 1, \ldots, p$, $p \geq 2$, are independent, and the loss is given by

$$L(\boldsymbol{\lambda}, \mathbf{d}) = \sum_{i=1}^{r}(\lambda_i - d_i)^2/\lambda_i,$$

then Clevensen and Zidek (1975) showed that the estimator

$$\mathbf{d}^{CZ}(\mathbf{x}) = \left(1 - \frac{c(\Sigma x_i)}{\Sigma x_i + b}\right)\mathbf{x}$$

is minimax if

(i). $c(\cdot)$ is nondecreasing

(ii). $0 \leq c(\cdot) \leq 2(p-1)$

(iii). $b \geq p - 1$.

This result highlights two differences between the normal and Poisson case. First, domination only requires $p \geq 2$, and the loss is now scaled squared error, instead of ordinary squared error. (The fact that we only now require $p \geq 2$ is discussed by Brown 1979, who described it as a "dimension doubling" phenomenon; see also Johnstone and McGibbon 1992.) Shrinkage estimators continue to dominate in many other discrete families. Using a different method of proof than Clevensen and Zidek (1975), Hwang (1982)

[see also Ghosh et al. 1987] demonstrated dominance of shrinkage estimators in a many discrete families.

An interesting exception is the binomial distribution, where Johnson (1973) demonstrated that no shrinkage estimator will dominate the usual estimator. This result was extended by Brown (1981), and later Guttmann(1982ab) established the somewhat surprising result that shrinkage estimators can never dominate in any problem with a finite sample space. (Domination by shrinkage is often referred to as *the Stein effect*, so there is no Stein effect in problems with finite sample spaces.)

Even with this limitation from finite sample spaces, shrinkage estimation has played a large role in developments in both theory and practice. On the practical side, the previously mentioned connection with empirical Bayes methods (and also hierarchical Bayes methods) has allowed the application of shrinkage estimators in a wide variety of problems. The theoretical developments have also been numerous, and have sometimes been accompanied by advances in the mathematical attack on the problem.

In the normal case all restrictions on the covariance matrix can be removed (see, for example, Gleser 1986). Outside of the normal case shrinkage estimators exist for spherically symmetric distributions (Cellier et al 1989, Brandwein and Strawderman 1990), and some results apply to the entire exponential family (Hudson 1978). For the case of estimating a gamma scale parameter, Berger (1980) obtained some interesting domination results, including domination by some "expanders" rather than shrinkers. The implications of this are further discussed by Brown (1980).

The theory of *superharmonic functions*, a type of multivariate concave function, which was originally applied to minimax estimation by Stein (1981), has also been valuable in extending shrinkage domination. George (1986ab) used it to establish dominion by *multiple shrinkage* estimators, estimators that can shrink to more than one target. More recently, Fourdrinier, et al.(1996) applied it to construct new families of proper Bayes minimax estimators based on Cauchy prior distributions.

Although the use of squared error loss is analytically convenient, shrinkage domination extends to other losses as well. For example, variations on squared error loss that allow weight matrices can easily be accommodated. Domination under an entire class of weighted squared error loss functions can be achieved (Brown 1975, Shinozaki 1980), as well as more general universal domination (Hwang 1985). Other results include those of Brandwein and

Strawderman (1980), who established domination results for concave losses and Berger (1976), who derived necessary conditions for dominance under a wide variety of losses. A more complete discussion of this, and many other aspects of shrinkage estimation, can be found in Lehmann and Casella (1997).

References

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

2. Berger, J. (1980a). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters *Ann. Statist.* **8**, 545-571.

3. Brandwein, A. C. and Strawderman, W. E. (1980). Minimax estimation of location parameters for spherically symmetric distributions with concave loss. *Ann. Statist.* **8**, 279- 284.

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

5. Brown, L. D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters). *J. Amer. Statist. Assoc.* **70**, 417-427.

6. Brown, L. D. (1979). A heuristic method for determining admissibility of estimators - with applications. *Ann. Statist.* **7**, 960-994.

7. Brown, L. D. (1980). Examples of Berger's phenomenon in the estimation of independent normal means. *Ann. Statist.* **8**, 572-585.

8. Brown, L. D. (1981). A complete class theorem for statistical problems with finite sample spaces. *Ann. Statist.* **9**, 1289-1300.

9. Casella, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.* **39**, 83-87.

10. Casella, G. (1992). Illustrating empirical Bayes methods. *Chemolab* **16**, 107-125.

11. Cellier, D., Fourdrinier, D. and Robert, C. (1989). Robust shrinkage estimators of the location parameter for elliptically symmetric distributions. *J. Mult. Anal.* **29**, 39-42.

12. Clevensen, M. L. and Zidek, J. (1975). Simultaneous estimation of the mean of independent Poisson laws. *J. Amer. Statist. Assoc.* **70**, 698-705.

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

14. Efron, B. and Morris, C. (1973b) . Combining possibly related estimation problems (with discussion). *J. Roy. Statist. Soc. Ser. B* **35**, 379-421.

15. Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**, 311-319.

16. Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (1996). On the construction of proper Bayes minimax estimators. Technical Report, Statistics Center, Cornell University.

17. George, E.I. (1986a). Minimax multiple shrinkage estimators. *Ann. Statist.* **14**, 188- 205.

18. George, E.I. (1986b). Combining minimax shrinkage estimators. *J. Amer. Statist. Assoc.* **81**, 437-445.

19. Ghosh, M., Hwang, J. T. and Tsui, K-W. (1983). Construction of improved estimators in multiparameter estimation for discrete exponential families. *Ann. Statist.* **11**, 351-367.

20. Gleser, L. J. (1986). Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Ann. Statist.* **14**, 1625-1633.

21. Guttman, S. (1982a). Stein's paradox is impossible in problems with finite parameter spaces. *Ann. Statist.* **10**, 1017-1020.

22. Guttman, S. (1982b). Stein's paradox is impossible in the nonanticipative context. *J. Amer. Statist. Assoc.* **77**, 934-935.

23. Hudson, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6**, 473-484.

24. Hwang, J. T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases. *Ann. Statist.* **10**, 857- 867.

25. Hwang, J. T. (1985). Universal domination and stochastic domination: Estimation simultaneously under a broad class of loss functions. *Ann. Statist.* **13**, 295-314.

26. Johnson, B. McK. (1971). On the admissible estimators for certain fixed sample binomial problems. *Ann. Math. Statist.* **42**, 1579-1587.

27. Johnstone, I. and MacGibbon, K. B. (1992). Minimax estimation of a constrained Poisson vector. *Ann. Statist.* **20**, 807-831.

28. Lehmann, E. L. and Casella, G. (1997). *Theory of Point Estimation, Second Edition.* New York: Springer-Verlag.

29. Lindley, D. V. (1962). Discussion of the paper by Stein. *J. Roy. Statist. Soc. Ser. B* **24** 265-296.

30. Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78**, 47-65.

31. Shinozaki, N. (1980). Estimation of a multivariate normal mean with a class of quadratic loss functions. *J. Amer. Statist. Assoc.* **75**, 973-976.

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

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

34. Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42**, 385-388.