OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

Biometrika Trust

Empirical Bayes on Vector Observations: An Extension of Stein's Method
Author(s): Bradley Efron and Carl Morris
Source: *Biometrika,* Vol. 59, No. 2 (Aug., 1972), pp. 335-347
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: https://www.jstor.org/stable/2334578
Accessed: 19-09-2018 17:59 UTC

Empirical Bayes on vector observations: An extension of Stein's method

BY BRADLEY EFRON
Stanford University
AND
CARL MORRIS
Rand Corporation, Santa Monica

SUMMARY

The statistician is considering several independent normal linear models with identical structures, and desires to estimate the vector of unknown parameters in each of them. An estimator is constructed which dominates the usual Gauss–Markov estimator in terms of total squared error loss. This estimator is shown to have good efficiency in the Bayesian situation where the parameter vectors themselves have a normal prior distribution. A practical example is given.

Some key words: Empirical Bayes; Vectors; Linear models; Two-way analysis of variance; Stein's estimator; Multivariate analysis.

1. INTRODUCTION

Suppose the statistician is simultaneously considering k independent problems, each of which involves estimating a vector $\boldsymbol{\theta}_i$ having p co-ordinates $(i = 1, \ldots, k)$. The ith problem provides data $\mathbf{x}_i$ which we will assume is also in the form of a p co-ordinate vector, normally distributed with mean vector $\boldsymbol{\theta}_i$ and covariance matrix the identity matrix, all k problems yielding mutually independent data vectors, i.e.

$$\mathbf{x}_i \sim \mathcal{N}_p(\boldsymbol{\theta}_i, \mathbf{I}) \quad (i = 1, \ldots, k). \tag{1·1}$$

For example, $\boldsymbol{\theta}_i$ might represent the p parameters of a regression model run separately in k different situations, and $\mathbf{x}_i$ the usual Gauss–Markov estimator for $\boldsymbol{\theta}_i$. In this case the covariance matrix of $\mathbf{x}_i$ would usually not equal $\mathbf{I}$, and might have to be estimated from the data. This complication does not affect the theory which follows, but for the sake of simple presentation will be dealt with separately in § 4.

The usual estimator for this situation is the maximum likelihood estimator, which estimates $\boldsymbol{\theta}_i$ by $\boldsymbol{\delta}_i^0 = \mathbf{x}_i$ $(i = 1, \ldots, k)$. Suppose we use a normalized squared error loss function in estimating the entire $p \times k$ matrix of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ by the corresponding matrix of estimates $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k)$,

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{pk} \operatorname{tr}\{(\boldsymbol{\theta} - \boldsymbol{\delta})'(\boldsymbol{\theta} - \boldsymbol{\delta})\} = \frac{1}{pk} \sum_{i=1}^{k} \sum_{j=1}^{p} (\theta_{ij} - \delta_{ij})^2. \tag{1·2}$$

Then the maximum likelihood estimator $\boldsymbol{\delta}^0$ has risk, expected loss, $R(\boldsymbol{\theta}, \boldsymbol{\delta}^0) = 1$ for all values of $\boldsymbol{\theta}$.

If the statistician had the additional information that the $\boldsymbol{\theta}_i$ vectors were themselves selected independently from a multivariate normal distribution, say

$$\boldsymbol{\theta}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{A}) \quad (i = 1, \ldots, k), \tag{1·3}$$

where $\mathbf{A}$ was a known $p \times p$ covariance matrix, he could use the Bayes rule $\boldsymbol{\delta}_i^* = (\mathbf{I} - \mathbf{B})\,\mathbf{x}_i$ $(i = 1, \ldots, k)$, where $\mathbf{B} = (\mathbf{A} + \mathbf{I})^{-1}$. Since given $\mathbf{x}_i$, $\boldsymbol{\theta}_i$ has the conditional distribution

$$\boldsymbol{\theta}_i | \mathbf{x}_i \sim \mathcal{N}_p\{(\mathbf{I} - \mathbf{B})\mathbf{x}_i, (\mathbf{I} - \mathbf{B})\} \quad (i = 1, \ldots, k), \tag{1·4}$$

the Bayes rule has Bayes risk

$$R(\mathbf{A}, \boldsymbol{\delta}^*) \equiv E_{\mathbf{A}}\{R(\boldsymbol{\theta}, \boldsymbol{\delta}^*)\} = \frac{1}{p}\operatorname{tr}(\mathbf{I} - \mathbf{B}), \tag{1·5}$$

where for any rule $\boldsymbol{\delta}$, $R(\mathbf{A}, \boldsymbol{\delta}) = E_{\mathbf{A}}\{R(\boldsymbol{\theta}, \boldsymbol{\delta})\}$ indicates the expectation of $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ with respect to the distribution on $\boldsymbol{\theta}$ given in (1·3). In general the notation $E_{\mathbf{A}}\{h(\boldsymbol{\theta}, \mathbf{x})\}$ will indicate the expectation of $h(\boldsymbol{\theta}, \mathbf{x})$ with respect to the joint distribution on $(\boldsymbol{\theta}, \mathbf{x})$ defined by (1·3) and (1·1). We will also use the simple notations $E_{\boldsymbol{\theta}}\{h(\boldsymbol{\theta}, \mathbf{x})\}$ and $E_{\mathbf{x}}\{h(\boldsymbol{\theta}, \mathbf{x})\}$ in place of $E_{\mathbf{A}}\{h(\boldsymbol{\theta}, \mathbf{x}) | \boldsymbol{\theta}\}$ and $E_{\mathbf{A}}\{h(\boldsymbol{\theta}, \mathbf{x}) | \mathbf{x}\}$, respectively.

In this paper we suggest the rule

$$\boldsymbol{\delta}_i^1 = \{\mathbf{I} - (k - p - 1)\,\mathbf{S}^{-1}\}\,\mathbf{x}_i, \tag{1·6}$$

where, if $\mathbf{X}$ be the $p \times k$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, $\mathbf{S} = \mathbf{X}\mathbf{X}'$. We are assuming $p + 1 < k$, and will continue to do so in the remainder of this paper.

We will show that this rule dominates the maximum likelihood estimator in the sense that

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^1) < R(\boldsymbol{\theta}, \boldsymbol{\delta}^0) = 1$$

for every value of $\boldsymbol{\theta}$. Moreover, in the Bayesian situation (1·3) the rule $\boldsymbol{\delta}^1$, which does not require knowledge of $\mathbf{A}$, compares well with $\boldsymbol{\delta}^*$ if $(p + 1)/k$ is small. To be specific, define the *relative savings loss* of any rule $\boldsymbol{\delta}$ *versus* the prior (1·3) as

$$RSL(\mathbf{A}, \boldsymbol{\delta}) = \frac{R(\mathbf{A}, \boldsymbol{\delta}) - R(\mathbf{A}, \boldsymbol{\delta}^*)}{R(\mathbf{A}, \boldsymbol{\delta}^0) - R(\mathbf{A}, \boldsymbol{\delta}^*)} = \frac{p}{\operatorname{tr}(\mathbf{B})}\left\{R(\mathbf{A}, \boldsymbol{\delta}) - \frac{1}{p}\operatorname{tr}(\mathbf{I} - \mathbf{B})\right\}. \tag{1·7}$$

Slightly different definitions of relative savings loss are used by Efron & Morris (1972). In words the relative savings loss is the additional risk incurred for using $\boldsymbol{\delta}$ when $\boldsymbol{\delta}^*$ is optimal, divided by the corresponding quantity when $\boldsymbol{\delta} = \boldsymbol{\delta}^0$. We show in § 2 that

$$RSL(\mathbf{A}, \boldsymbol{\delta}^1) = (p + 1)/k$$

for every value of $\mathbf{A}$. For example, if $p = 3$, $k = 20$ and $\mathbf{A}$ is such that $(p - 1)\operatorname{tr}(\mathbf{I} - \mathbf{B}) = 0·25$, then the maximum likelihood estimator has savings loss $1 - 0·25 = 0·75$ while $\boldsymbol{\delta}^1$ loses only $4/20$ as much, implying $R(\mathbf{A}, \boldsymbol{\delta}^1) = 0·40$ for this situation.

For convenience we have assumed the prior distribution (1·3) has mean vector zero, but in § 7 it is shown that there is no added difficulty in assuming

$$\boldsymbol{\theta}_i \sim \mathcal{N}_p(\mathbf{m}, \mathbf{A}) \quad (i = 1, \ldots, k), \tag{1·8}$$

$\mathbf{m}$ being an unknown vector. The modified version of $\boldsymbol{\delta}^1$ appropriate for this situation is an empirical Bayes rule in the sense that it does well against any prior of the form (1·8) without requiring knowledge of **m** or **A**, provided only that $(p+2)/k$ is reasonably small. This includes the null hypothesis situation $\mathbf{A} = \mathbf{0}$, i.e. the case where all the $\boldsymbol{\theta}_i$ are equal.

When $p = 1$, $\boldsymbol{\delta}^1$ reduces to the James–Stein estimator for k univariate parameters (James & Stein, 1960),

$$\delta_i^1(\mathbf{x}) = \left(1 - \frac{k-2}{\sum\limits_{i=1}^{k} x_i^2}\right) x_i \quad (i = 1, \ldots, k). \tag{1·9}$$

A. Baranchik in his unpublished thesis showed that this estimator can be improved upon for all values of $\boldsymbol{\theta}$ by restricting the factor multiplying x_i to be nonnegative, that is by using

$$\delta_i^{1+} = \left(1 - \frac{k-2}{\sum\limits_{i=1}^{k} x_i^2}\right)^{+} x_i; \tag{1·10}$$

see Stein (1966). In §6, we give the analogue of the plus rule for (1·6). In §8, we show the connexion between the plus-rule version of (1·6) and models that have been suggested for estimating interactions in a two-way analysis of variance.

The practical utility of Stein's estimator has been questioned ever since Stein's original presentation. In the review of this paper the Editor has raised this point again and asked us to comment on it. Section 9 is our defence, and consists of a practical example where Stein's approach in general, not just the variation discussed in this paper, yields a considerable improvement over the maximum likelihood estimator.

2. Relative savings loss

The term $(k-p-1)\,\mathbf{S}^{-1}$ in (1·6) can be seen to be an unbiased estimate of $\mathbf{B} = (\mathbf{A}+\mathbf{I})^{-1}$. As a matter of fact, since the $\mathbf{x}_i$ have the marginal distribution

$$\mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{A}+\mathbf{I}) \quad (i = 1, \ldots, k), \tag{2·1}$$

the marginal distribution of $\mathbf{S} = \mathbf{XX}'$ is Wishart,

$$\mathbf{S} \sim W(\mathbf{A}+\mathbf{I}, k, p), \tag{2·2}$$

and a standard calculation shows that

$$E_{\mathbf{A}}\{(k-p-1)\,\mathbf{S}^{-1}\} = \mathbf{B}. \tag{2·3}$$

Notice that in this case $E_{\mathbf{A}}$, as defined previously, reduces to an expectation over the marginal distribution (2·2) of **S**.

Suppose now we consider an arbitrary rule of the form

$$\hat{\boldsymbol{\delta}}_i = (\mathbf{I} - \hat{\mathbf{B}})\,\mathbf{x}_i \quad (i = 1, \ldots, k), \tag{2·4}$$

where $\hat{\mathbf{B}}$ is any symmetric matrix function of the data.

Lemma 1. *For any rule of the form* (2·4), *the relative savings loss is given by*

$$RSL(\mathbf{A}, \hat{\boldsymbol{\delta}}) = \frac{E_{\mathbf{A}}\{\operatorname{tr}(\hat{\mathbf{B}} - \mathbf{B})\,\mathbf{S}(\hat{\mathbf{B}} - \mathbf{B})\}}{k \operatorname{tr}(\mathbf{B})}. \tag{2·5}$$

Proof. We have

$$\begin{aligned} R(\mathbf{A}, \hat{\boldsymbol{\delta}}) &= E_{\mathbf{A}}\left[\frac{1}{pk}\operatorname{tr}\{\boldsymbol{\theta} - (\mathbf{I} - \hat{\mathbf{B}})\,\mathbf{X}\}'\{\boldsymbol{\theta} - (\mathbf{I} - \hat{\mathbf{B}})\,\mathbf{X}\}\right] \\ &= \frac{1}{pk}\operatorname{tr}\{E_{\mathbf{A}}(E_{\mathbf{x}}[\{\boldsymbol{\theta} - (\mathbf{I} - \hat{\mathbf{B}})\,\mathbf{X}\}'\{\boldsymbol{\theta} - (\mathbf{I} - \hat{\mathbf{B}})\,\mathbf{X}\}])\}. \end{aligned} \tag{2·6}$$

Using (1·4), we see that the $E_{\mathbf{x}}$ expectation in (2·6) is evaluated to be

$$k(\mathbf{I}-\mathbf{B})+(\hat{\mathbf{B}}-\mathbf{B})\mathbf{X}\mathbf{X}'(\hat{\mathbf{B}}-\mathbf{B}),$$

giving

$$R(\mathbf{A},\hat{\boldsymbol{\delta}}) = \frac{1}{p}\operatorname{tr}(\mathbf{I}-\mathbf{B})+\frac{1}{pk}E_{\mathbf{A}}\{\operatorname{tr}(\hat{\mathbf{B}}-\mathbf{B})\,\mathbf{S}(\hat{\mathbf{B}}-\mathbf{B})\}.$$

The lemma follows from the definition (1·7) of the relative savings loss.

THEOREM 1. *The rule* $\boldsymbol{\delta}^1_i = \{\mathbf{I}-(k-p-1)\,\mathbf{S}^{-1}\}\mathbf{x}_i$ $(i = 1,\ldots,k)$ *has* $RSL(\mathbf{A},\boldsymbol{\delta}^1) = (p+1)/k$ *for every value of* $\mathbf{A}$.

Theorem 1 follows directly from Lemma 1 by substituting $\hat{\mathbf{B}} = (k-p-1)\,\mathbf{S}^{-1}$ and using equation (2·3).

Lemma 1 reduces the problem of estimating $\boldsymbol{\theta}$ given $\mathbf{X}$ in the Bayesian situation (1·3), but with $\mathbf{A}$ unknown, to the problem of estimating $\mathbf{B}$ with a loss function proportional to

$$L(\mathbf{B},\hat{\mathbf{B}}) = \operatorname{tr}(\hat{\mathbf{B}}-\mathbf{B})\,\mathbf{S}(\hat{\mathbf{B}}-\mathbf{B}).$$

The matrix $\mathbf{S}$ is sufficient and complete for estimating $\mathbf{B}$, but there is no reason to believe that the unbiased estimator $(k-p-1)\,\mathbf{S}^{-1}$ is particularly good for this purpose, although it is easy to show that it is optimal among all estimators of the form $c\mathbf{S}^{-1}$. In § 6 we show that the plus-rule modification yields a uniformly better estimate of $\mathbf{B}$. It is tempting to consider using a Bayesian estimator for $\mathbf{B}$ itself, presumably one that would give added weight to special cases such as $\mathbf{A}$ diagonal or proportional to $\mathbf{I}$, but no calculations have been carried out in this direction. However, Professor C. Stein has pointed out to the authors that estimators of the form $\hat{\mathbf{B}} = a\mathbf{S}^{-1}+b\mathbf{I}/\operatorname{tr}(\mathbf{S})$ uniformly dominate $\hat{\mathbf{B}} = (k-p-1)\,\mathbf{S}^{-1}$ for certain choices of the constants a and b, a fact which will be discussed in a forthcoming paper.

3. THE RISK FUNCTION $R(\boldsymbol{\theta},\boldsymbol{\delta}^1)$

We now show that the rule $\boldsymbol{\delta}^1$ given by (1·6) dominates the maximum likelihood estimator $\boldsymbol{\delta}^0$.

THEOREM 2. *The rule* $\boldsymbol{\delta}^1$ *has risk function*

$$R(\boldsymbol{\theta},\boldsymbol{\delta}^1) = 1-\frac{(k-p-1)^2}{pk}E_{\boldsymbol{\theta}}\{\operatorname{tr}(\mathbf{X}\mathbf{X}')^{-1}\} \tag{3·1}$$

which is less than $R(\boldsymbol{\theta},\boldsymbol{\delta}^0) = 1$ *for all values of* $\boldsymbol{\theta}$.

Proof. Define $\boldsymbol{\sigma} = \boldsymbol{\theta}\boldsymbol{\theta}'$, so that under the Bayesian assumption (1·3)

$$\boldsymbol{\sigma} \sim W(\mathbf{A},k,p). \tag{3·2}$$

Both sides of equation (3·1) are easily verified to be invariant under transformations $\boldsymbol{\theta}\to\boldsymbol{\theta}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is any $k\times k$ orthogonal matrix. Therefore, both are functions of $\boldsymbol{\sigma}$, say $R(\boldsymbol{\theta},\boldsymbol{\delta}^1) = f(\boldsymbol{\sigma})$ and $1-\{(k-p-1)^2/pk\}E_{\boldsymbol{\theta}}\{\operatorname{tr}(\mathbf{X}\mathbf{X}')^{-1}\} = g(\boldsymbol{\sigma})$.

By definition $E_{\mathbf{A}}\{f(\boldsymbol{\sigma})\} = R(\mathbf{A},\boldsymbol{\delta}^1)$. Using Theorem 1 and the definition of the relative savings loss (1·7), we have

$$E_{\mathbf{A}}\{f(\boldsymbol{\sigma})\} = 1-\frac{k-p-1}{pk}\operatorname{tr}(\mathbf{A}+\mathbf{I})^{-1}. \tag{3·3}$$

We also calculate

$$E_{\mathbf{A}}\{g(\boldsymbol{\sigma})\} = 1 - \frac{k-p-1}{pk}\operatorname{tr}(\mathbf{A}+\mathbf{I})^{-1}, \tag{3·4}$$

this relation following from (2·3) and the fact that $E_{\mathbf{A}} E_{\boldsymbol{\theta}}\{\operatorname{tr}(\mathbf{XX}')^{-1}\} = E_{\mathbf{A}}\{\operatorname{tr}(\mathbf{S}^{-1})\}$. Because $\boldsymbol{\sigma}$ is complete for $\mathbf{A}$, the theorem follows by comparison of (3·3) and (3·4).

We have already noted that $R(\boldsymbol{\theta}, \boldsymbol{\delta}^1)$ is a function of $\boldsymbol{\theta}$ only through $\boldsymbol{\sigma}$, but actually it depends only on the eigenvalues of $\boldsymbol{\sigma}$. This can be seen by verifying directly from (1·6) and (1·2) that $R(\boldsymbol{\theta}, \boldsymbol{\delta}^1)$ is invariant under all transformations $\boldsymbol{\theta} \to \boldsymbol{\Gamma\theta\Lambda}$, where $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ are orthogonal of sizes $p \times p$ and $k \times k$, respectively.

4. Unknown covariance matrix

Suppose we relax (1·1) to

$$\mathbf{x}_i \sim \mathcal{N}_p(\boldsymbol{\theta}_i, \mathbf{D}) \quad (i = 1, \ldots, k), \tag{4·1}$$

where the $p \times p$ covariance matrix $\mathbf{D}$ is assumed to be of full rank. If $\mathbf{D}$ is known to the statistician then the previous theory goes through with only minor modifications in the definitions. The loss function (1·2) should now be written as

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{pk}\operatorname{tr}(\boldsymbol{\theta}-\boldsymbol{\delta})'\mathbf{D}^{-1}(\boldsymbol{\theta}-\boldsymbol{\delta}) \tag{4·2}$$

in order that the loss be invariant under linear transformations on the observed variables. Under the Bayesian assumption (1·3) the conditional distribution of $\boldsymbol{\theta}_i$, given $\mathbf{x}_i$, is

$$\boldsymbol{\theta}_i | \mathbf{x}_i \sim \mathcal{N}_p\{(\mathbf{I}-\mathbf{B})\,\mathbf{x}_i, (\mathbf{I}-\mathbf{B})\,\mathbf{D}\} \quad (i = 1, \ldots, k), \tag{4·3}$$

where now

$$\mathbf{B} = \mathbf{D}(\mathbf{A}+\mathbf{D})^{-1}. \tag{4·4}$$

Therefore the Bayes estimator $\boldsymbol{\delta}^* = (\mathbf{I}-\mathbf{B})\,\mathbf{X}$ has Bayes risk which is still given by (1·5), namely

$$R(\mathbf{A}, \boldsymbol{\delta}^*) = \frac{1}{p}\operatorname{tr}(\mathbf{I}-\mathbf{B}). \tag{4·5}$$

The definition of $\boldsymbol{\delta}^1$ is now

$$\boldsymbol{\delta}_i^1 = \{\mathbf{I} - (k-p-1)\,\mathbf{DS}^{-1}\}\,\mathbf{x}_i \quad (i = 1, \ldots, k) \tag{4·6}$$

with $\mathbf{S} = \mathbf{XX}'$ as before. With relative savings loss still defined by (1·7) as

$$RSL(\mathbf{A}, \boldsymbol{\delta}) = \{R(\mathbf{A}, \boldsymbol{\delta}) - R(\mathbf{A}, \boldsymbol{\delta}^*)\}/\{R(\mathbf{A}, \boldsymbol{\delta}^0) - R(\mathbf{A}, \boldsymbol{\delta}^*)\}, \tag{4·7}$$

the rule $\boldsymbol{\delta}^1$ continues to have $RSL(\mathbf{A}, \boldsymbol{\delta}^1) = (p+1)/k$ for all $\mathbf{A}$. This is derived from the generalized form of Lemma 1, i.e.

$$RSL(\mathbf{A}, \boldsymbol{\delta}) = \frac{E_{\mathbf{A}}\{\operatorname{tr}(\hat{\mathbf{B}}-\mathbf{B})\,\mathbf{S}(\hat{\mathbf{B}}-\mathbf{B})'\,\mathbf{D}^{-1}\}}{k\operatorname{tr}(\mathbf{B})}. \tag{4·8}$$

Notice that $\mathbf{B}$ is no longer necessarily symmetric. The risk function of $\boldsymbol{\delta}^1$ is

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^1) = 1 - \frac{(k-p-1)^2}{pk} E_{\boldsymbol{\theta}}\{\operatorname{tr}(\mathbf{XX}')^{-1}\mathbf{D}\}. \tag{4·9}$$

These assertions may be derived from their counterparts in the previous sections by a linear transformation on $\mathbf{X}$, or as special cases of the remainder of this section.

In many situations $\mathbf{D}$ will not be known to the statistician, but he will have data to estimate it. Suppose $\hat{\mathbf{D}}$, a symmetric matrix independent of $\mathbf{X}$, is available as an estimate of $\mathbf{D}$. This is the situation in most multivariate normal models. By analogy to (4·6), consider the estimator

$$\hat{\delta}_i^1 = \{\mathbf{I}-(k-p-1)\hat{\mathbf{D}}\mathbf{S}^{-1}\}\mathbf{x}_i \quad (i = 1, \ldots, k). \tag{4·10}$$

THEOREM 3. *The rule* (4·10) *has relative savings loss* (4·7) *given by*

$$RSL(\mathbf{A}, \hat{\delta}^1) = \frac{p+1}{k} + \left(1-\frac{p+1}{k}\right)\frac{1}{\operatorname{tr}(\mathbf{B})} E\{\operatorname{tr}(\hat{\mathbf{D}}\mathbf{D}^{-1}-\mathbf{I})\,\mathbf{B}(\hat{\mathbf{D}}\mathbf{D}^{-1}-\mathbf{I})\}. \tag{4·11}$$

Proof. Let $q \equiv k-p-1$. Then $\hat{\mathbf{B}} = q\hat{\mathbf{D}}\mathbf{S}^{-1}$, where $\mathbf{S} = \mathbf{XX}'$. Note that $\mathbf{S} \sim W\{(\mathbf{D}+\mathbf{A}), k, p\}$. Therefore, $E_{\mathbf{A}}(\mathbf{S}) = k(\mathbf{A}+\mathbf{D}) = k\mathbf{B}^{-1}\mathbf{D}$ and $E_{\mathbf{A}}(q\mathbf{S}^{-1}) = (\mathbf{A}+\mathbf{D})^{-1} = \mathbf{D}^{-1}\mathbf{B}$ by (2·3). The symbol E will indicate expectation over $\hat{\mathbf{D}}$ alone. Then (4·8) gives

$$\begin{aligned} k\operatorname{tr}(\mathbf{B})\,RSL(\mathbf{A}, \hat{\delta}^1) &= E_{\mathbf{A}}\{\operatorname{tr}(q\hat{\mathbf{D}}\mathbf{S}^{-1}-\mathbf{B})\,\mathbf{S}(q\mathbf{S}^{-1}\hat{\mathbf{D}}-\mathbf{B}')\,\mathbf{D}^{-1}\} \\ &= qE_{\mathbf{A}}[\operatorname{tr}\{\hat{\mathbf{D}}(q\mathbf{S}^{-1})\,\hat{\mathbf{D}}\mathbf{D}^{-1}\}]-2qE\{\operatorname{tr}(\mathbf{B}\hat{\mathbf{D}}\mathbf{D}^{-1})\}+E_{\mathbf{A}}\{\operatorname{tr}(\mathbf{BSBD}^{-1})\} \\ &= qE\{\operatorname{tr}(\hat{\mathbf{D}}\mathbf{D}^{-1}\mathbf{B}\hat{\mathbf{D}}\mathbf{D}^{-1})\}-2qE\{\operatorname{tr}(\mathbf{B}\hat{\mathbf{D}}\mathbf{D}^{-1})\}+k\operatorname{tr}(\mathbf{DBD}^{-1}) \\ &= qE\{\operatorname{tr}(\hat{\mathbf{D}}\mathbf{D}^{-1}-\mathbf{I})\,\mathbf{B}(\hat{\mathbf{D}}\mathbf{D}^{-1}-\mathbf{I})\}+(k-q)\operatorname{tr}(\mathbf{B}), \end{aligned}$$

which is (4·11).

In many situations the estimator $\hat{\mathbf{D}}$ will satisfy $E(\hat{\mathbf{D}}) = a_1\mathbf{D}$ and $E(\hat{\mathbf{D}}\mathbf{D}^{-1}\hat{\mathbf{D}}) = a_2\mathbf{D}$. We can assume that $a_1 = a_2 = a$, say, since if not, we can redefine $\hat{\mathbf{D}}$ to be $(a_1/a_2)\,\hat{\mathbf{D}}$,

$$E(\hat{\mathbf{D}}) = a\mathbf{D}, \quad E(\hat{\mathbf{D}}\mathbf{D}^{-1}\hat{\mathbf{D}}) = a\mathbf{D}. \tag{4·12}$$

The constant a satisfies $0 \leqslant a \leqslant 1$ since if we let $\mathbf{M} = \mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{D}}\mathbf{D}^{-\frac{1}{2}}$, (4·12) yields $E(\mathbf{M}^2) = a\mathbf{I}$, $E(\mathbf{M}-E\mathbf{M})^2 = (a-a^2)\,\mathbf{I}$.

COROLLARY 3.1. *Under assumptions* (4·12), *the rule* (4·10) *has relative savings loss*

$$RSL(\mathbf{A}, \hat{\delta}^1) = \frac{p+1}{k} + \left(1-\frac{p+1}{k}\right)(1-a) = 1-\frac{k-p-1}{k}a \tag{4·13}$$

and risk function

$$R(\theta, \hat{\delta}^1) = 1-a\frac{(k-p-1)^2}{pk}E_\theta[\operatorname{tr}\{(\mathbf{XX}')^{-1}\mathbf{D}\}]. \tag{4·14}$$

The proof of (4·13) is immediate from (4·12) and (4·11). Then (4·14) follows by the completeness argument of § 3. It is also easy to see that replacing $\hat{\mathbf{D}}$ by $c\hat{\mathbf{D}}$ will uniformly increase both the relative savings loss and the risk function for any choice of $c \neq 1$.

Familiar assumptions about $\mathbf{D}$ with a complete sufficient statistic $\hat{\mathbf{D}}$ independent of $\mathbf{X}$ and satisfying (4·12) include

(i) $\mathbf{D}$ known, $\hat{\mathbf{D}} = \mathbf{D}$, $a = 1$, $RSL(\mathbf{A}, \hat{\delta}^1) = (p+1)/k$; (4·15)

(ii) $\mathbf{D} = \sigma\mathbf{G}$, $\mathbf{G}$ known, σ unknown, $v \sim \sigma\chi_n^2$,

$$\hat{\sigma} = v/(n+2), \quad \hat{\mathbf{D}} = \hat{\sigma}\mathbf{I}, \quad a = n/(n+2),$$

$$RSL(\mathbf{A}, \hat{\delta}^1) = \frac{p+1}{k} + \left(1-\frac{p+1}{k}\right)\frac{2}{n}; \tag{4·16}$$

(iii) $\mathbf{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$, all σ_i unknown, $v_i \sim \sigma_i\chi_n^2$ independent $(i = 1, \ldots, p)$;

$$\hat{\sigma}_i = v_i/(n+2), \quad \hat{\mathbf{D}} = \operatorname{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_p), \quad a = \frac{n}{n+2},$$

$$RSL(\mathbf{A}, \hat{\delta}^1) = \frac{p+1}{k} + \left(1-\frac{p+1}{k}\right)\frac{2}{n}; \tag{4.17}$$

(iv) $\mathbf{D} > 0$, entirely unknown, $\mathbf{V} \sim W(\mathbf{D}, n, p)$, $n \geqslant p$,

$$\hat{\mathbf{D}} = \mathbf{V}/(n+p+1), \quad a = n/(n+p+1),$$

$$RSL(\mathbf{A}, \hat{\boldsymbol{\delta}}^1) = \frac{p+1}{k} + \left(1 - \frac{p+1}{k}\right) \frac{p+1}{n+p+1} = \frac{p+1}{k} \frac{n+k}{n+p+1}.$$

5. Univariate James–Stein rule applied co-ordinate-wise

Assume once again that $\mathbf{x}_i \sim \mathcal{N}_p(\boldsymbol{\theta}_i, \mathbf{I})$, i.e. that $\mathbf{D} = \mathbf{I}$. We could apply the univariate James–Stein rule (1·9) separately to each co-ordinate of the $\mathbf{x}$ vectors, that is estimate θ_{ij} by

$$\tilde{\delta}_{ij} = \left(1 - \frac{k-2}{\sum\limits_{i=1}^{k} x_{ij}^2}\right) x_{ij} \quad (i = 1, \ldots, k; j = 1, \ldots, p). \tag{5·1}$$

Application of Lemma 1 gives the relative savings loss of this rul e as

$$RSL(\mathbf{A}, \tilde{\boldsymbol{\delta}}) = \frac{2}{k} + \left(1 - \frac{2}{k}\right) \frac{\operatorname{tr}(\mathbf{B}) - \sum\limits_{j=1}^{p} B_j}{\operatorname{tr}(\mathbf{B})},$$

where $\mathbf{B} = (\mathbf{A}+\mathbf{I})^{-1}$ as before, and $B_j \equiv 1/(A_{jj}+1)$, A_{jj} being the jth diagonal element of $\mathbf{A}$.

Let ρ_j be the multiple correlation between x_{ij} and $(x_{i1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{ip})$ under the marginal distribution $\mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{A}+\mathbf{I})$. Then standard manipulations (Anderson, 1958, p. 32) show that $1 - \rho_j^2 = B_j/B_{jj}$ and hence

$$\operatorname{tr}(\mathbf{B}) - \sum_{j=1}^{p} B_i = \sum_{j=1}^{p} B_{jj} \rho_j^2,$$

which gives the following comparison.

Theorem 4. *The rule* $\boldsymbol{\delta}^1$ *given by* (1·6) *is preferable to the rule* $\tilde{\boldsymbol{\delta}}$ *given by* (5·1) *if and only if*

$$\frac{p-1}{k-2} \leqslant \frac{\sum\limits_{j=1}^{p} B_{jj}\rho_j^2}{\sum\limits_{j=1}^{p} B_{jj}}.$$

Theorem 4 is a quantitative statement of the fact that the rule $\boldsymbol{\delta}^1$ picks up information between co-ordinates as well as between the k repetitions of the problem. If there is little or no such information, the limiting case being $\mathbf{A}$ a diagonal matrix, the use of $\tilde{\boldsymbol{\delta}}$ gives a better value of the relative savings loss.

6. An improved version of the rule $\boldsymbol{\delta}^1$

In this section we discuss the multivariate equivalent of the plus-rule (1·10). At first we continue to assume that $\mathbf{x}_i \sim \mathcal{N}_p(\boldsymbol{\theta}_i, \mathbf{I})$ $(i = 1, \ldots, k)$ as in § 2.

Definition. Let $\mathscr{B}$ be the set of all $p \times p$ symmetric matrices $\mathbf{B}$ which are positive definite and such that $\mathbf{I} - \mathbf{B}$ is nonnegative definite, i.e.

$$\mathscr{B} = \{\mathbf{B}: \mathbf{0} < \mathbf{B} \leqslant \mathbf{I}\}. \tag{6·1}$$

Every matrix $\mathbf{B} = (\mathbf{A}+\mathbf{I})^{-1}$ is in $\mathscr{B}$, and conversly, whereas the estimator

$$\hat{\mathbf{B}} = (k-p-1)\,\mathbf{S}^{-1}$$

need not lie in $\mathscr{B}$. It turns out that we can improve any rule $\hat{\boldsymbol{\delta}}_i = (\mathbf{I}-\hat{\mathbf{B}})\,\mathbf{x}_i$ by forcing the estimator $\hat{\mathbf{B}}$ to always lie in $\mathscr{B}$.

The set $\mathscr{B}$ is compact and convex in $E^{\frac{1}{2}p(p+1)}$, the Euclidean space of dimension $\frac{1}{2}p(p+1)$ which represents every $p \times p$ symmetric matrix by its entries on and above the main diagonal. For any given positive definite matrix $\mathbf{S}$, we can define an inner product in this space by $(\mathbf{B}, \tilde{\mathbf{B}})_{\mathbf{S}} = \operatorname{tr}(\mathbf{B}\mathbf{S}\tilde{\mathbf{B}})$, which in turn defines a distance function $\Delta_{\mathbf{S}}$ between matrices,

$$\Delta^2_{\mathbf{S}}(\mathbf{B}, \tilde{\mathbf{B}}) = \operatorname{tr}\{(\mathbf{B}-\tilde{\mathbf{B}})\,\mathbf{S}(\mathbf{B}-\tilde{\mathbf{B}})\}. \tag{6·2}$$

DEFINITION. For $\mathbf{S}$ positive definite and $\hat{\mathbf{B}}$ any symmetric matrix, let $\hat{\mathbf{B}}_{\mathbf{S}}$ be the point in $\mathscr{B}$ nearest $\hat{\mathbf{B}}$ in distance $\Delta_{\mathbf{S}}$, that is

$$\Delta_{\mathbf{S}}(\hat{\mathbf{B}}_{\mathbf{S}}, \hat{\mathbf{B}}) = \inf_{\mathbf{B}\in\mathscr{B}} \Delta_{\mathbf{S}}(\mathbf{B}, \hat{\mathbf{B}}). \tag{6·3}$$

THEOREM 5. *If $\hat{\boldsymbol{\delta}}$ is any rule of the form $\hat{\boldsymbol{\delta}}_i = (\mathbf{I}-\hat{\mathbf{B}})\,\mathbf{x}_i$ $(i = 1, \ldots, k)$, then the rule $\hat{\boldsymbol{\delta}}_{\mathbf{S}}$ defined by $\hat{\boldsymbol{\delta}}_{\mathbf{S}i} = (\mathbf{I}-\hat{\mathbf{B}}_{\mathbf{S}})\,\mathbf{x}_i$ $(i = 1, \ldots, k)$ has $RSL(\mathbf{A}, \hat{\boldsymbol{\delta}}_{\mathbf{S}}) \leqslant RSL(\mathbf{A}, \hat{\boldsymbol{\delta}})$ for every value of* $\mathbf{A}$.

Proof. Theorem 5 follows from Lemma 1 by noticing that

$$\operatorname{tr}\{(\hat{\mathbf{B}}_{\mathbf{S}}-\mathbf{B})\,\mathbf{S}(\hat{\mathbf{B}}_{\mathbf{S}}-\mathbf{B})\} \leqslant \operatorname{tr}\{(\hat{\mathbf{B}}-\mathbf{B})\,\mathbf{S}(\hat{\mathbf{B}}-\mathbf{B})\} \tag{6·4}$$

for every realization of $\mathbf{X}$: if $\hat{\mathbf{B}} \in \mathscr{B}$, then $\hat{\mathbf{B}}_{\mathbf{S}} = \hat{\mathbf{B}}$ and (6·4) is trivially true. For $\hat{\mathbf{B}} \notin \mathscr{B}$, (6·4) reflects the following easily verified fact. If $\mathscr{B}$ is a convex set in Euclidean space, $\hat{\mathbf{B}}$ is a point not in $\mathscr{B}$ and $\hat{\mathbf{B}}_{\mathbf{S}}$ is the nearest point to $\hat{\mathbf{B}}$ in $\mathscr{B}$ in some Euclidean metric $\Delta_{\mathbf{S}}$, then

$$\Delta_{\mathbf{S}}(\hat{\mathbf{B}}_{\mathbf{S}}, \mathbf{B}) < \Delta_{\mathbf{S}}(\hat{\mathbf{B}}, \mathbf{B})$$

for every $\mathbf{B} \in \mathscr{B}$. More precisely,

$$\Delta^2_{\mathbf{S}}(\hat{\mathbf{B}}, \mathbf{B}) - \Delta^2_{\mathbf{S}}(\hat{\mathbf{B}}_{\mathbf{S}}, \mathbf{B}) \geqslant \Delta^2_{\mathbf{S}}(\hat{\mathbf{B}}, \hat{\mathbf{B}}_{\mathbf{S}}). \tag{6·5}$$

For the particular estimator $\hat{\mathbf{B}} = (k-p-1)\,\mathbf{S}^{-1}$ we have been using, it is easy to write down an expression for $\hat{\mathbf{B}}_{\mathbf{S}}$ in terms of the eigenvalues and eigenvectors of $\mathbf{S}$, say $\mathbf{S} = \boldsymbol{\Gamma}\mathbf{E}\boldsymbol{\Gamma}'$, where $\boldsymbol{\Gamma}$ is a $p \times p$ orthogonal matrix and $\mathbf{E}$ the diagonal matrix with e_{ii} equal to the ith largest eigenvalue. Then

$$\hat{\mathbf{B}} = (k-p-1)\,\boldsymbol{\Gamma}\mathbf{E}^{-1}\boldsymbol{\Gamma}', \tag{6·6}$$

$$\hat{\mathbf{B}}_{\mathbf{S}} = (k-p-1)\,\boldsymbol{\Gamma}\mathbf{E}_{\mathbf{S}}^{-1}\boldsymbol{\Gamma}', \tag{6·7}$$

where $\mathbf{E}_{\mathbf{S}}$ has diagonal elements $\max\{(k-p-1), e_{ii}\}$. Notice that $\hat{\mathbf{B}}_{\mathbf{S}}$ is just the expression (6·6) for $\hat{\mathbf{B}}$ modified so that all the roots of the estimated $\mathbf{B}$ matrix are between 0 and 1. To prove that $\hat{\mathbf{B}}_{\mathbf{S}}$ of (6·7) satisfies (6·3) write an arbitrary matrix $\mathbf{B}$ in $\mathscr{B}$ as

$$\mathbf{B} = (k-p-1)\,\boldsymbol{\Gamma}\mathbf{Q}^{-1}\boldsymbol{\Gamma}',$$

where $\mathbf{Q}^{-1}$ is a positive definite matrix with maximum root less than or equal to $1/(k-p-1)$. If $\mathbf{Q}^{-1}$ is not diagonal, the diagonal matrix $\mathbf{Q}_0^{-1}$ with the elements of $\mathbf{Q}^{-1}$ on the main diagonal also is positive definite with maximum root less than or equal to $1/(k-p-1)$ and direct computation shows that $\mathbf{B}_0 = (k-p-1)\,\boldsymbol{\Gamma}\mathbf{Q}_0^{-1}\boldsymbol{\Gamma}'$ satisfies $\Delta^2_{\mathbf{S}}(\mathbf{B}_0, \hat{\mathbf{B}}) < \Delta^2_{\mathbf{S}}(\mathbf{B}, \hat{\mathbf{B}})$. Therefore $\mathbf{Q}^{-1}$ can be taken to be diagonal, giving

$$\Delta^2_{\mathbf{S}}(\mathbf{B}, \hat{\mathbf{B}}) = (k-p-1)^2 \sum_{i=1}^{p} e_{ii}\left(\frac{1}{e_{ii}} - \frac{1}{q_{ii}}\right)^2.$$

The restriction that the maximum root of $\mathbf{Q}^{-1}$ be less than or equal to $1/(k-p-1)$ is equivalent to $\min\limits_{1 \leqslant i \leqslant p} q_{ii} \geqslant k-p-1$, and (6·7) is verified. From (6·5) we see that

$$\Delta^2_{\mathbf{S}}(\hat{\mathbf{B}}, \mathbf{B}) - \Delta^2_{\mathbf{S}}(\hat{\mathbf{B}}_{\mathbf{S}}, \mathbf{B}) \geqslant (k-p-1)^2 \Sigma \frac{1}{e_{ii}} \left(1 - \frac{e_{ii}}{k-p-1}\right)^2,$$

the summation being over all values of i with $e_{ii} < k-p-1$.

In the case $p = 1$, it is possible to calculate the increase in savings of the plus-rule $\boldsymbol{\delta}^{1+}$ over $\boldsymbol{\delta}^1$, i.e.

$$RSL(A, \boldsymbol{\delta}^1) - RSL(A, \boldsymbol{\delta}^{1+}) = \left(\frac{2}{k} - A^2\right) I_{\frac{1}{2}k}\{\tfrac{1}{2}(k-2)\,B\} + \frac{k-2}{k}\, A i_{\frac{1}{2}k}\{\tfrac{1}{2}(k-2)\,B\}.$$

Here $I_{\frac{1}{2}k}$ is the incomplete gamma function and $i_{\frac{1}{2}k}$ its derivative, $i_{\frac{1}{2}k}(t) = t^{\frac{1}{2}k-1}\, e^{-t}/\Gamma(\frac{1}{2}k)$. The maximum improvement occurs at $A = 0$. In the case $k = 4$, for example,

$$RSL(0, \delta^{1+}) = 0{\cdot}368$$

as compared to $RSL(0, \delta^1) = 0{\cdot}500$.

It seems likely that the improvement of $\boldsymbol{\delta}^{1+}$ over $\boldsymbol{\delta}^1$ is more substantial when $p > 1$ since the probability $\hat{\mathbf{B}} \notin \mathscr{B}$ can be much higher, but so far the authors have not carried out calculations of the actual savings.

Suppose more generally that $\mathbf{x}$ is distributed as (4·1) and the loss function (4·2) is used. If $\mathbf{D}$ is known, then it is simple to transform back to the case $\mathbf{D} = \mathbf{I}$. The transformations $\tilde{\mathbf{x}}_i = \mathbf{D}^{-\frac{1}{2}}\mathbf{x}_i$ and $\tilde{\boldsymbol{\theta}}_i = \mathbf{D}^{-\frac{1}{2}}\boldsymbol{\theta}_i$ $(i = 1, \ldots, k)$ induce the relationships

$$\tilde{\mathbf{S}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}, \quad \tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}, \quad \tilde{\mathbf{B}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{B}\mathbf{D}^{\frac{1}{2}}.$$

The estimate $\hat{\mathbf{B}} = (k-p-1)\,\mathbf{D}\mathbf{S}^{-1}$ of $\mathbf{B}$ used in $\boldsymbol{\delta}^1$ transforms into

$$\hat{\tilde{\mathbf{B}}} = \mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{B}}\mathbf{D}^{\frac{1}{2}} = (k-p-1)\,\tilde{\mathbf{S}}^{-1},$$

and if $\hat{\tilde{\mathbf{B}}} \notin \mathscr{B}$, we can uniformly reduce $\operatorname{tr}\{(\hat{\tilde{\mathbf{B}}} - \tilde{\mathbf{B}})\,\tilde{\mathbf{S}}(\hat{\tilde{\mathbf{B}}} - \tilde{\mathbf{B}})\}$ by replacing $\hat{\tilde{\mathbf{B}}}$ with $\hat{\tilde{\mathbf{B}}}_{\tilde{\mathbf{S}}}$ as in (6·7). The resulting rule $\boldsymbol{\delta}^1_{\tilde{\mathbf{S}}}$ has $RSL(\mathbf{A}, \boldsymbol{\delta}^1_{\tilde{\mathbf{S}}}) < RSL(\mathbf{A}, \boldsymbol{\delta}^1)$ for every $\mathbf{A} > 0$.

The matrix $\hat{\tilde{\mathbf{B}}}_{\tilde{\mathbf{S}}}$ is the nearest point in $\mathscr{B}$ to the matrix $\hat{\tilde{\mathbf{B}}}$ in the metric $\Delta_{\tilde{\mathbf{S}}}$. If $\mathbf{D}$ is unknown but we have some estimate $\hat{\mathbf{D}}$ of it as in § 4, then we can proceed by defining $\hat{\mathbf{B}} = (k-p-1)\,\hat{\mathbf{D}}\mathbf{S}^{-1}$ or equivalently $\hat{\tilde{\mathbf{B}}} = (k-p-1)\,\tilde{\mathbf{S}}^{-1}$ with $\tilde{\mathbf{S}} = \hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{S}\hat{\mathbf{D}}^{-\frac{1}{2}}$. If $\hat{\tilde{\mathbf{B}}} \notin \mathscr{B}$, we can still obtain $\hat{\tilde{\mathbf{B}}}_{\tilde{\mathbf{S}}}$ as before. However, there is now no guarantee that the relative savings loss will be uniformly reduced by using $\hat{\tilde{\mathbf{B}}}_{\tilde{\mathbf{S}}}$ in place of $\hat{\tilde{\mathbf{B}}}$. The difficulty is that minimizing distance in the metric $\Delta_{\tilde{\mathbf{S}}} = \Delta_{\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{S}\hat{\mathbf{D}}^{-\frac{1}{2}}}$ is not in general the same as minimizing distance in the metric $\Delta_{\mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}}$. In the important special case (4·16), where $\mathbf{D}$ is known up to a constant, the two metrics are equivalent and we do get $RSL(\mathbf{A}, \hat{\boldsymbol{\delta}}_{\mathbf{S}}) < RSL(\mathbf{A}, \hat{\boldsymbol{\delta}})$ for all $\mathbf{A} > 0$.

7. Unknown prior mean

If instead of (1·3) we assume that *a priori* the $\boldsymbol{\theta}_i$ are independently distributed as

$$\boldsymbol{\theta}_i \sim \mathscr{N}_p(\mathbf{m}, \mathbf{A}) \quad (i = 1, \ldots, k), \tag{7·1}$$

where $\mathbf{A}$ is an unknown $p \times p$ covariance matrix and $\mathbf{m}$ is an unknown $p \times 1$ mean vector, then we can modify the definition of $\boldsymbol{\delta}^1$ as follows:

$$\boldsymbol{\delta}^1_i = \bar{\mathbf{x}} + \{\mathbf{I} - (k-p-2)\,\tilde{\mathbf{S}}^{-1}\}(\mathbf{x}_i - \bar{\mathbf{x}}) \quad (i = 1, \ldots, k). \tag{7·2}$$

In this expression,

$$\bar{\mathbf{x}} = \sum_{i=1}^{k} \mathbf{x}_i/k, \quad \tilde{\mathbf{S}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}',$$

where, letting $\mathbf{e}_k' = (1, \ldots, 1)$,

$$\tilde{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{x}}\mathbf{e}_k'. \tag{7·3}$$

Equation (7·2) is derived by writing

$$\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}\mathbf{e}_k' + \tilde{\boldsymbol{\theta}}, \tag{7·4}$$

where $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\mathbf{e}_k'$, and using $\boldsymbol{\delta}^1(\tilde{\mathbf{x}})$ as defined in (1·6) to estimate $\tilde{\boldsymbol{\theta}}$. The degrees of freedom are reduced by one since the rows of $\tilde{\mathbf{X}}$ and $\tilde{\boldsymbol{\theta}}$ lie in the $k-1$ dimensional space orthogonal to $\mathbf{e}_k$. We are now assuming that $p+2 < k$. The vector $\bar{\boldsymbol{\theta}}$ is estimated by maximum likelihood as $\bar{\mathbf{x}}$. Then the estimates for $\bar{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are combined as in (7·4) to give (7·2). It is a simple consequence of Theorem 1 that

$$RSL\{(\mathbf{m}, \mathbf{A}), \boldsymbol{\delta}^1\} = (p+2)/k \tag{7·5}$$

for all values of $\mathbf{m}$ and $\mathbf{A}$. We have modified the relative savings loss notation to include the parameter vector $\mathbf{m}$, but with only that change the definitions (1·7) remain valid. Of course we do not have to use maximum likelihood to estimate $\bar{\boldsymbol{\theta}}$. If $p \geqslant 3$, we could use the univariate James–Stein estimator to estimate $\bar{\boldsymbol{\theta}}$ from $\bar{\mathbf{x}}$ and thereby produce an estimator of $\boldsymbol{\theta}$ with uniformly lower risk than (7·2).

More generally we can write $\boldsymbol{\theta} = \boldsymbol{\phi}\mathbf{F} + \tilde{\boldsymbol{\theta}}$, where $\mathbf{F}$ is some given structure matrix of size $r \times k$ and rank r, $r < p$, and $\boldsymbol{\phi}$ is the $p \times r$ matrix $\boldsymbol{\phi} = \boldsymbol{\theta}\mathbf{F}'(\mathbf{F}\mathbf{F}')^{-1}$, so that the rows of $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\phi}\mathbf{F}$ are orthogonal to the rows of $\mathbf{F}$. Let $\hat{\boldsymbol{\phi}}$ be the usual maximum likelihood multivariate analysis of variance estimate of $\boldsymbol{\phi}$, $\hat{\boldsymbol{\phi}} = \mathbf{X}\mathbf{F}'(\mathbf{F}\mathbf{F}')^{-1}$ and $\tilde{\mathbf{X}} = \mathbf{X} - \hat{\boldsymbol{\phi}}\mathbf{F}$, $\tilde{\mathbf{S}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$. Assume $p+r+1 < k$. Then the version of $\boldsymbol{\delta}^1$ defined by

$$\boldsymbol{\delta}_i^1 = \hat{\boldsymbol{\phi}}\mathbf{F}_i + \{\mathbf{I} - (k-p-r-1)\,\tilde{\mathbf{S}}^{-1}\}\tilde{\mathbf{x}}_i \quad (i = 1, \ldots, k) \tag{7·6}$$

has $RSL\{(\mathbf{m}, \mathbf{A}), \boldsymbol{\delta}^1\} = (p+r+1)/k$ for any prior distribution $\boldsymbol{\theta}_i \sim \mathcal{N}_p(\mathbf{m}_i, \mathbf{A})$ $(i = 1, \ldots, k)$ provided that the matrix of mean vectors $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_k)$ can be written as $\mathbf{m} = \mathbf{f}\mathbf{F}$ for some $p \times r$ matrix $\mathbf{f}$.

What are the advantages of using (7·6), or (7·2), instead of (1·6)? If the structure matrix $\mathbf{F}$ is particularly appropriate for the problem at hand then the residual matrix $\tilde{\boldsymbol{\theta}}$ should have much smaller elements than $\boldsymbol{\theta}$. Stein's method reaps its greatest savings when the parameters have small magnitude; in the Bayesian framework when $\mathbf{A}$ is small. We increase the relative savings loss from $(p+2)/k$ to $(p+r+1)/k$, but this disadvantage may be overcome by a larger increase in $1/\{p\,\mathrm{tr}\,(\mathbf{B})\}$, which is the amount of possible savings the savings loss is computed from.

8. Application to two-way analysis of variance

There is a relationship between the estimator $\boldsymbol{\delta}^{1+}$ and methods that have been suggested for analyzing interactions in a two-way analysis of variance. We will point this out specifically with reference to the interesting model proposed by Mandel (1969). A similar model is proposed by Gollob (1968). For the purposes of the comparison it is sufficient to assume again that $\mathbf{x}_i \sim \mathcal{N}_p(\boldsymbol{\theta}_i, \mathbf{I})$ $(i = 1, \ldots, k)$ although the case $\mathbf{x}_i \sim \mathcal{N}_p(\boldsymbol{\theta}_i, \sigma^2\mathbf{I})$ with σ^2 unknown is more common.

We write the $\boldsymbol{\theta}$ matrix in the form that emphasizes the usual main effects,

$$\boldsymbol{\theta} = \mu\mathbf{e}_p\mathbf{e}_k' + \boldsymbol{\alpha}\mathbf{e}_k' + \mathbf{e}_p\boldsymbol{\beta}' + \tilde{\boldsymbol{\theta}}.$$

Here μ is the scalar grand mean, $\boldsymbol{\alpha}$ is the $p\times 1$ vector of row effects satisfying $\mathbf{e}_p'\boldsymbol{\alpha} = 0$ and $\boldsymbol{\beta}$ is the $k\times 1$ vector of column effects satisfying $\mathbf{e}_k'\boldsymbol{\beta} = 0$. As before $\mathbf{e}_p$ and $\mathbf{e}_k$ are vectors of ones of the indicated lengths. The residual matrix $\tilde{\boldsymbol{\theta}}$ satisfies $\tilde{\boldsymbol{\theta}}\mathbf{e}_k = 0$ and $\mathbf{e}_p'\tilde{\boldsymbol{\theta}} = 0$.

If we estimate μ, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by their usual maximum likelihood estimates $\hat{\mu}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, and subtract the fitted matrix $\hat{\mu}\mathbf{e}_p\mathbf{e}_k'+\hat{\boldsymbol{\alpha}}\mathbf{e}_k'+\mathbf{e}_p\hat{\boldsymbol{\beta}}'$ from $\mathbf{X}$, the residual matrix $\tilde{\mathbf{X}}$, with components

$$\tilde{x}_{ij} = x_{ij}-x_{i.}-x_{.j}+x_{..}$$

in the usual analysis of variance notation, has expected value $\tilde{\boldsymbol{\theta}}$. Because of the linear constraints $\tilde{\mathbf{X}}\mathbf{e}_k = \mathbf{0}$ and $\mathbf{e}_p'\tilde{\mathbf{X}} = \mathbf{0}$, $\tilde{\mathbf{X}}$ has $(p-1)\times(k-1)$ degrees of freedom, and with a proper change of co-ordinates can be rewritten as that many independent normal variates each with variance one. Stein (1966) has proposed using the univariate James–Stein estimator to estimate $\tilde{\boldsymbol{\theta}}$ in this situation.

The multivariate estimator $\boldsymbol{\delta}^{1+}$ can also be applied here. Without going into the details, which amount to finding a pseudo inverse for the singular matrix $\tilde{\mathbf{S}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}$ for use in (6·7) the estimate of the matrix $\tilde{\boldsymbol{\theta}}$ given by $\boldsymbol{\delta}^{1+}$ is

$$\boldsymbol{\delta}^{1+} = (\mathbf{I}-\boldsymbol{\Gamma}\mathbf{M}\boldsymbol{\Gamma}')\,\tilde{\mathbf{X}}, \tag{8·1}$$

where $\mathbf{M}$ is a $(p-1)\times(p-1)$ diagonal matrix having ith diagonal element

$$m_{ii} = \min\{1, (k-p-1)/e_{ii}\} \quad (i = 1,\ldots,p-1). \tag{8·2}$$

As in § 6, e_{ii} is the ith largest eigenvalue of $\tilde{\mathbf{S}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$, and the ith column of $\boldsymbol{\Gamma}$ is the corresponding eigenvector, so that $\boldsymbol{\Gamma}$ is a $p\times(p-1)$ matrix satisfying $\boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \mathbf{I}$. If $\tilde{\boldsymbol{\theta}} = \mathbf{0}$, that is if there is no real interaction in the table, then $\tilde{\mathbf{S}}$ will have a Wishart distribution which under a reparameterization to remove redundant co-ordinates can be written as $W(\mathbf{I}, k-1, p-1)$. In this case (8·1) will tend to estimate $\tilde{\boldsymbol{\theta}}$ as $\mathbf{0}$ or close to zero. From our previous calculations, we know that $R(\tilde{\boldsymbol{\theta}}, \boldsymbol{\delta}^{1+}) < p/(k-1)$ for $\tilde{\boldsymbol{\theta}} = \mathbf{0}$. This follows from Theorems 2 and 5 and the fact that $R(\mathbf{A}, \boldsymbol{\delta})$ for $\mathbf{A} = \mathbf{0}$ is the same as $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ with $\boldsymbol{\theta} = \mathbf{0}$ for any rule $\boldsymbol{\delta}$. We have $p/(k-1)$ instead of $(p+1)/k$ because of the reduced degrees of freedom in $\tilde{\mathbf{X}}$. Since both k and p have been reduced by one, the constant $k-p-1$ in (8·2) is the same as in (6·7).

Mandel proposes an estimate of $\tilde{\boldsymbol{\theta}}$ of exactly the form (8·1) but with $\mathbf{M}$ a diagonal matrix defined by $m_{ii} = 0$ or 1 $(i = 1,\ldots,p-1)$. Essentially those eigenvalues e_{ii} which are small enough to have occurred reasonably under the null distribution, $\tilde{\boldsymbol{\theta}} = \mathbf{0}$, of $\tilde{\mathbf{S}}$ correspond to choices of $m_{ii} = 1$, while the others are assigned $m_{ii} = 0$. Mandel does not give a formal procedure for making this decision, but he does provide tables of the expected values of the 3 largest roots under the null hypothesis.

From the point of view of this paper, Mandel's procedure amounts to using an unnatural estimate of $\mathbf{B}$, the estimated matrix $\hat{\mathbf{B}}$ having all its roots equal to 0 or 1. Equivalently, the estimated matrix $\hat{\mathbf{A}}$ defined by $\hat{\mathbf{B}} = (\hat{\mathbf{A}}+\mathbf{I})^{-1}$ has all its roots equal to infinity or 0. We can expect (8·1) to yield poor relative savings loss properties. A possible improvement would estimate those roots of $\mathbf{B}$ which pass the significance test and are hence judged to be less than 1 by a better estimator than 0, for example by $(k-p-1)/e_{ii}$. However, experience in simpler situations (Sclove, Morris & Radhakrishnan, 1972) has shown that rules of the form 'first test the null hypothesis and if you reject then use some standard estimation rule' are outperformed by Stein-type estimators.

9. APPLICATION OF EMPIRICAL BAYES ESTIMATORS FOR IMPROVING A COMPUTER SIMULATION

We give below an example of a Monte-Carlo experiment where Stein's method yields a substantial improvement over the usual estimator. The example is realistic in that the normality and variance assumptions of the earlier sections are only approximations to the true situation. It should be said that this is the first and only example considered for this paper, and that the favourable results are typical of our previous experience. We are currently preparing a paper which will present several applications of Stein's method to real data.

Let Y_1 and Y_2 be independent binomial random variables, $Y_i \sim \text{bin}\,(m, \pi_i)$, so $E(Y_i) = m\pi_i$. Pearson's chi-squared statistic is commonly used to test the composite null hypothesis H_0: $\pi_1 = \pi_2$ against all alternatives. If the nominal size is $\alpha = 0{\cdot}05$, then using the chi-squared approximation with one degree of freedom requires that we reject H_0 if

$$T = \frac{2m(Y_1-Y_2)^2}{(Y_1+Y_2)\,(2m-Y_1-Y_2)} > 3{\cdot}84. \tag{9·1}$$

We denote the true size of the test (9·1), which depends on m and $\pi = \pi_1 = \pi_2$, by $\alpha(m, \pi)$. We consider $m_i \equiv 5+5i$ and $\pi_j \equiv 0{\cdot}525 - 0{\cdot}025j$ $(i = 1, 2, 3; j = 1, \ldots, 17)$. Thus $p = 3$ and $k = 17$. For each of these $pk = 51$ cases, $n = 500$ computer simulations using pseudo random numbers were obtained, and the number of times that (9·1) held was recorded as N_{ij}. Hence $N_{ij} \sim \text{bin}\,\{n, \alpha(m_i, \pi_j)\}$ independently for all i and j. The maximum likelihood and unbiased estimate of $\alpha_{ij} \equiv \alpha(m_i, \pi_j)$ is $Z_{ij} = N_{ij}/n$. The standard deviation of Z_{ij} is $\{\alpha_{ij}(1-\alpha_{ij})/n\}^{\frac{1}{2}}$ which is approximately $\sigma \equiv \{(0{\cdot}05)\,(0{\cdot}95)/500\}^{\frac{1}{2}} \equiv 0{\cdot}009747$. Since Z_{ij} is approximately normal, the variables $x_{ij} \equiv (Z_{ij} - 0{\cdot}05)/\sigma$, $E(x_{ij}) = \theta_{ij} \equiv (\alpha_{ij} - 0{\cdot}05)/\sigma$, have approximately the distribution (1·1).

We estimated the θ_{ij} by seven different estimators. The first of these was the maximum likelihood estimator x_{ij}. The second was Stein's univariate plus-rule (1·10) with $k = 51$, that is applied simultaneously to all the data. The third was (1·10) with $k = 17$, applied separately to the cases $m = 10$, 15 and 20. The fourth was the multivariate rule (1·6) with the plus-rule modification (6·7).

The fifth, sixth and seventh cases consisted of applying the three Stein rules to the residuals of a linear regression fitted separately to each row of the **X** matrix. This is the model of § 7 with $r = 2$, $F_{1j} = 1$ and $F_{2j} = \pi_j$ $(j = 1, \ldots, 17)$. By using a Stein rule in this way we compromise between extreme linear smoothing, i.e. the estimates obtained by simply using the values predicted by the linear regression, and the maximum likelihood estimates.

Table 1 gives the losses

$$\sum_{j=1}^{17} (\hat{\theta}_{ij} - \theta_{ij})^2 \quad (i = 1, 2, 3)$$

and for each of the seven estimators. We use the term standardized losses since

$$\sum_{j=1}^{17} (\hat{\theta}_{ij} - \theta_{ij})^2 = \sum_{j=1}^{17} (\hat{\alpha}_{ij} - \alpha_{ij})^2/\sigma^2$$

under the obvious definition $\hat{\alpha}_{ij} = 0{\cdot}05 + \sigma\hat{\theta}_{ij}$. For comparison purposes the losses incurred by using the linear regression values themselves are listed in the last column, even though this is not a bona fide estimation rule without further specification of when the linear hypothesis would be acceptable to the statistician.

The main result is clear, that for this data the maximum likelihood estimator is worst for every $m = 10$, 15, 20 and has roughly twice the total loss of the other estimators. Equivalently, the maximum likelihood estimator requires twice as much data to be as accurate as the other estimators. Since all the Stein-type estimators dominate the maximum likelihood estimator, the improvement is achieved without having risked poorer performance in other situations.

Table 1. *Standardized losses in the Monte-Carlo experiment*

					With linear smoothing			
m	Maximum likeli- hood esti- mation	Stein's rule (1·10) applied to all 51 comp.	Stein's rule (1·10) applied separately for $m = 10$, 15, 20	Multi- variate rule (1·6) with mod. (6·7)	(1·10) 51 comp.	(1·10) separately $m = 10$, 15, 20	(1·6) as modified in (6·7)	Linear regression values
10	13·0	10·0	8·7	8·2	5·3	5·3	5·6	5·3
15	10·8	6·6	7·8	4·1	10·0	10·0	7·3	10·0
20	21·4	10·0	4·5	7·6	13·0	12·1	12·6	13·0
Total	45·2	26·6	21·0	19·9	28·3	27·4	25·5	28·3

The approximation that var $(x_{ij}) = \sigma^2$ is not very good when α_{ij} differs greatly from 0·05, as it does for small values of π_j, but the rules worked well in spite of this. When we stabilized the variance of the Z_{ij} by an arcsin transformation and applied our methods to the transformed data the estimators were not significantly changed.

References

Anderson, T. W. (1958). *An Introduction to Multivariate Analysis*. New York: Wiley.

Efron, B. & Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators. *J. Am. Statist. Ass.* **67**, 103–9.

Gollob, H. G. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika* **33**, 73–116.

James, W. & Stein, C. (1960). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp.* **1**, 361–79.

Mandel, J. (1969). The partitioning of interaction in analysis of variance. *J. Res. Nat Bur. Stand.* B, **73**B, 309–28.

Sclove, S. L., Morris, C. & Radhakrishnan, R. (1972). Nonoptimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* To appear.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In *Festschrift for J. Neyman*, ed. F. N. David, pp. 351–66. New York: Wiley.

[*Received June* 1971. *Revised October* 1971]