The Annals of Statistics
1976, Vol. 4, No. 1, 22–32

MULTIVARIATE EMPIRICAL BAYES AND ESTIMATION OF COVARIANCE MATRICES

BY BRADLEY EFRON AND CARL MORRIS

Stanford University and The Rand Corporation

The problem of estimating several normal mean vectors in an empirical Bayes situation is considered. In this case, it reduces to the problem of estimating the inverse of a covariance matrix in the standard multivariate normal situation using a particular loss function. Estimators which dominate any constant multiple of the inverse sample covariance matrix are presented. These estimators work by shrinking the sample eigenvalues toward a central value, in much the same way as the James–Stein estimator for a mean vector shrinks the maximum likelihood estimators toward a common value. These covariance estimators then lead to a class of multivariate estimators of the mean, each of which dominates the maximum likelihood estimator.

1. Introduction and summary. Let S be an observed $p \times p$ covariance matrix having the Wishart distribution with k degrees of freedom and mean $k\mathbf{\Sigma}$

$$\mathbf{S} \sim W_p(\mathbf{\Sigma}, k)\,. \tag{1.1}$$

In the context of this paper, the problem of finding multivariate empirical Bayes estimators will be shown to reduce to estimation of the inverse of the covariance matrix $\mathbf{\Sigma}$ from S, using the loss function

$$L(\mathbf{\Sigma}^{-1}, \hat{\mathbf{\Sigma}}^{-1}; \mathbf{S}) = \frac{\operatorname{tr}[(\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1})^2\mathbf{S}]}{k \operatorname{tr}(\mathbf{\Sigma}^{-1})}\,. \tag{1.2}$$

Throughout, $\mathbf{\Sigma}^{-1}$ is assumed to exist, and $k > p + 1$.

The usual estimator of $\mathbf{\Sigma}^{-1}$ is the best multiple of $\mathbf{S}^{-1}$, which for this loss function is

$$\hat{\mathbf{\Sigma}}^{-1} = (k - p - 1)\mathbf{S}^{-1}\,. \tag{1.3}$$

The estimator (1.3) is the best unbiased estimator of $\mathbf{\Sigma}^{-1}$ and is minimax with constant risk $(p + 1)/k$. We used (1.3) in [1] to derive a multivariate empirical Bayes estimator, a generalization of the James–Stein estimator [3], for cases $p \geqq 2$.

In the first main theorem,

$$\hat{\mathbf{\Sigma}}_0^{-1} \equiv (k - p - 1)\mathbf{S}^{-1} + \frac{(p^2 + p - 2)}{\operatorname{tr}(\mathbf{S})}\mathbf{I} \tag{1.4}$$

Received Octover 1974; revised May 1975.

AMS 1970 *subject classifications.* Primary 62F10; Secondary 62C99.

Key words and phrases. Multivariate empirical Bayes, Stein's estimator, minimax estimation, mean of a multivariate normal distribution, estimating a covariance matrix, James–Stein estimator, simultaneous estimation, combining estimates.

is shown to be uniformly better than (1.3) if $p \geqq 2$. Note that $\hat{\Sigma}_0^{-1}$ increases (1.3) by an amount proportional to the estimator

$$\hat{\Sigma}_1^{-1} \equiv \frac{pk-2}{\operatorname{tr}(\mathbf{S})}\mathbf{I} \tag{1.5}$$

which is the best unbiased estimator of Σ^{-1} when Σ is known to be proportional to the identity matrix. The risk functions of these estimators and their mixtures,

$$\hat{\Sigma}_\alpha^{-1} = (1-\alpha)\hat{\Sigma}_0^{-1} + \alpha\hat{\Sigma}_1^{-1} \qquad 0 \leqq \alpha \leqq 1\,, \tag{1.6}$$

which are also of interest, are considered in Sections 3 and 5.

The other main theorem, Section 4, shows that the empirical Bayes estimators derived from (1.6) are minimax, all dominating the maximum likelihood estimator $\mathbf{X}$ of a $p \times k$ matrix of means $\boldsymbol{\theta}$ for fixed $\boldsymbol{\theta}$. The case $\alpha = 1$ corresponds to the James–Stein estimator applied to all pk values θ_{ij} simultaneously while the new estimator with $\alpha = 0$ uniformly improves the multivariate empirical Bayes estimator of [1].

2. The relationship between multivariate empirical Bayes estimation and estimating the inverse of a covariance matrix. Let $\mathbf{X}_1, \cdots, \mathbf{X}_k$ be independent p-dimensional normal column vectors with $\mathbf{X}_i$ having conditional mean vector $\boldsymbol{\theta}_i$ and the identity covariance matrix $\mathbf{I}$,

$$\mathbf{X}_i \mid \boldsymbol{\theta}_i \sim_{\text{ind}} N_p(\boldsymbol{\theta}_i, \mathbf{I}) \qquad i = 1, \cdots, k\,. \tag{2.1}$$

Suppose also that the unknown parameter vectors $\boldsymbol{\theta}_i$ are an independent sample from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{A}$

$$\boldsymbol{\theta}_i \sim_{\text{ind}} N_p(\mathbf{0}, \mathbf{A}) \qquad i = 1, \cdots, k\,. \tag{2.2}$$

Then the multivariate Bayes estimator of $\boldsymbol{\theta}_i$ with respect to squared error loss is

$$\boldsymbol{\theta}_i^* \equiv (\mathbf{I} - \Sigma^{-1})\mathbf{X}_i \qquad i = 1, \cdots, k \tag{2.3}$$

with Σ defined by

$$\Sigma \equiv \mathbf{I} + \mathbf{A}\,. \tag{2.4}$$

In the empirical Bayes situation $\mathbf{A}$ and Σ are unknown, so the Bayes estimator (1.3) cannot be computed. The matrix Σ^{-1} may be estimated, however, since (2.1) and (2.2) give the marginal distribution

$$\mathbf{X}_i \sim N_p(\mathbf{0}, \Sigma) \tag{2.5}$$

to $\mathbf{X}_i$. A complete sufficient statistic for estimating Σ is $\mathbf{S} \equiv \mathbf{X}\mathbf{X}'$ having the Wishart distribution (1.1), with $\mathbf{X}$ being the $p \times k$ matrix $(\mathbf{X}_1, \cdots, \mathbf{X}_k)$.

Suppose now that the $p \times k$ matrix $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k)$ is to be estimated with normalized squared error loss function

$$D(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{pk} \sum_{i=1}^k \sum_{j=1}^p (\hat{\theta}_{ij} - \theta_{ij})^2\,, \tag{2.6}$$

by a rule similar to (2.3)

$$\hat{\boldsymbol{\theta}} \equiv (\mathbf{I} - \hat{\Sigma}^{-1})\mathbf{X}\,, \tag{2.7}$$

with $\hat{\Sigma}^{-1}$ depending only on S. Then the risk R of (2.7), which is computed by averaging (2.6) over both distributions (2.1) and (2.2), may be written

$$R = R^* + (R^0 - R^*)EL(\Sigma^{-1}, \hat{\Sigma}^{-1}; \mathbf{S})\,. \tag{2.8}$$

The value $R^0 = 1$ is the risk of the maximum likelihood estimator $\hat{\boldsymbol{\theta}} = \mathbf{X}$, i.e. $\hat{\Sigma}^{-1} = 0$, and $R^* = 1 - \operatorname{tr}(\Sigma^{-1})/p$ is the risk of the Bayes estimator (2.3) with $\hat{\Sigma}^{-1} = \Sigma^{-1}$ known. Here $L(\Sigma^{-1}, \hat{\Sigma}^{-1}; \mathbf{S})$ is the loss function (1.2). The proof of (2.8) follows easily by averaging D first over its conditional distribution

$$\boldsymbol{\theta}_i \,|\, \mathbf{X}_i \sim N_p((\mathbf{I} - \Sigma^{-1})\mathbf{X}_i, \mathbf{I} - \Sigma^{-1})\,, \tag{2.9}$$

as shown in [1, Lemma 1].

Because of (2.8), the problem of evaluating multivariate empirical Bayes estimators of the form (2.7) reduces to evaluating estimators of the inverse of an unknown covariance matrix Σ because R^0 and R^* are unaffected by the particular estimator $\hat{\Sigma}^{-1}$ under consideration and because the risk $EL(\Sigma^{-1}, \hat{\Sigma}^{-1}; \mathbf{S})$, called the "relative savings loss" in [1], only involves an expectation over S having the Wishart distribution (1.1) with Σ defined by (2.4). The special feature, that $\Sigma > \mathbf{I}$, will be ignored until Section 6.

3. An estimator of the covariance matrix Σ which dominates any multiple of S. Assume the distribution (1.1) and the loss function (1.2). We consider estimators $\hat{\Sigma}_\alpha^{-1}$ of the form (1.6). Denote $\omega \equiv \operatorname{tr}(\Sigma^{-1})/p$ and let

$$\varphi = \frac{1}{\omega} E \frac{pk - 2}{\operatorname{tr}(\mathbf{S})}\,. \tag{3.1}$$

We will show in Section 5 that $0 < \varphi \leqq 1$ for all Σ and also that

$$\varphi = \frac{1}{\omega} E \frac{\operatorname{tr}(\Sigma^{-1}\mathbf{S})}{\operatorname{tr}(\mathbf{S})}\,. \tag{3.2}$$

In the special case $\Sigma = \sigma\mathbf{I}$, the maximum value $\varphi = 1$ is attained. Denote $c \equiv (p^2 + p - 2)/(pk - 2)$ so $0 \leqq c \leqq 1$ and $0 < c < 1$ if both $p > 1$ and $k > p + 1$.

THEOREM 1. *The risk of* $\tilde{\Sigma}_\alpha^{-1}$ *is*

$$\begin{aligned} R_\alpha &\equiv EL(\Sigma^{-1}, \hat{\Sigma}_\alpha^{-1}; \mathbf{S}) \\ &= \frac{p+1}{k} + \frac{k-p-1}{k}\alpha^2 - \frac{pk-2}{pk}(c + \alpha - c\alpha)^2\varphi\,. \end{aligned} \tag{3.3}$$

In particular, $\tilde{\Sigma}_0^{-1}$ *is minimax, having risk*

$$R_0 = \frac{p+1}{k} - \frac{pk-2}{pk}c^2\varphi \tag{3.4}$$

which is uniformly smaller than the risk $(p + 1)/k$ *of the best multiple of* $\mathbf{S}^{-1}$, $(k - p - 1)\mathbf{S}^{-1}$.

PROOF. The risk of

$$\hat{\Sigma}^{-1} = a\mathbf{S}^{-1} + b\mathbf{I}/\operatorname{tr}(\mathbf{S}) \tag{3.5}$$

is computed from (1.2) as

$$\begin{aligned} &\frac{1}{pk\omega} E \operatorname{tr}(a\mathbf{S}^{-1} + b\mathbf{I}/\operatorname{tr}(\mathbf{S}) - \Sigma^{-1})^2\mathbf{S} \\ &= \frac{a^2}{pk\omega} E \operatorname{tr}(\mathbf{S}^{-1}) + \frac{2ab}{k\omega} E \frac{1}{\operatorname{tr}(\mathbf{S})} - \frac{2a}{k} \\ &\quad + \frac{b^2}{pk\omega} E \frac{1}{\operatorname{tr}(\mathbf{S})} - \frac{2b}{pk\omega} E \frac{\operatorname{tr}(\Sigma^{-1}\mathbf{S})}{\operatorname{tr}(\mathbf{S})} + \frac{1}{pk\omega} E \operatorname{tr}(\Sigma^{-2}\mathbf{S}) \\ &= \frac{a^2}{k(k-p-1)} + \frac{2ab}{k(pk-2)}\varphi - \frac{2a}{k} + \frac{b^2}{pk(pk-2)}\varphi \\ &\quad - \frac{2b}{pk}\varphi + 1 \end{aligned} \tag{3.6}$$

where we have used (3.1), (3.2) and $E(k - p - 1)\mathbf{S}^{-1} = \Sigma^{-1}$. The minimizing value of b is obtained by differentiating (3.6) and is $b^* = pk - 2 - ap$ which is independent of the unknown parameters. Inserting b^* into (3.6) and simplifying gives

$$R = \frac{p+1}{k} + \frac{(k-p-1-a)^2}{k(k-p-1)} - \frac{(pk-2-ap)^2}{pk(pk-2)}\varphi\,. \tag{3.7}$$

Reparameterizing with $a = (k - p - 1)(1 - \alpha)$ and substituting this value into (3.7) yields (3.3). Assertion (3.4) follows by setting $\alpha = 0$ in (3.3). The proof is complete.

DISCUSSION. If φ is known, then R_α is minimized at

$$\alpha^* = c\varphi/[1 - \varphi + c\varphi] \tag{3.8}$$

which increases monotonically from 0 to 1 as φ increases from 0 to 1. The risk of (1.6) with $\alpha = \alpha^*$ is

$$R_{\alpha^*} = \alpha^* \frac{2}{pk} + (1 - \alpha^*)\frac{p+1}{k}\,. \tag{3.9}$$

The case $\varphi = 1$ (Σ proportional to the identity), $\alpha^* = 1$, yields the rule (1.5) as an estimate.

More generally, if a prior distribution on Σ is given, then the rule of the form (3.5) that minimizes the average risk takes the form (1.6) with

$$\alpha^{**} = cE\varphi/[1 - E\varphi + cE\varphi]\,. \tag{3.10}$$

This depends only on the a priori mean $E\varphi$ of φ. Then $R_{\alpha^{**}}$ is given by (3.9) with α^* replaced by α^{**}. Formulas (3.8)—(3.10) are proven by averaging (3.3) over the prior distribution, and then by differentiating (3.3), perhaps most easily in the form

$$R_\alpha = \frac{p+1}{k} + \frac{pk-2}{pk}[(1-c)\alpha^2 - (c+\alpha-c\alpha)^2 E\varphi]\,. \tag{3.11}$$

The minimal complete subclass of the class of all rules of the form (3.5) with $-\infty < a,\ b < \infty$ is the class of rules $\hat{\Sigma}_\alpha^{-1}$ (1.6) with $0 \leqq \alpha \leqq 1$. This follows from the fact that the Bayes rules are given by (3.10) and that $b^* = pk - 2 - ap$ is the optimal choice for b.

There are many minimax estimators (rules with risk not exceeding $(p+1)/k$) in the class (3.5). The best such estimator is $\hat{\Sigma}_0^{-1}$ because the minimax estimators must have $a = k - p - 1$ to perform well at $\varphi = 0$, and then $b = p^2 + p - 2$ is the best choice for b.

4. Using the covariance estimators in a simultaneous estimation problem: minimax estimation. In the context of Section 2, estimators of the $p \times k$ matrix $\boldsymbol{\theta}$ of the form

$$\hat{\boldsymbol{\theta}}_\alpha = (\mathrm{I} - \hat{\Sigma}_\alpha^{-1})\mathrm{X} \tag{4.1}$$

are suggested, $\hat{\Sigma}_\alpha$ given by (1.6). For fixed $\boldsymbol{\theta}$, Theorem 2 will show that each $\hat{\boldsymbol{\theta}}_\alpha$ is minimax (dominates the maximum likelihood estimator X) as an estimator of the mean $\boldsymbol{\theta}$ of a multivariate normal distribution. Furthermore, $\hat{\boldsymbol{\theta}}_0$ dominates the estimator implied by (1.3), which was presented in [1].

The risk of (4.1) averaged over the distribution (2.2) of $\boldsymbol{\theta}$ will be needed. It is

$$ED(\boldsymbol{\theta}, \boldsymbol{\theta}_\alpha) = 1 - \omega + \omega R_\alpha \tag{4.2}$$

from (2.8) and Theorem 1, $\omega \equiv \mathrm{tr}\,(\Sigma^{-1})/p$.

THEOREM 2. *As a function of* $\boldsymbol{\theta}$ *the rule* $\hat{\boldsymbol{\theta}}_\alpha$ *of* (4.1) *has risk*

$$E_{\boldsymbol{\theta}} D(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_\alpha) = 1 - \frac{(k-p-1)^2}{pk}(1-\alpha^2)E_{\boldsymbol{\theta}}\,\mathrm{tr}\,(\mathrm{S}^{-1}) - \frac{(pk-2)^2}{pk}(c+\alpha-c\alpha)^2 E_{\boldsymbol{\theta}}\frac{1}{\mathrm{tr}\,(\mathrm{S})}\,. \tag{4.3}$$

PROOF. First, (4.2) may be written as'

$$ED(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_\alpha) = 1 - \frac{(k-p-1)^2}{pk}(1-\alpha^2)E\,\mathrm{tr}\,(\mathrm{S}^{-1}) - \frac{(pk-2)^2}{pk}(c+\alpha-c\alpha)^2 E\frac{1}{\mathrm{tr}\,(\mathrm{S})}\,. \tag{4.4}$$

This follows because

$$\omega = (k-p-1)E\,\mathrm{tr}\,(\mathrm{S}^{-1})/p\,, \tag{4.5}$$

while (3.1) provides an expression for $\omega\varphi$. Both sides of (4.4) involve first an expectation $E_{\boldsymbol{\theta}}$ over the distribution (2.1) of $\mathbf{X}$ given $\boldsymbol{\theta}$, this expectation being a function of $\boldsymbol{\Lambda} \equiv \boldsymbol{\theta\theta}'$ only, followed by an expectation $E_{\mathbf{A}}$ over the distribution (2.2) of $\boldsymbol{\theta}$ for fixed $\mathbf{A}$. Since the family of distributions of $\boldsymbol{\Lambda}$ is complete for $\mathbf{A}$, (4.4) holds even when the $E_{\mathbf{A}}$ expectation is removed. This ends the proof.

Note that the right-hand side of (4.3), with the expectation sign removed, provides an unbiased estimate of both risks (4.2) and (4.3) of $\hat{\boldsymbol{\theta}}_\alpha$. The James–Stein estimator is the rule $\alpha = 1$ with risk

$$(4.6) \qquad 1 - \frac{(pk-2)^2}{pk} E_0 \frac{1}{\operatorname{tr}(\mathbf{S})} .$$

The estimator with $\alpha = 0$,

$$(4.7) \qquad \hat{\boldsymbol{\theta}}_0 \equiv \left(\mathbf{I} - (k-p-1)\mathbf{S}^{-1} - \frac{p^2+p-2}{\operatorname{tr}(\mathbf{S})}\mathbf{I}\right)\mathbf{X} ,$$

is the best in the class $\hat{\boldsymbol{\theta}}_\alpha$ as $\tau \equiv \operatorname{tr}(\boldsymbol{\theta\theta}') \to \infty$ and improves the risk $1 - (k - p - 1)^2 E_{\boldsymbol{\theta}} \operatorname{tr}(\mathbf{S}^{-1})/pk$ of $\hat{\boldsymbol{\theta}} = (\mathbf{I} - (k-p-1)\mathbf{S}^{-1})\mathbf{X}$ by the amount

$$(4.8) \qquad \frac{(p^2+p-2)^2}{pk} E_{\boldsymbol{\theta}} \frac{1}{\operatorname{tr}(\mathbf{S})} .$$

The improvement (4.8) is largest at $\boldsymbol{\theta} = 0$ where it is $(p^2 + p - 2)^2/pk(pk - 2)$.

Bounds on the last term of (4.3), (4.6), and (4.8) may be computed for any $\boldsymbol{\theta}$ from the fact that

$$(4.9) \qquad \frac{1}{1+\tau/(pk-2)} \leqq E_{\boldsymbol{\theta}} \frac{pk-2}{\operatorname{tr}(\mathbf{S})} \leqq \frac{1}{1+\tau/pk} .$$

Formula (4.9) is sharp if either $\boldsymbol{\theta} = \mathbf{0}$ or if $pk \to \infty$.

An upper bound for the risk (4.3) as a function of $\boldsymbol{\theta}$ may be derived from (4.9) and (4.10), which follows from Jensen's inequality.

$$(4.10) \qquad E \operatorname{tr}(\mathbf{S}^{-1}) \geqq \operatorname{tr}(k\mathbf{I} + \boldsymbol{\theta\theta}')^{-1} .$$

This expression is not sharp since $E \operatorname{tr}(\mathbf{S}^{-1}) = p/(k-p-1)$ when $\boldsymbol{\theta} = \mathbf{0}$, while (4.10) gives p/k.

When $\boldsymbol{\theta} = \mathbf{0}$, then the risk (4.3) is identical to R_α', the value (3.3) with $\varphi = 1$. This follows from (4.2) with $\omega = 1$. Then

$$(4.11) \qquad E_0 D(\mathbf{0}, \hat{\boldsymbol{\theta}}_\alpha) = R_\alpha' = 2/pk + c(1-c)(1-\alpha)^2(pk-2)/pk .$$

Assertion (4.9) needs proof. Since $\operatorname{tr}(\mathbf{S})$ has a noncentral chi-square distribution with mean $pk + \tau$, $\operatorname{tr}(\mathbf{S}) \sim \chi^{2\prime}_{pk}(\tau)$, it can be written as a Poisson mixture of central chi-squares as in [5], say $\operatorname{tr}(\mathbf{S}) \sim \chi^2_{pk+2J}$, $J \sim$ Poisson with mean $\tau/2$. Letting E_τ indicate expectation with respect to the Poisson distribution, then

$$(4.12) \qquad E_{\boldsymbol{\theta}} \frac{1}{\operatorname{tr}(\mathbf{S})} = E_\tau \frac{1}{pk+2J-2} .$$

The left-hand side of (4.9) follows from Jensen's inequality applied to (4.12).

To obtain the right-hand inequality, write $E_\tau \cdot 1/(pk + 2J - 2)$ as

$$\frac{1}{pk - 2}\left[1 - \sum_{j=0}^{\infty} \frac{e^{-\tau/2}(\tau/2)^j}{j!} \frac{2j}{pk + 2j - 2}\right]$$

and notice that this also can be expressed as $[1 - \tau E_\tau \cdot 1/(pk + 2J)]/(pk - 2)$. Jensen's inequality $E \cdot 1/(pk + 2J) \geqq 1/(pk + \tau)$ gives the result.

5. Risk functions and the function φ. We will now give a more explicit evaluation of the function φ which appears in the risk formula (3.3). Let $W_1, \cdots, W_p$ be independent χ_k^2 random variables and $U_i = W_i/\sum W_j$. Let $\sigma_1, \cdots, \sigma_p$ be the eigenvalues of $\mathbf{\Sigma}$, $\omega = \operatorname{tr}(\mathbf{\Sigma}^{-1})/p = \sum (1/\sigma_j)/p$ and define

$$\varphi \equiv \frac{1}{\omega} E\left(\sum_{j=1}^{p} \sigma_j U_j\right)^{-1} . \tag{5.1}$$

The value (5.1) agrees with (3.2) because orthogonal invariance permits the assumption $\mathbf{\Sigma}$ diagonal with elements $\sigma_1, \cdots, \sigma_p$ and then (3.2) with $W_i = S_{ii}/\sigma_i$ reduces to $(1/\omega)E(\sum W_i/\sum \sigma_i W_i)$, being (5.1). Because $\sum W_j$ is independent of $(U_1, \cdots, U_p)$,

$$\begin{aligned} \varphi &= \frac{1}{\omega} E \frac{1}{\sum \sigma_i U_i} E \frac{pk - 2}{\sum W_j} \\ &= \frac{1}{\omega} E \frac{pk - 2}{\sum \sigma_i W_i} = \frac{1}{\omega} E \frac{pk - 2}{\operatorname{tr}(\mathbf{S})} \end{aligned} \tag{5.2}$$

establishing the equivalence of (5.1) and (3.1), and also (3.1) and (3.2). Note $0 < \varphi \leqq 1$ since $1/\sum \sigma_i U_i \leqq \sum U_i/\sigma_i$ and $E \sum U_i/\sigma_i = (\sum 1/\sigma_i)/p = \omega$.

Define

$$\rho \equiv p/\omega \operatorname{tr}(\mathbf{\Sigma}) \tag{5.3}$$

as the squared cosine of the angle beween $\mathbf{\Sigma}^{\frac{1}{2}}$ and $\mathbf{\Sigma}^{-\frac{1}{2}}$, so $0 \leqq \rho \leqq 1$. Jensen's inequality applied to (5.1) shows $\varphi \geqq \rho$. We have bounds

$$\rho \leqq \varphi \leqq \min\left(1, \frac{kp - 2}{kp - 2p} \cdot \rho\right) , \tag{5.4}$$

since letting $\pi_i = \sigma_i/\sum \sigma_j$ in (5.1) gives

$$\frac{1}{\sum \sigma_i U_i} = \frac{1}{\sum \sigma_i} \frac{1}{\sum \pi_i U_i} \leqq \frac{1}{\sum \sigma_i} \sum \pi_i/U_i . \tag{5.5}$$

Taking expectations of (5.5) and using $E \cdot 1/U_i = (kp - 2)/(k - 2)$ for all i proves (5.4). The bounds (5.4) become tight as k increases and for any p,

$$\lim_{k \to \infty} \varphi_k = \rho . \tag{5.6}$$

The index henceforth will be used to indicate the dependence of φ on k. The values φ_k and ρ are unity only when $\mathbf{\Sigma} = \sigma \mathbf{I}$, i.e., only when all σ_i are equal, and the lower bound of (5.4), ρ, is the better approximation when the σ_i are nearly equal. Dispersed σ_i cause φ_k and ρ both to approach zero with the upper bound of (5.4) being attained asymptotically if at least one σ_i is finite and one σ_i approaches infinity.

In the special case $p = 2$, φ_k depends on Σ only through the ratio $\lambda = \sigma_2/\sigma_1$ of the largest to the smallest eigenvalue. Then values of φ_k are generated recursively for $\lambda \neq 1$ by

$$(5.7) \qquad \varphi_1 = 2\lambda^{\frac{1}{2}}/(\lambda + 1) = \rho^{\frac{1}{2}}\,, \qquad \varphi_2 = 2\lambda \log(\lambda)/(\lambda^2 - 1)$$

$$\varphi_k = \frac{k-1}{k-2}\,\frac{4\lambda}{(\lambda-1)^{\cdot}}(1 - \varphi_{k-2}) = \frac{k-1}{k-2}\,\frac{\rho}{1-\rho}(1 - \varphi_{k-2})\,, \qquad k \geqq 3\,.$$

Obviously $\varphi_k = 1$ if $\lambda = 1$. We omit the proof of (5.7) to save space. The limiting value of φ_k as $k \to \infty$ is $\rho = 4\lambda/(1 + \lambda)^2$.

The function φ_6 is plotted in Figure 1 for the case $p = 2$, $k = 6$ together with the four risks, from (3.3),

$$(5.8) \qquad R_\alpha = .5 + .5\alpha^2 - \tfrac{2}{15}(1 + 1.5\alpha)^2\varphi_6$$

for $\alpha = 0, .25, .50, 1$. Figure 1 illustrates that $\alpha = 0$ is best if $\varphi = 0$ and $\alpha = 1$ is best if $\varphi = 1$ as confirmed by (3.8), while intermediate values like $\alpha = .25$ and $\alpha = .5$ are effective compromises if the extremes $\varphi = 0$ or $\varphi = 1$ are not especially likely. It is tempting to estimate φ, say by a function $C/[\mathrm{tr}\,(\mathrm{S}^{-1})\,\mathrm{tr}\,(\mathrm{S})]$,

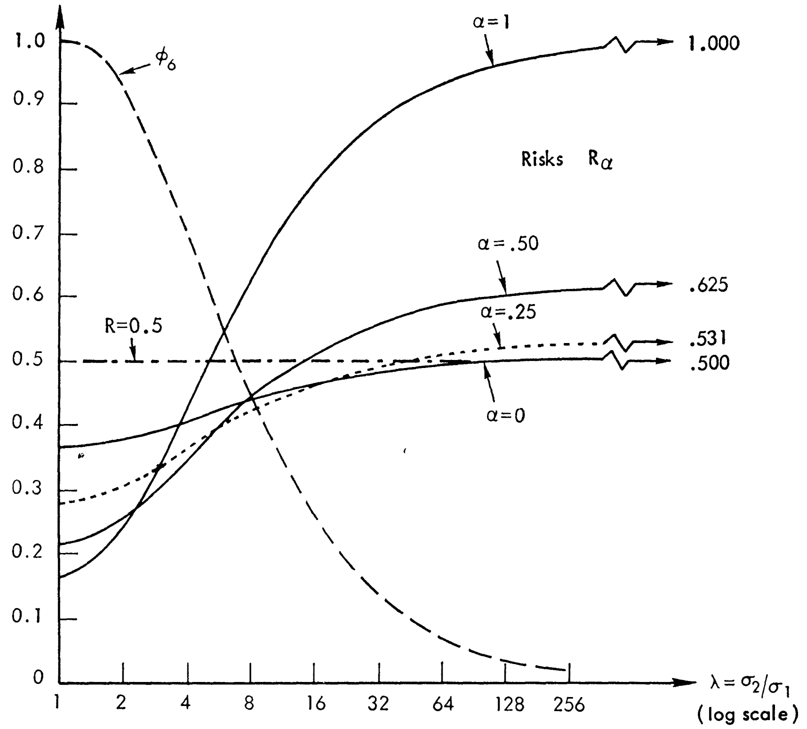

FIG. 1. A plot of ϕ_6 and the risks (relative savings losses) R_0, $R_{.25}$, $R_{.5}$, R_1 of (3.6) against the ratio of the largest to the smallest eigenvalue for the case $p = 2$, $k = 6$.

C close to $C_0 \equiv p(pk-2)/(k-p-1)$, and to use this to determine an estimated value $\hat{\alpha}$ from (3.8). In the situation of Figure 1, for example, the hope would be to produce a rule with risk function close to the lower envelope of the risk functions graphed. Our calculations for the case $p=2$ show that the rule with $C=C_0$ works fairly well, provided $\hat{\alpha}$ is forced to be less than unity, but that smaller values of C are even better. However, no clear guidelines for the use of such "adaptive" rules are available at this time.

The improvement of the rule $\alpha = 0$ over the best multiple of $\mathbf{S}^{-1}$ is measured by the distance between the R_0 curve and the horizontal line $R = .5$ in Figure 1. This is a 27 percent improvement in risk at $\lambda = 1$; larger improvements can occur in cases with k large and p near k.

For any p, k, $\hat{\Sigma}_0^{-1}$ has lower risk than $\hat{\Sigma}_1^{-1}$ provided $\varphi \leqq 1/(1+c)$. This holds for $p=2$, $k=6$ provided $\lambda^{\frac{1}{2}} \geqq 1.90$. Note that $\lambda^{\frac{1}{2}}$ is the ratio of the standard deviations of the major and the minor principal components defined by the two rows of $\mathbf{X}$.

6. The restriction $\Sigma^{-1} \leqq \mathbf{I}$. We know $\Sigma^{-1} \leqq \mathbf{I}$ since $\Sigma = \mathbf{I} + \mathbf{A}$ with $\mathbf{A}$ nonnegative definite, but the estimators $\hat{\Sigma}_\alpha^{-1}$ of (1.6) do not obey this inequality. This undesirable feature may be overcome as follows. Diagonalize $\hat{\Sigma}_\alpha^{-1} = \mathbf{\Gamma}'\mathbf{\Delta}\mathbf{\Gamma}$ with $\mathbf{\Gamma}$ a $p \times p$ orthogonal matrix and $\mathbf{\Delta}$ the diagonal matrix of eigenvalues δ_i. A preferred estimate is $\overset{*}{\Sigma}_\alpha^{-1} \equiv \mathbf{\Gamma}'\mathbf{\Delta}^*\mathbf{\Gamma}$ with $\delta_i^* \equiv \min(1, \delta_i)$, $i = 1, \cdots, p$, since this estimate satisfies the restriction $\overset{*}{\Sigma}_\alpha^{-1} \leqq \mathbf{I}$. The loss function (1.2) is either unchanged or reduced for every $\mathbf{S}$, Σ by this modification,

$$L(\Sigma^{-1}, \overset{*}{\Sigma}_\alpha^{-1}; \mathbf{S}) \leqq L(\Sigma^{-1}, \hat{\Sigma}_\alpha^{-1}; \mathbf{S}) \tag{6.1}$$

for all $\mathbf{S}$.

The improved estimator $\overset{*}{\Sigma}_\alpha^{-1}$ has risk uniformly lower than R_α of (3.3) because of (6.1). In the simultaneous estimation context of Section 4, the estimator

$$\overset{*}{\boldsymbol{\theta}}_\alpha \equiv (I - \overset{*}{\Sigma}_\alpha^{-1})\mathbf{X} \tag{6.2}$$

therefore has risk as a function of $\mathbf{A}$, $E_{\mathbf{A}} E_{\boldsymbol{\theta}} D(\boldsymbol{\theta}, \overset{*}{\boldsymbol{\theta}}_\alpha)$, strictly lower than (4.4). The risk as a function of $\boldsymbol{\theta}$, $E_{\boldsymbol{\theta}} D(\boldsymbol{\theta}, \overset{*}{\boldsymbol{\theta}}_\alpha)$, is likely to be lower than (4.5) for all $\boldsymbol{\theta}$, and is known to be for $p = 1$. This conjecture is not proved for $p \geqq 2$ however because the completeness argument used to establish (4.5) does not apply with $\overset{*}{\boldsymbol{\theta}}_\alpha$ (there is no convenient expression for its risk as a function of $\mathbf{A}$).

The proof of (6.1) notes the convexity of the set of matrices $\mathbf{0} \leqq \Sigma^{-1} \leqq \mathbf{I}$, the fact that the loss function L is a metric derived from an Euclidean inner product, and that in this metric $\overset{*}{\Sigma}_\alpha^{-1}$ is the closest matrix in the convex set to $\hat{\Sigma}_\alpha^{-1}$. The precise argument is given in [1, Section 6].

7. Discussion. The fact that $\hat{\Sigma}_0^{-1}$ dominates the best fully invariant estimator $(k-p-1)\mathbf{S}^{-1}$ of Σ^{-1} for our fully invariant loss function suggests that shrinking the best multiple of $\mathbf{S}$ toward the identity matrix may be effective in more general situations of estimating a covariance matrix. All of the estimators of $\mathbf{\Sigma}$ in this paper are orthogonally invariant, of the form

$$\hat{\mathbf{\Sigma}}(\mathbf{S}) = \mathbf{\Gamma}'\hat{\sigma}\mathbf{\Gamma} \tag{7.1}$$

with $\mathbf{\Gamma}$ the matrix of eigenvectors of $\mathbf{S}$, say $\mathbf{S} = \mathbf{\Gamma}'\mathbf{D}\mathbf{\Gamma}$, $\mathbf{D}$ diagonal, and $\hat{\boldsymbol{\sigma}}$ a diagonal matrix whose entries are functions of the eigenvalues $\mathbf{D}$ of $\mathbf{S}$, $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}(\mathbf{D})$. Explicitly, the best linear multiple of $\mathbf{S}$, $\hat{\mathbf{\Sigma}}(\mathbf{S}) = \mathbf{S}/(k - p - 1)$, estimates the ith eigenvalue of $\mathbf{\Sigma}$ by $\hat{\sigma}_i = d_i/(k - p - 1)$, while $\hat{\mathbf{\Sigma}}_0 = ((k - p - 1)\mathbf{S}^{-1} + (p^2 + p - 2)\mathbf{I}/\mathrm{tr}\,())\mathbf{S}^{-1}$ uses

$$\hat{\sigma}_i^{(0)} = \frac{1}{1 + \left(\frac{p^2 + p - 2}{k - p - 1}\right)\frac{d_i}{\sum d_j}}\hat{\sigma}_i\,, \tag{7.2}$$

so improves on $\hat{\sigma}_i$ by shrinking all the estimated eigenvalues toward zero, the larger eigenvalues being shrunk proportionately more than the smaller. This is reminiscent of the James-Stein estimator of k means [3], and the basic phenomenon seems to be the same: the eigenvalues of $\mathbf{S}$, considered as an ensemble of p numbers, are distorted in a systematic nonlinear way from the eigenvalues of $\mathbf{\Sigma}$. A universally improved estimator is obtained by undoing this distortion.

For the general problem of estimating a covariance matrix, it would be more satisfying to show that estimators of the form

$$\mathbf{\Sigma} = (a\mathbf{S}^{-1} + b\mathbf{I}/\mathrm{tr}\,(\mathbf{S}))^{-1} \tag{7.3}$$

dominate the best fully invariant estimator of $\mathbf{\Sigma}$ when the loss function is also fully invariant, but the computations are difficult for such loss functions. The loss function used here leads to nicely computable risk expressions for rules of the form (7.3), permitting a comparison of their operating characteristics, and more important, shows where the additional information lies for improving the best fully invariant estimator. It also has the virtue of arising naturally from the squared error estimation problem for $\boldsymbol{\theta}$.

In Section 5 of [3], Stein considered a covariance estimation example with a fully invariant loss function and found a constant-risk estimator (invariant under the lower triangular group of matrices, but not orthogonally invariant) which is uniformly better than the best fully invariant estimator. The expected value of his estimator, like $\hat{\mathbf{\Sigma}}_0$ here, is always closer to $\mathbf{0}$ than the mean of the best fully invariant estimator. He has recently made further progress on the problem of covariance estimation by using a method for finding unbiased estimators of the risk function [7].

In the empirical Bayes and the simultaneous estimation of means situations the loss function L is natural, as the derivation in Section 2 shows, and the simple estimators of $\boldsymbol{\theta}$ (2.7) based on the form (7.3) have computable risks. This simplicity also leads to risk expressions as a function of $\boldsymbol{\theta}$ (Theorem 2) and yields unbiased estimates of the risk. These estimators may be criticized for being inadmissible since they ignore the restriction $\Sigma^{-1} \leqq \mathbf{I}$. The rules of Section 6 may be nearly admissible though; at least in the case $p = 1$ they reduce to the James–Stein positive-part estimator for which no uniform improvement has ever been offered.

Orthogonally invariant estimators of $\boldsymbol{\theta}$ take the form (2.7) with $\hat{\Sigma}$ as in (7.1), and are not necessarily of the form (7.3). One approach to finding alternatives to (7.3) was suggested at the end of Section 5. Stein [7] offers another method by producing unbiased estimates of the risk of arbitrary orthogonally invariant rules. Other rules having this orthogonality property are offered by Gollob [2] and Mandel [4]. Their estimates of $\boldsymbol{\theta}$ correspond to using (7.1) in (2.7) where $1/\hat{\sigma}_i = 1$ if d_i fails to pass a significance test and otherwise is zero, forcing $\mathbf{0} \leqq \hat{\Sigma}^{-1} \leqq \mathbf{I}$. When $p = 1$ their rule is equivalent to estimation following a preliminary test that $\boldsymbol{\theta} = \mathbf{0}$, a procedure that is known not to be minimax and to be dominated uniformly by the positive-part version of a Stein-type estimator [6].

Acknowledgment. This paper is a refinement of a technical report [8] written jointly with Charles Stein. Most of the ideas presented here are contained in that earlier document.

REFERENCES

[1] EFRON, B. and MORRIS, C. (1972). Empirical Bayes on vector observations—An extension of Stein's method. *Biometrika* **59** 335-347.

[2] GOLLOB, H. F. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika* **33** 73-116.

[3] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 361-379. Univ. of California Press.

[4] MANDEL, J. (1969). The partitioning of interaction in analysis of variance. *J. Res. Nat. Bur. Standards Sect. B* **73** 309-328.

[5] ROBBINS, H. and PITMAN, E. J. G. (1949). Application of the method of mixtures to quadratic forms in normal variables. *Aun. Math. Statist.* **20** 552-560.

[6] SCLOVE, S. L., MORRIS, C. and RADHAKRISHNAN, R. (1972). Nonoptimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **45** 1481-1490.

[7] STEIN, C. (1973). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. Asymptotic Statist.* 345-381.

[8] STEIN, C., EFRON, B. and MORRIS, C. (1972). Improving the usual estimator of a normal covariance matrix. Technical Report No. 37, Department of Statistics, Stanford Univ.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

THE RAND CORPORATION
1700 MAIN STREET
SANTA MONICA, CALIFORNIA 90406