Statistical Science
1990, Vol. 5, No. 1, 147–155

The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators

Stephen M. Stigler

Abstract. More than 30 years ago, Charles Stein discovered that in three or more dimensions, the ordinary estimator of the vector of means of a multivariate normal distribution is inadmissible. This article examines Stein's paradox from the perspective of an earlier century and shows that from that point of view the phenomenon is transparent. Furthermore, this earlier perspective leads to a relatively simple rigorous proof of Stein's result, and the perspective can be extended to cover other situations, such as the simultaneous estimation of several Poisson means. The relationship of this perspective to other earlier work, including the empirical Bayes approach, is also discussed.

Key words and phrases: Stein paradox, regression, James–Stein estimation, Poisson distribution, admissibility, empirical Bayes.

1. INTRODUCTION

One of the most provocative results in mathematical statistics of the past 35 years is the phenomenon known variously as Stein's paradox, shrinkage estimation, or the James–Stein estimator. In its simplest form (the only one considered here), the situation is this: a collection of independent measurements $X_1, X_2, \cdots, X_k$ is available, each measuring a different θ_i, each normally distributed $N(\theta_i, 1)$. The θ_i's are fixed unknown parameters which need bear no relation to one another, and it is desired to estimate all the θ_i's with composite loss function

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{k} (\theta_i - \hat{\theta}_i)^2,$$

where

$$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)',$$
$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)'.$$

The performance of the joint estimator $\hat{\boldsymbol{\theta}}$ is to be judged by the risk function,

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = EL(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}).$$

The startling discovery of Stein was that the obvious or "ordinary" estimator $\hat{\theta}_i^{\circ} = X_i$ is inadmissible if $k \geq 3$; in fact, for $k \geq 3$ any estimator of the form

$$\hat{\theta}_i^{\mathrm{JS}} = \left(1 - \frac{c}{S^2}\right) X_i$$

has uniformly smaller risk for all $\boldsymbol{\theta}$, where

$$S^2 = \sum_{j=1}^{k} X_j^2,$$

and c is any constant with $0 < c < 2(k-2)$. (The best choice of c is $k - 2$.) Because $\hat{\theta}_i^{\mathrm{JS}}$ may be considered as a weighted average of 0 and X_i, it has been described as "shrinking" the ordinary estimator $\hat{\theta}_i^{\circ}$ toward 0 despite the fact that if $S^2 < c$ it "shrinks past" 0. Variants of this have been devised, including the Efron–Morris estimators which "shrink" X_i toward $\bar{X}$ and dominate the ordinary estimator as long as $k \geq 4$; these estimators are of the form

$$\hat{\theta}_i^{\mathrm{EM}} = \bar{X} + \left(1 - \frac{c}{S'^2}\right)(X_i - \bar{X}),$$

where

$$S'^2 = \sum_{i=1}^{k} (X_i - \bar{X})^2$$

and c is here any constant with $0 < c < 2(k-3)$. (The best c is $k - 3$.) These results date from the work of Stein (1956), James and Stein (1961), Lindley (1962), and Efron and Morris (1973).

When this phenomenon is first encountered it can seem preposterous—how can (to use a variant of an early illustration) information about the price of apples in Washington and about the price of oranges in

Stephen M. Stigler is Professor, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science*. www.jstor.org

Florida be used to improve an estimate of the price of French wine, when it is assumed that they are unrelated? The best heuristic explanation that has been offered is a Bayesian argument: If the θ_i are a priori independent $N(0, \tau^2)$, then the posterior mean of θ_i is of the same form as $\hat{\theta}_i^{JS}$, and hence $\hat{\boldsymbol{\theta}}^{JS}$ can be viewed as an empirical Bayes estimator (Efron and Morris, 1973; Lehmann, 1983, page 299). Another explanation that has been offered is that $\hat{\boldsymbol{\theta}}^{JS}$ can be viewed as a relative of a "pre-test" estimator; if one performs a preliminary test of the null hypothesis that $\boldsymbol{\theta} = \mathbf{0}$, and one then uses $\hat{\boldsymbol{\theta}} = \mathbf{0}$ or $\hat{\theta}_i = X_i$ depending on the outcome of the test, the resulting estimator is a weighted average of $\mathbf{0}$ and $\hat{\boldsymbol{\theta}}^0$ of which $\hat{\boldsymbol{\theta}}^{JS}$ is a smoothed version (Lehmann, 1983, pages 295–296). But neither of these explanations is fully satisfactory (although both help render the result more plausible); the first because it requires special a priori assumptions where Stein did not, the second because it corresponds to the result only in the loosest qualitative way. The difficulty of understanding the Stein paradox is compounded by the fact that its proof usually depends on explicit computation of the risk function or the theory of complete sufficient statistics, by a process that convinces us of its truth without really illuminating the reasons that it works. (The best presentation I know is that in Lehmann (1983, pages 300–302) of a proof due to Efron and Morris (1973); Berger (1980, page 165, example 54) outlines a short but unintuitive proof; the one shorter proof I have encountered in a textbook is vitiated by a major noncorrectable error.) The purpose of this paper is to show how a different perspective, one developed by Francis Galton over a century ago (Stigler, 1986, chapter 8), can render the result transparent, as well as lead to a simple, full proof. This perspective is perhaps closer to that of the period before 1950 than to subsequent approaches, but it has points in common with more recent works, particularly those of Efron and Morris (1973), Rubin (1980), Dempster (1980) and Robbins (1983).

2. STEIN ESTIMATION AS A REGRESSION PROBLEM

The estimation problem involves pairs of values (X_i, θ_i), $i = 1, \cdots, k$, where one element of each pair (X_i) is known and one (θ_i) is unknown. Since the θ_i's are unknown, the pairs cannot in fact be plotted, but it will help our understanding of the problem and suggest a means of approaching it if we imagine what such a plot would look like. Figure 1 is hypothetical, but some aspects of it accurately reflect the situation. Since X is $N(\theta, 1)$, we can think of the X's as being generated by adding $N(0, 1)$ "errors" to the given θ's. Thus the horizontal deviations of the points from the 45° line $\theta = X$ are independent $N(0, 1)$, and in that respect they should cluster around the line as indicated. Also, $E(\bar{X}) = \bar{\theta}$ and $\mathrm{Var}(\bar{X}) = 1/k$, so we should

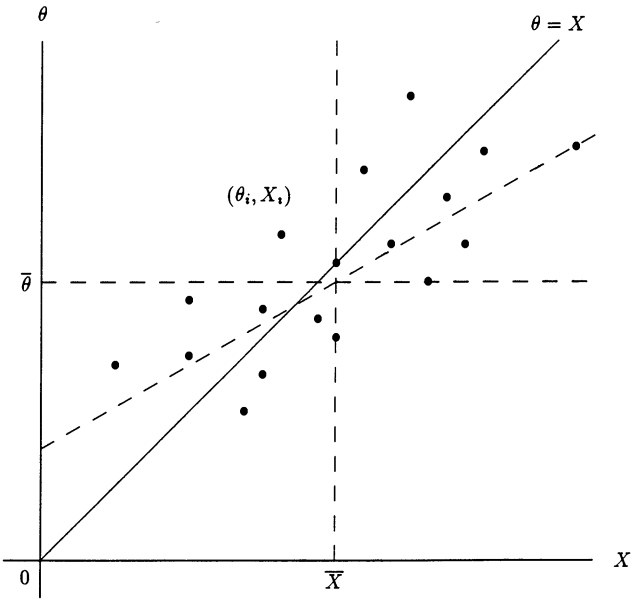

FIG. 1. *Hypothetical bivariate plot of θ_i vs. X_i, for $i = 1, \cdots, k$.*

expect the point of means $(\bar{X}, \bar{\theta})$ to lie near the 45° line.

Now our goal is to estimate all of the θ_i's given all of the X_i's, with no assumptions about a possible distributional structure for the θ_i's—they are simply to be viewed as unknown constants. Nonetheless, to see why we should expect that the ordinary estimator θ^0 can be improved upon, it helps to think about what we would do if this were not the case. If the θ_i's, and hence the pairs (X_i, θ_i), had a known joint distribution, a natural (and in some settings even optimal) method of proceeding would be to calculate $\hat{\theta}(X) = E(\theta \mid X)$ and use this, the theoretical regression function of θ on X, to generate estimates of the θ_i's by evaluating it for each X_i. We may think of this as an unattainable ideal, unattainable because we do not know the conditional distribution of θ given X. Indeed, we will not assume that our uncertainty about the unknown constants θ_i can be described by a probability distribution at all; our view is not that of either the Bayesian or empirical Bayesian approach. We *do* know the conditional distribution of X given θ, namely $N(\theta, 1)$, and we *can* calculate $E(X \mid \theta) = \theta$. Indeed this, the theoretical regression line of X on θ, corresponds to the line $\theta = X$ in Figure 1, and it is this line which gives the ordinary estimators $\hat{\theta}_i^0 = X_i$. Thus the ordinary estimator may be viewed as being based on the "wrong" regression line, on $E(X \mid \theta)$ rather than $E(\theta \mid X)$. Since, as Francis Galton already knew in the 1880's, the regressions of X on θ and of θ on X can be markedly different, this suggests that the ordinary estimator can be improved upon and even suggests how this might be done—by attempting to approximate "$E(\theta \mid \mathrm{X})$"—or whatever that might mean in a setting where the θ's do not have a distribution.

With no distributional assumptions about the θ's,

we are of course prevented from looking for an optimal estimate of "$E(\theta \mid X)$". Instead, we note that $\hat{\theta}_i^0 = X_i$ is a linear function of X_i, and we may look for a "best linear" estimator of the θ_i's, an estimator of the form

$$\hat{\theta}_i = a + bX_i, \quad i = 1, \cdots, k.$$

Now our goal is to minimize the loss function

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{k} (\theta_i - \hat{\theta}_i)^2,$$

and so *if* the θ_i's were actually available to us (that is, if the "data" plotted in Figure 1 were given to us), then we would be faced with a standard simple linear regression problem and the "best linear estimator" would clearly be the least squares line found by regressing $\boldsymbol{\theta}$ on $\mathbf{X}$, namely

$$\hat{\theta}_i = \bar{\theta} + \hat{\beta}(X_i - \bar{X}),$$

where

$$\sum \hat{\beta} \equiv \hat{\beta}_\theta = \frac{\sum (X_i - \bar{X})(\theta_i - \bar{\theta})}{\sum (X_i - \bar{X})^2}.$$

The θ_i's are not available, but if we can estimate the functions $\bar{\theta}$ and $\hat{\beta}$ of these unknown parameters, we will have an estimate of the regression line of $\boldsymbol{\theta}$ on $\mathbf{X}$; that regression line is both optimum for our loss function and a reasonable linear estimator of the ideal regression function $E(\theta \mid X)$.

The obvious (and Uniform Minimum Variance Unbiased) estimator for $\bar{\theta}$ is $\bar{X}$. To construct an estimator of the random parametric function $\hat{\beta}$, consider its numerator, $\sum (X_i - \bar{X})(\theta_i - \bar{\theta})$. Our approach is non-Bayesian, but we can nonetheless use Bayesian calculations to help guide us. Suppose then for a moment that the θ_i's are independently distributed according to some distribution, *any* distribution (known or unknown) with a finite second moment. (For example, motivated by Efron's Bootstrap, the θ_i's could even be supposed to be distributed as a random sample taken with replacement from the list of the actual (unknown, fixed) values of the θ_i's.)

Then the sample covariance

$$\frac{1}{k-1} \sum (X_i - \bar{X})(\theta_i - \bar{\theta})$$

is an unbiased estimator of $\operatorname{cov}(X, \theta)$. Furthermore, since

$$X = \theta + \varepsilon$$

where ε is $N(0, 1)$, independent of θ, $\operatorname{var}(X) = \operatorname{var}(\theta) + \operatorname{var}(\varepsilon)$ and we have

$$\begin{aligned} \operatorname{cov}(X, \theta) &= \operatorname{var}(\theta) \\ &= \operatorname{var}(X) - \operatorname{var}(\varepsilon) \\ &= \operatorname{var}(X) - 1, \end{aligned}$$

where $\operatorname{var}(X)$ is computed for the marginal distribution of X. But an unbiased estimator of $\operatorname{var}(X)$ is

$$\frac{1}{k-1} \sum (X_i - \bar{X})^2,$$

and thus an unbiased estimator of $\operatorname{cov}(X, \theta)$ is

$$\frac{1}{k-1} \sum (X_i - \bar{X})^2 - 1.$$

That is, regardless of the supposed distribution of the θ_i,

$$\sum (X_i - \bar{X})(\theta_i - \bar{\theta})$$

and

$$\sum (X_i - \bar{X})^2 - (k-1)$$

both have the same expectation. Indeed, reverting to our non-Bayesian perspective, where the θ_i's are simply fixed constants, it is easy to see that the same is true there: $E[\sum (X_i - \bar{X})^2 - (k-1)] = E[\sum (X_i - \bar{X})(\theta_i - \bar{\theta})] = \sum (\theta_i - \bar{\theta})^2$. This suggests estimating the random parametric function $\hat{\beta}$ by

$$\begin{aligned} \frac{\sum (X_i - \bar{X})^2 - (k-1)}{\sum (X_i - \bar{X})^2} &= 1 - \frac{k-1}{\sum (X_i - \bar{X})^2} \\ &= 1 - \frac{k-1}{S'^2}, \end{aligned}$$

which leads to the estimated least squares line

$$\hat{\theta}_i^{\mathrm{EM}} = \bar{X} + \left(1 - \frac{k-1}{S'^2}\right)(X_i - \bar{X}).$$

But this is just the Efron–Morris estimator, with $c = k - 1$; it is not the best choice of c, but it has risk uniformly smaller than the "ordinary" estimator as long as $k - 1 < 2(k-3)$, or $k > 5$.

The James–Stein estimator can be derived by a similar route, by considering the class of estimators that are linear in X with zero intercept,

$$\hat{\theta}_i = bX_i.$$

Then the least squares estimator has

$$\hat{\beta} = \frac{\sum \theta_i X_i}{\sum X_i^2},$$

and $\theta_i X_i$ and $\sum X_i^2 - k$ have the same expectation ($\sum \theta_i^2$), leading to the James–Stein estimator with $c = k$,

$$\hat{\theta}_i^{\mathrm{JS}} = \left(1 - \frac{k}{S^2}\right) X_i.$$

3. INTERPRETATION AND EXTENSION

This Galtonian perspective on the Stein paradox renders it nearly transparent. The "ordinary"

estimators $\hat{\theta}_i^{\circ} = X_i$ are derived from the theoretical regression line of X on θ. That line would be useful if our goal were to predict X from θ, but our problem is the reverse, namely to predict θ from X using the sum of squared errors $\sum (\theta_i - \hat{\theta}_i)^2$ as a criterion. For that criterion, the optimum linear estimators are given by the least squares regression line of θ on X, and the James–Stein and Efron–Morris estimators are themselves estimators of that optimum linear estimator. The "ordinary" estimators are derived from the wrong regression line, the James–Stein and Efron–Morris estimators are derived from approximations to the right regression line. We can even see why $k \geq 3$ is necessary: if $k = 1$ or 2, the least squares line of θ on X must pass through the points (X_i, θ_i), and hence for $k = 1$ or 2, the two regression lines (of X on θ and of θ on X) must agree at each X_i. Thus the ordinary estimators $\hat{\theta}_i^{\circ} = X_i$, although they lie on the wrong theoretical regression line, approximate either least squares line equally well.

This regression perspective not only makes the logic of the procedures clear, it also leads to a short, rigorous proof of the phenomenon. *If* we could actually use (for $k \geq 3$) estimators derived from either regression line

$$\hat{\theta}_i = \bar{\theta} + \hat{\beta}(X_i - \bar{X})$$

or

$$\hat{\theta}_i = \hat{\beta} X_i,$$

then we would improve upon the ordinary estimators not only in the sense that we would have lower risk or *expected* loss (averaged over the possible values of the X_i), but even in the much stronger sense that we would have lower *actual* loss $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ for all possible values of the X_i (barring the unlikely event that the ordinary estimators actually fall on the regression line).

To see what is involved in turning this perspective into a full proof of the phenomenon, and to gain deeper understanding of why the approximation works, let us look more carefully at the case of the James–Stein estimator. We consider three representatives of the class of linear estimators with zero intercept, $\hat{\theta}_i^b = bX_i$, namely the ordinary estimator

$$\hat{\theta}_i^{\circ} = X_i,$$

the least squares estimator

$$\hat{\theta}_i^{\mathrm{LS}} = \hat{\beta} X_i,$$

where

$$\hat{\beta} = \sum X_i \theta_i / \sum X_i^2,$$

and the James–Stein estimators we found to approximate $\hat{\boldsymbol{\theta}}^{\mathrm{LS}}$,

$$\hat{\theta}_i^{\mathrm{JS}} = \left(1 - \frac{c}{S^2}\right) X_i$$

$$= \hat{b} X_i, \quad \text{say.}$$

Now let

$$L(\boldsymbol{\theta}, \hat{\beta}\mathbf{X}) = RSS_{\mathrm{LS}},$$

the minimum attainable loss within this class of estimators. For other estimators $\hat{\boldsymbol{\theta}}^b$ we have

$$\begin{aligned} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^b) &= \sum (\theta_i - \hat{\theta}_i^b)^2 \\ &= \sum (\theta_i - \hat{\theta}_i^{\mathrm{LS}} + \hat{\theta}_i^{\mathrm{LS}} - \hat{\theta}_i^b)^2 \\ &= RSS_{\mathrm{LS}} + \sum (\hat{\theta}_i^{\mathrm{LS}} - \hat{\theta}_i^b)^2 \\ &= RSS_{\mathrm{LS}} + (\hat{\beta} - b)^2 S^2. \end{aligned}$$

Thus

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^b) = E(RSS_{\mathrm{LS}}) + E[(\hat{\beta} - b)^2 S^2],$$

and we see that a James–Stein estimator will improve on the ordinary estimator if and only if

$$E[(\hat{\beta} - \hat{b})^2 S^2] < E[(\hat{\beta} - 1)^2 S^2];$$

that is, if and only if $\hat{b}$ is closer to the least squares slope $\hat{\beta}$ than is the constant 1, in this average (weighted by S^2) sense, for all $\boldsymbol{\theta}$. Since $\hat{b}$ is a "reasonable" estimator of $\hat{\beta}$, while 1.0 is not, we should not be surprised that this is the case.

A relatively simple proof that this is indeed the case can be obtained as follows: Let $\hat{b}_c = 1 - c/S^2$, so that $1 - \hat{b}_c = c/S^2$, and look at

$$\begin{aligned} &E[(\hat{\beta} - \hat{b}_c)^2 S^2] - E[(\hat{\beta} - 1)^2 S^2] \\ &\quad = E[(\hat{\beta} - \hat{b}_c)^2 - (\hat{\beta} - 1)^2] S^2 \\ &\quad = E((1 - \hat{b}_c)(2\hat{\beta} - 1 - \hat{b}_c) S^2) \\ &\quad = 2c\, E\left(\frac{\sum X_i \theta_i - S^2 + c/2}{S^2}\right) \\ &\quad = 2c \left[E\left(\frac{\sum X_i \theta_i + c/2}{S^2}\right) - 1\right]. \end{aligned}$$

To prove the superiority of the James–Stein estimator, it is clearly enough to show that

$$E\left(\frac{\sum X_i \theta_i + c/2}{S^2}\right) \leq 1$$

for all θ_i, all $k \geq 3$, $0 \leq c \leq 2(k-2)$. The left-hand side of this inequality is monotone increasing in c, so it is sufficient to establish that

$$E\left(\frac{\sum X_i \theta_i + (k-2)}{S^2}\right) \equiv 1$$

for all θ, all $k \geq 3$. The lemma given in the Appendix shows just that.

Essentially the same proof also shows that the Efron–Morris estimator $\hat{\theta}^{EM}$ dominates the ordinary estimator $\hat{\theta}^{\circ}$ if $k \geq 4$. Let

$$\hat{\theta}_i^{a,b} = a + b(X_i - \bar{X}).$$

This class includes the Efron—Morris estimators, for which

$$a^{EM} = \bar{X}, \quad b^{EM} = \left(1 - \frac{c}{S'^2}\right),$$

and the ordinary estimator, where

$$a^{\circ} = \bar{X}, \quad b^{\circ} = 1,$$

and the least squares line for θ on $\mathbf{X}$, where

$$a^{LS} = \bar{\theta},$$
$$b^{LS} = \sum (X_i - \bar{X})(\theta_i - \bar{\theta})/S'^2.$$

Then, as before, we have

$$\begin{aligned} L(\theta, \hat{\theta}^{ab}) &= \sum (\theta_i - \hat{\theta}_i^{a,b})^2 \\ &= RSS_{LS} + \sum (\hat{\theta}_i^{LS} - \hat{\theta}_i^{a,b})^2 \\ &= RSS_{LS} \\ &\quad + \sum ((b^{LS} - b)(X_i - \bar{X}) + (a^{LS} - a))^2 \\ &= RSS_{LS} + (b^{LS} - b)^2 S'^2 + (a^{LS} - a)^2. \end{aligned}$$

Then since $a^{EM} = a^{\circ}$, the risk function of $\hat{\theta}^{EM}$ will dominate that of $\hat{\theta}^{\circ}$ if and only if

$$E[(b^{LS} - b^{EM})^2 S'^2] \leq E[(b^{LS} - 1)^2 S'^2],$$

or, equivalently,

$$E[(1 - b^{EM})(2b^{LS} - 1 - b^{EM})S'^2] \leq 0$$

for all θ. But $(1 - b^{EM})S'^2 = c$, and so this is equivalent to having

$$2c\left\{E\left[\frac{\sum (X_i - \bar{X})(\theta_i - \bar{\theta}) + c/2}{S'^2}\right] - 1\right\} \leq 0,$$

or

$$E\left[\frac{\sum (X_i - \bar{X})(\theta_i - \bar{\theta}) + c/2}{S'^2}\right] \leq 1.$$

Now an elementary matrix decomposition gives us $\mathbf{X} - \bar{X}\mathbf{1} = \mathbf{U}'\mathbf{D}\mathbf{U}$, where $\mathbf{U}$ is orthogonal and $\mathbf{D} = \text{diag}(1, 1, \cdots, 1, 0)$. (See, e.g., Searle, 1982, page 352). Then if $\mathbf{Y} = \mathbf{U}\mathbf{X}$, the Y_i's are normal with expectations μ_i (say) and variances 1, and we have

$$\frac{\sum (X_i - \bar{X})(\theta_i - \bar{\theta}) + c/2}{S'^2} = \frac{\sum_{i=1}^{k-1} Y_i \mu_i + c/2}{\sum_{i=1}^{k-1} Y_i^2}.$$

It follows from the lemma of the Appendix that for $0 < c \leq 2(k-3)$,

$$\begin{aligned} &E\left(\frac{\sum (X_i - \bar{X})(\theta_i - \bar{\theta}) + c/2}{S'^2}\right) \\ &\leq E\left(\frac{\sum (X_i - \bar{X})(\theta_i - \bar{\theta}) + (k-3)}{S'^2}\right) \\ &= E\left(\frac{\sum_{i=1}^{k-1} Y_i \mu_i + (k-3)}{\sum_{i=1}^{k-1} Y_i^2}\right) \\ &\equiv 1 \text{ all } \theta, \quad \text{all } k \geq 4. \end{aligned}$$

It may be noted that the essence of these proofs is a "swindle" akin to those that have proved useful in Monte Carlo studies (e.g., Andrews et al., 1972; Simon, 1976). A basic property of least squares estimators (the orthogonality of the fitted values and the residuals) is exploited to separate out the common term RSS_{LS} from $L(\theta, \hat{\theta}^{ab})$ for all a, b, permitting different linear estimators to be compared without the necessity of evaluating $E(RSS_{LS})$.

In the above development we have supposed the variances of the X_i's are equal. The case of unequal variances, which would be of interest in many practical situations, presents serious mathematical difficulties that have not been surmounted except for specially weighted loss functions. It is not clear how the present approach can shed additional light upon this problem. The issues involved, and an approach that is useful in practice, are discussed in Morris (1983).

4. THE POISSON CASE

The regression perspective can also be used to motivate shrinkage estimators for the Poisson case. Suppose (following Clevenson and Zidek, 1975) that $Z_1, Z_2, \cdots, Z_k$ are independent, where Z_i has a Poisson (λ_i) distribution, and we wish to estimate $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_k)'$ with loss function

$$L^*(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) = \sum_{i=1}^{k} \frac{(\hat{\lambda}_i - \lambda_i)^2}{\lambda_i}.$$

Clevenson and Zidek (1975) have shown that the "ordinary" estimator $\hat{\lambda}_i^{\circ} = Z_i$ is dominated by the shrinkage estimator

$$\hat{\boldsymbol{\lambda}}^{CZ} = b^{CZ}\mathbf{Z}, \tag{4.1}$$

where

$$b^{CZ} = \frac{\sum Z_i}{c + \sum Z_i},$$

as long as $k-1 \le c \le 2(k-1)$ and $k \ge 2$, in the sense that

$$E(L^*(\hat{\lambda}^{CZ}, \lambda)) \le E(L^*(\hat{\lambda}^{\circ}, \lambda))$$

for all λ. (See also Berger, 1985, §5.44.)

As was true in the normal case, $E(Z_i \mid \lambda_i) = \lambda_i$, and the estimator $\hat{\lambda}^{\circ}$ can be viewed as based upon the "wrong" regression line, that of Z on λ. The proper regression here would be a weighted regression of λ on Z. If we limit our attention to the linear estimators $\hat{\lambda}^b = b\mathbf{Z}$, then we see that within this class, the loss

$$L^*(\hat{\lambda}^b, \lambda) = \sum \frac{(bZ_i - \lambda_i)^2}{\lambda_i}$$

$$= \sum [b(Z_i/\sqrt{\lambda_i}) - \sqrt{\lambda_i}]^2$$

is minimized by the weighted least squares choice,

$$b^{LS} = \sum Z_i / \sum (Z_i^2/\lambda_i).$$

Now, $E(\sum (Z_i^2/\lambda_i)) = \sum (\lambda_i + \lambda_i^2)/\lambda_i = k + \sum \lambda_i$ can be estimated by $k + \sum Z_i$, suggesting the estimator with

$$b = \sum Z_i / (k + \sum Z_i),$$

which is a Clevenson-Zidek estimator with $c = k$, and dominates $\hat{\lambda}_i^{\circ}$ as long as $k \ge 2$.

Similarly to the normal case, we can write for the class of linear estimators

$$L^*(\hat{\lambda}^{\circ}, \lambda) - L^*(\hat{\lambda}^b, \lambda) = (1-b)(1+b-2b^{LS}) \sum_{i=1}^{k} \frac{Z_i^2}{\lambda_i}, \tag{4.2}$$

and if $b < 1$ (as is true for Clevenson–Zidek estimators), we have

$$L^*(\hat{\lambda}^{\circ}, \lambda) \le L^*(\hat{\lambda}^b, \lambda)$$

as long as $(1+b)/2 \ge b^{LS}$; that is, as long as $|b - b^{LS}| \le |1-b|$, or as long as b is closer to the weighted least squares slope b^{LS} than to the slope of the wrong regression line, 1.

Robbins (1983, page 722) outlines (in the tantalizing manner of Fermat, who could not fit the crucial details of his "last theorem" into the margin of a book) an "elementary" proof that $\hat{\lambda}^{CZ}$ dominates $\hat{\lambda}^{\circ}$ for $c = k$, and indeed Clevenson and Zidek's original proof is short, elegant and covers a broader class of estimators. A simple proof based on (4.2) is not hard to derive: First exploit the relationship between the Poisson and binomial distributions (i.e., given $\bar{Z}$, Z_i is binomial) to show that $\sum E(Z_i^2/\lambda_i \mid \bar{Z}) = \bar{Z}[k-1+k\bar{Z}]/\bar{\lambda}$; then (4.2) gives

$$E(L^*(\hat{\lambda}, \lambda) - L^*(\hat{\lambda}^b, \lambda) \mid \bar{Z}) = \frac{c}{\bar{\lambda}}(k-1+k\bar{Z})(1+b^{CZ})b^{CZ} - 2kcb^{CZ}. \tag{4.3}$$

Now (4.3) is a convex function of $\bar{Z}$ as long as $\bar{Z} \ge 0$ and $c \ge k-1$ (differentiate), and Jensen's Inequality implies that the expected value of (4.3) is bounded below by this same expression, where $\bar{Z}$ is replaced by $\bar{\lambda}$; this lower bound is easily seen to be positive as long as $c \le 2(k-1)$. The best recent treatment of this topic in a general setting is that of Ghosh, Hwang and Tsui (1983).

5. SOME HISTORICAL BACKGROUND

The perspective advanced here is far from new, although the proof that emerges from this development appears to be new. The use of least squares estimators for the adjustment of data of course goes back well into the previous century, as does Galton's more subtle idea that there are two regression lines (Stigler, 1986, Chapter 8). Earlier in this century, regression was employed in educational psychology in a setting quite like that considered here. Truman Kelley developed models for ability which hypothesized that individuals had true scores (think of our θ_i's) measured by fallible testing instruments to give observed scores (our X_i's); the observed scores could be improved as estimates of the true scores by allowing for the regression effect and shrinking toward the average, by a procedure quite similar to the Efron–Morris estimator. (Kelley, 1923, pages 212–214; Kelley in effect assumed the means and covariances known, and of course proved no result of the type Stein was to discover.) The approach of the present paper is directly in line with this literature and recent developments of it, in particular by Rubin (1980), as was most clearly realized by Dempster (1980).

Modern work on this topic from a decision theoretic point of view was initiated by the pathbreaking work of Stein (1956) and James and Stein (1961). Stein's original paper seems to have been motivated by the observation that (in our notation)

$$E\left(\sum_{i=1}^{k} X_i^2\right) = \sum_{i=1}^{k} \theta_i^2 + k,$$

and so when $\sum X_i^2 = C$ is observed, we should estimate the θ_i's to be such that $\sum \theta_i^2 \cong C - k$; since the "ordinary" estimator would estimate $\sum \theta_i^2 \cong C$, we should shrink the components to compensate for this over-estimation.

In a series of important papers in the early 70's, Efron and Morris recast the problem in the empirical Bayes framework and explored the properties of the Efron–Morris estimator $\hat{\theta}^{EM}$ (Efron and Morris, 1972, 1973), which had been earlier suggested by Lindley (1962). In his 1982 Neyman lecture, Herbert Robbins (the originator of much of the area of empirical Bayes) developed the Efron–Morris estimator within the

empirical Bayes framework by a route that parallels that of Section 2 above quite closely (Robbins, 1983). It was the excellent review of parametric empirical Bayes by Morris (1983) that led me to consider the problem from the regression point of view (Stigler, 1983) and eventually to the present paper.

The vast literature on the Stein paradox is surveyed by Berger (1985, pages 359–369; 1988). Good textbook treatments can be found; for example Lehmann (1983, §4.6) and, from a Bayesian perspective, Hartigan (1983, Chapter 9). A nice general introduction was given by Efron and Morris (1977).

APPENDIX

LEMMA. *Let X_i, $i = 1, 2, \cdots, k+1$ be independent normally distributed random variables, $X_i \sim N(\theta_i, 1)$. Then*

$$E\left(\frac{\sum_{i=1}^{k+1} \theta_i X_i + (k-1)}{\sum_{i=1}^{k+1} X_i^2}\right) = 1$$

for all $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{k+1})'$, all $k \geq 2$.

PROOF. Let $\theta = (\sum_{i=1}^{k+1} \theta_i^2)^{1/2}$. Now without loss of generality we may take $\boldsymbol{\theta} = (\theta, 0, 0, \cdots, 0)'$, since the problem can be reduced to this case by a simple rotation transformation in E_{k+1} (because $\sum \theta_i X_i$ depends only upon the angle between $\boldsymbol{\theta}$ and $\mathbf{X}$). Let $X = X_1$ and $Y = \sum_{i=2}^{k+1} X_i^2$; then the problem becomes this: Show that

$$E\left(\frac{X\theta + (k-1)}{X^2 + Y}\right) = 1,$$

for all $k \geq 2$, all $\theta \in R$,

where X and Y are independent, $X \sim N(\theta, 1)$ and $Y \sim \chi^2(k)$. The proof is essentially a simple one: Transform to polar coordinates, integrate by parts, transform back to rectangular coordinates, and the identity is obvious.

The joint distribution of X and Y has density

$$f_{X,Y}(x, y) = C_k \exp(-\tfrac{1}{2}[x^2 + y - 2\theta x + \theta^2])\, y^{k/2-1},$$

for $y > 0$, $-\infty < x < \infty$,

where

$$C_k = \left(2^{k/2}\Gamma\left(\frac{k}{2}\right)\sqrt{2\pi}\right)^{-1}.$$

Make the transformation

$$w = \frac{x}{\sqrt{x^2 + y}}, \quad z = \sqrt{x^2 + y},$$

whose inverse transformation is given by

$$x = wz, \quad y = z^2(1 - w^2),$$

and whose Jacobian is

$$|J| = \begin{vmatrix} z & w \\ -2wz^2 & 2z(1-w^2) \end{vmatrix} = 2z^2.$$

Thus the joint density of W and Z is

$$f_{W,Z}(w, z) = C_k \exp\left(-\frac{z^2}{2} + \theta wz - \frac{\theta^2}{2}\right) \cdot [z^2(1-w^2)]^{k/2-1}\, 2z^2,$$

for $-1 \leq w < 1$, $z > 0$,

and

$$E\left(\frac{\theta X + (k-1)}{X^2 + Y}\right)$$

$$= C_k \int_{-1}^{1}\int_{0}^{\infty} \frac{(\theta wz + (k-1))}{z^2} f_{W,Z}(w, z)\, dz\, dw$$

$$= 2C_k \int_{-1}^{1} (1-w^2)^{k/2-1} e^{-(\theta^2/2)(1-w^2)} \cdot \left[\theta w \int_0^\infty z^{k-1} e^{-(z-\theta w)^2/2}\, dz + \int_0^\infty e^{-(z-\theta w)^2/2}\, dz^{k-1}\right] dw$$

$$= 2C_k \int_{-1}^{1} (1-w^2)^{k/2-1} e^{-(\theta^2/2)(1-w^2)} \int_0^\infty z^k e^{-(z-\theta w)^2/2}\, dz\, dw$$

(after integration by parts)

$$= 2C_k \int_0^1 \int_{-\infty}^{\infty} (1-w^2)^{k/2-1} (z^2)^{k/2} e^{-z^2/2 + zw\theta - \theta^2/2}\, dz\, dw,$$

by taking advantage of symmetry of the integrand (its integral over $\int_{-1}^{0}\int_{0}^{\infty}$ equals that over $\int_0^1 \int_{-\infty}^0$). Now transform back to X and Y; there the Jacobian is

$$|J| = \begin{vmatrix} y(x^2+y)^{-3/2} & -\tfrac{1}{2}x(x^2+y)^{-3/2} \\ x(x^2+y)^{-1/2} & \tfrac{1}{2}(x^2+y)^{-1/2} \end{vmatrix} = \tfrac{1}{2}(x^2+y)^{-1},$$

and so (since $1 - w^2 = y/(x^2+y)$ and $z^2 = x^2 + y$)

$$E\left(\frac{\theta X + (k-1)}{X^2 + Y}\right)$$

$$= 2C_k \int_0^\infty \int_{-\infty}^{\infty} \left(\frac{y}{x^2+y}\right)^{k/2-1} (x^2+y)^{k/2} \cdot \exp\left(-\frac{(x-\theta)^2}{2} - \frac{y}{2}\right)\frac{1}{2}(x^2+y)^{-1}\, dx\, dy$$

$$= C_k \int_0^\infty \int_{-\infty}^{\infty} y^{k/2-1} \exp\left(-\frac{y}{2}\right)\exp\left(-\frac{(x-\theta)^2}{2}\right) dx\, dy$$

$$= \int_0^\infty \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy$$

$$= 1. \qquad \square$$

ANOTHER PROOF. A shorter proof of this lemma can be based upon a classical identity of the theory of estimation, although it has the disadvantage of pushing the magical property of the normal distribution that makes the lemma "work" further out of sight. The identity in question is

$$\frac{d}{d\theta_i} Eh(\mathbf{X}) = \operatorname{cov}\left(h(\mathbf{X}), \frac{d}{d\theta_i} \log p_{\boldsymbol{\theta}}(\mathbf{X})\right),$$

where $p_{\boldsymbol{\theta}}(\mathbf{X})$ is the joint density of $\mathbf{X}$, and h is any function of $\mathbf{X}$ (subject to mild regularity conditions). This identity was in common use by the mid-1940's in connection with proofs of the information inequality; see Lehmann (1983, pages 117, 129 and 145) for references. For the special case considered here, where the X_i's have independent normal densities with unit variances, this identity becomes

$$\begin{aligned}\frac{d}{d\theta_i} Eh(\mathbf{X}) &= \operatorname{cov}(h(\mathbf{X}), (X_i - \theta_i)) \\ &= Eh(\mathbf{X})(X_i - \theta_i).\end{aligned}$$

Then (letting $\mathbf{U} = \mathbf{X} - \boldsymbol{\theta}$, and using Fubini's theorem)

$$\begin{aligned}\frac{d}{d\theta_i} Eh(\mathbf{X}) &= \frac{d}{d\theta_i} Eh(\mathbf{U} + \boldsymbol{\theta}) \\ &= E \frac{d}{d\theta_i} h(\mathbf{U} + \boldsymbol{\theta}) \\ &= E \frac{d}{dX_i} h(\mathbf{X}),\end{aligned}$$

and we have

$$E \frac{d}{dX_i} h(\mathbf{X}) = Eh(\mathbf{X})(X_i - \theta_i).$$

In this form, the identity has been ingeniously exploited by Stein and his students in studying the estimation of the multivariate normal mean (Stein, 1981). This identity leads to the lemma as follows: Let

$$h(\mathbf{X}) = \frac{X_i}{S^2},$$

where $S^2 = \sum_{i=1}^{k+1} X_i^2$. Then $(d/dX_i)h(\mathbf{X}) = (S^2 - 2X_i^2)/S^4$, and the identity gives

$$E\left(\frac{1}{S^2} - \frac{2X_i^2}{S^4}\right) = E\left(\frac{X_i^2}{S^2} - \frac{\theta_i X_i}{S^2}\right).$$

Sum both sides from $i = 1$ to $k + 1$ to get

$$E\left(\frac{k+1}{S^2} - \frac{2S^2}{S^4}\right) = E\left(\frac{S^2}{S^2} - \frac{\sum \theta_i X_i}{S^2}\right)$$

or

$$E\left(\frac{k-1}{S^2}\right) = 1 - E\left(\frac{\sum \theta_i X_i}{S^2}\right),$$

which is what we wish to prove. □

ACKNOWLEDGMENTS

This work was presented as a portion of the author's Jerzy Neyman Memorial Lecture at the Annual Meeting of the I.M.S. in Fort Collins, Colorado in August 1988. I am grateful to R. R. Bahadur, Carl Morris and Wing Wong for their insightful comments. This research was supported in part by National Science Foundation Grant DMS-86-01732. This manuscript was prepared using computer facilities supported in part by National Science Foundation Grants DMS-86-01732 and DMS-87-03942 to the Department of Statistics at The University of Chicago.

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location.* Princeton Univ. Press, Princeton, N.J.

BERGER, J. O. (1980). *Statistical Decision Theory.* Springer, New York.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERGER, J. O. (1988). The Stein effect. *Encyclopedia of the Statistical Sciences* **8** 757–761. Wiley, New York.

CLEVENSON, M. L. and ZIDEK, J. V. (1975). Simultaneous estimation of the mean of independent Poisson laws. *J. Amer. Statist. Assoc.* **70** 698–705.

DEMPSTER, A. P. (1980). Comment on "Using empirical Bayes techniques in the law school validity studies," by D. B. Rubin. *J. Amer. Statist. Assoc.* **75** 817.

EFRON, B. and MORRIS, C. N. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139.

EFRON, B. and MORRIS, C. N. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

EFRON, B. and MORRIS, C. N. (1977). Stein's paradox in statistics. *Scientific American* **236** (5) 119–127.

GHOSH, M., HWANG, J. T. and TSUI, K.-W. (1983). Construction of improved estimators in multiparameter estimation for discrete exponential families (with discussion). *Ann. Statist.* **11** 351–376.

HARTIGAN, J. A. (1983). *Bayes Theory.* Springer, New York.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–379. Univ. California Press.

KELLEY, T. L. (1923). *Statistical Method.* Macmillan, New York.

LEHMANN, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

LINDLEY, D. V. (1962). Discussion of "Confidence sets for the mean of a multivariate normal distribution," by C. Stein. *J. Roy. Statist. Soc. Ser. B* **24** 285–287.

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

ROBBINS, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713–723.

RUBIN, D. B. (1980). Using empirical Bayes techniques in the Law School Validity Studies (with discussion). *J. Amer. Statist. Assoc.* **75** 801–827.

SEARLE, S. R. (1982). *Matrix Algebra Useful for Statistics.* Wiley, New York.

SIMON, G. (1976). Computer simulation swindles, with applications to estimates of location and dispersion. *Appl. Statist.* **25** 266–274.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197–206. Univ. California Press.

STEIN, C. (1962). Confidence sets for the mean of a multivariate normal distribution (with discussion). *J. Roy. Statist. Soc. Ser. B* **24** 265–296.

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

STIGLER, S. M. (1983). Comment on "Parametric empirical Bayes inference: Theory and applications," by C. N. Morris. *J. Amer. Statist. Assoc.* **78** 62–63.

STIGLER, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* The Belknap Press of Harvard Univ. Press, Cambridge, Mass.