The Annals of Statistics
1980, Vol. 8, No. 3, 586–597

EMPIRICAL BAYES ESTIMATION OF THE MULTIVARIATE NORMAL COVARIANCE MATRIX

BY L. R. HAFF

University of California, San Diego, La Jolla

Let $\mathbf{S}_{p\times p}$ have a Wishart distribution with scale matrix Σ and k degrees of freedom. Estimators of Σ are given for each of the loss functions $L_1(\hat{\Sigma}, \Sigma) = \text{tr}\,(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1}) - p$ and $L_2(\hat{\Sigma}, \Sigma) = \text{tr}\,(\hat{\Sigma}\Sigma^{-1} - I)^2$. The obvious estimators of Σ are the scalar multiples of $\mathbf{S}$, i.e., $a\mathbf{S}$ where $0 < a \leqslant 1/k$. (Recall that $(1/k)\mathbf{S}$ is unbiased.) For each problem $(\Sigma, \hat{\Sigma}, L_i)$, $i = 1, 2$, we provide empirical Bayes estimators which dominate $a\mathbf{S}$ by a substantial amount. It is seen that the uniform reduction in the risk function determined by L_2 is at least $100(p + 1)/(k + p + 1)\%$. Dominance results for L_1 and L_2 were first given by James and Stein.

1. Introduction and summary. The problem under consideration here is that of estimating the multi-normal covariance matrix Σ. Empirical Bayes (EB) alternatives are derived which dominate *all* scalar multiples of the unbiased estimator. The risk of the unbiased estimator is reduced by a significant amount in each case.

In [3] and [4], the author treated the problem of estimating Σ^{-1}. An identity for the Wishart distribution was derived and used to compute an unbiased estimator for the risk function. Dominance results were then obtained by working with the unbiased estimator. The same technique is widely used in the present paper. (The Wishart identity is stated in line (2.4) below. Unknown to the author, Charles Stein essentially derived it several years ago. His approach was different, however, and his work remains unpublished—see [12].)

Our present results have the flavor of those given by James and Stein [6], pages 376, 377. We provide estimators which dominate the usual unbiased estimator for each of two invariant loss functions. Our present methods, however, supersede those of [6], and the implications are deeper. Whereas the estimator in [6] is only slightly better than the usual one, the EB estimators are substantially better.

Let $\mathbf{S}_{p\times p}$ have a Wishart distribution with unknown matrix Σ and k degrees of freedom, i.e.,

$$\mathbf{S} \sim W(\Sigma, k) \text{ for } k - p - 1 > 0. \tag{1.1}$$

Also, let $\hat{\Sigma}$ be an estimator of Σ. We assume that the loss function is

$$L_1(\hat{\Sigma}, \Sigma) = \text{tr}\,(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1}) - p \text{ or } L_2(\hat{\Sigma}, \Sigma) = \text{tr}\,(\hat{\Sigma}\Sigma^{-1} - I)^2 \tag{1.2}$$

(from [6]), and define the risk function by

$$R_i(\hat{\Sigma}, \Sigma) \equiv E\big[L_i(\hat{\Sigma}, \Sigma)|\Sigma, k\big], \qquad i = 1 \text{ or } 2.$$

Received January 1978; revised February 1979.

AMS 1970 *subject classifications.* Primary 62F10: secondary 62C99.

Key words and phrases. Covariance matrix, empirical Bayes estimators, unbiased estimation of risk function.

(The latter is an average with respect to the $W(\Sigma, k)$ distribution.) If $\hat{\Sigma}$ and $\hat{\Sigma}_*$ are competing estimators of Σ, then "$\hat{\Sigma}$ dominates $\hat{\Sigma}_*$ (mod L_i)" will mean $R_i(\hat{\Sigma}, \Sigma) \leqslant R_i(\hat{\Sigma}_*, \Sigma)$ $(\forall\Sigma)$.

Our EB estimators have the form

$$\hat{\Sigma} = a[\mathbf{S} + \mathbf{u}t(\mathbf{u})C] \tag{1.3}$$

with $0 < a \leqslant 1/k$, $\mathbf{u} = 1/\text{tr}\,(\mathbf{S}^{-1}C)$, $t(\cdot)$ nonincreasing, and C an arbitrary positive definite matrix. For $t(\cdot) \equiv 0$, we have the obvious estimators, the scalar multiples of $\mathbf{S}$. In particular, note that $E(1/k)\mathbf{S} = \Sigma$.

For loss function L_1, Stein [6] derived a minimax estimator and proved that for any $\beta > 0$, $\beta(\mathbf{S}/k)$ is not minimax. He obtained similar results for L_2, but was unable to get an explicit formula for a minimax estimator in this case. Selliah [10] gave additional results for L_2. Later, Perlman [8] showed that if $(kp - 2)/(kp + 2) \leqslant \beta < 1$, then $\beta(\mathbf{S}/k)$ dominates $\mathbf{S}/k$ with respect to a general quadratic loss function. All these estimators (from [6], [8], and [10]) are only slightly better than the unbiased estimator. Finally, in the 1975 Rietz lecture [12], Stein described one (for L_1) which is substantially better. His methods were similar to those used in the following.

A summary of the present work. Among the scalar multiples of $\mathbf{S}$, we prove that the best estimator (mod L_1) is the unbiased estimator

$$\hat{\Sigma}_1 = (1/k)\mathbf{S}, \tag{1.4}$$

and the best estimator (mod L_2) is

$$\hat{\Sigma}_2 = [1/(k + p + 1)]\mathbf{S}.$$

Our main results concern the estimators in line (1.3). These, (1.3), are derived as empirical Bayes estimators. Then, for each loss function, conditions are given under which they dominate the best scalar multiple of $\mathbf{S}$. Under L_2, we have the following numerical comparisons: it is seen that $R_2(\hat{\Sigma}_1, \Sigma) = p(p + 1)/k$ and $R_2(\hat{\Sigma}, \Sigma) \leqslant R_2(\hat{\Sigma}_2, \Sigma) = p(p + 1)/(k + p + 1)$ $(\forall\Sigma)$. Thus $R_2(\hat{\Sigma}, \Sigma)$ is less than $R_2(\hat{\Sigma}_1, \Sigma)$ by *at least* 100 $(p + 1)/(k + p + 1)\%$ $(\forall\Sigma)$. For $p = 3$ and $k = 8$, say, the uniform improvement is at least 33%. In addition, our Monte Carlo results (unpublished) indicate that $R_2(\hat{\Sigma}, \Sigma)$ is less than $R_2(\hat{\Sigma}_2, \Sigma)$ by as much as 16% ($p = 3$ and $k = 8$).

Finally, there is a parallel between our results and those of Stein [6]. Roughly speaking, the EB estimators which perform well (mod L_1) also perform well (mod L_2). Although we give explicit estimators under L_2, the calculations are difficult, and the results lack the generality of the L_1 results.

2. A brief description of the mathematics. Let δ be a real number, $V_{p\times p}$ a symmetric matrix. If we apply the expansion

$$\log\det(I + \delta V) = \Sigma_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}\delta^n\,\text{tr}\,(V^n) \tag{2.1}$$

to L_1, then a relationship is obtained between L_1 and L_2. (The series converges if the spectral radius of V is less than unity and $0 < \delta < 1$.) In particular, set $\delta = 1$, factor Σ^{-1} as $\Sigma^{-1} = \Omega^2$, and expand

$$\begin{aligned} \log\det(\hat{\Sigma}\Sigma^{-1}) &= \log\det(\Omega\hat{\Sigma}\Omega) \\ &= \log\det(I + V)\ (V = \Omega\hat{\Sigma}\Omega - I) \\ &= \operatorname{tr}(V) - \left(\tfrac{1}{2}\right)\operatorname{tr}(V^2) + \left(\tfrac{1}{3}\right)\operatorname{tr}(V^3) - \cdots \\ &= \operatorname{tr}(\hat{\Sigma}\Sigma^{-1} - I) - \left(\tfrac{1}{2}\right)\operatorname{tr}(\hat{\Sigma}\Sigma^{-1} - I)^2 + \cdots . \end{aligned} \tag{2.2}$$

From (2.2) loss function L_1 can be written as

$$L_1(\hat{\Sigma}, \Sigma) = \left(\tfrac{1}{2}\right)L_2(\hat{\Sigma}, \Sigma) - \left(\tfrac{1}{3}\right)\operatorname{tr}(\hat{\Sigma}\Sigma^{-1} - I)^3 + \cdots , \tag{2.3}$$

so it is plausible that estimators which perform well (mod L_1) also perform well (mod L_2).

The series (2.1) is used to obtain an approximation for $\alpha_1(\Sigma) \equiv R_1(\hat{\Sigma}, \Sigma) - R_1(\hat{\Sigma}_1, \Sigma)$. This approximation, call it $\underline{\alpha}_1(\Sigma)$, has the property that $\underline{\alpha}(\Sigma) < 0\ (\forall\Sigma)$ implies $\alpha_1(\Sigma) < 0\ (\forall\Sigma)$. We shall force $\underline{\alpha}_1(\Sigma) < 0\ (\forall\Sigma)$ by using an identity for the Wishart distribution.

The identity and its application. For suitable choices of a matrix $T = T(\mathbf{S}, \Sigma)_{p\times p}$, a scalar $h(S)$, and a constant $r \neq 0$; the identity is stated in terms of the following:

(i) $\operatorname{diag}(T)$, a diagonal matrix with diagonal elements equal to those of T;
(ii) $T_{(r)} \equiv rT + (1 - r)\operatorname{diag}(T)$;
(iii) $D^*T_{(r)} \equiv \sum_{i=1}^{p}\partial t_{ii}/\partial s_{ii} + r\sum_{i\neq j}\partial t_{ij}/\partial s_{ij}$; and
(iv) $\partial h/\partial S \equiv (\partial h/\partial s_{ij})_{p\times p}$.

The identity is given by

$$\begin{aligned} E\left[h(\mathbf{S})\operatorname{tr}(\mathbf{T}\Sigma^{-1})\right] &= 2E\left[h(\mathbf{S})D^*\mathbf{T}_{(\frac{1}{2})}\right] + 2E\operatorname{tr}\left[\frac{\partial h(\mathbf{S})}{\partial S}\cdot\mathbf{T}_{(\frac{1}{2})}\right] \\ &\quad + (k - p - 1)E\left[h(\mathbf{S})\operatorname{tr}(\mathbf{S}^{-1}T)\right]. \end{aligned} \tag{2.4}$$

First, we apply (2.4) to risk function R_1. It happens that $\underline{\alpha}_1(\Sigma)$ has terms under the expectation of the form $h(\mathbf{S})\operatorname{tr}(\mathbf{T}\Sigma^{-1})$ with $\mathbf{T} = \mathbf{T}(\mathbf{S})$ (see the left side of (2.4)). Consequently, we readily obtain $\hat{\underline{\alpha}}_1(\Sigma)$, an unbiased estimator of $\underline{\alpha}_1(\Sigma)$. Theorem 4.3 gives conditions under which $\hat{\underline{\alpha}}_1(\Sigma) < 0\ (\forall S)$. The inequality implies $\underline{\alpha}_1(\Sigma) < 0\ (\forall\Sigma)$. Risk function R_2 introduces additional difficulties. In particular, $\alpha_2(\Sigma) \equiv R_2(\hat{\Sigma}, \Sigma) - R_2(\hat{\Sigma}_2, \Sigma)$ has terms which are quadratic in Σ^{-1} (not merely linear like $\underline{\alpha}_1$). We obtain an unbiased estimator for $\alpha_2(\Sigma)$ by a method involving repeated application of (2.4).

Various combinations of $h(\mathbf{S})$ and $T(\mathbf{S}, \Sigma)$ are used in the following. For each, the identity (2.4) is verified by straightforward modification of [3], pages 377–379. Most of these details are omitted. However, some general conditions for the validity of (2.4) are established in Haff [5].

3. Empirical Bayes estimators. In this section, we describe the empirical Bayes character of the estimators

$$\hat{\Sigma} = a[\mathbf{S} + \mathbf{u}t(\mathbf{u})C] \tag{3.1}$$

where $0 < a \leqslant 1/k$, $\mathbf{u} = 1/\text{tr}\,(\mathbf{S}^{-1}C)$, $t(\mathbf{u}) \searrow$, and $C_{p\times p} > 0$. Our dominance results are stated in Section 4. Together, these sections provide a basis for choosing a, $t(\mathbf{u})$, and C.

Assume that

$$\mathbf{S} \sim W_{p\times p}(\Sigma, k) \text{ and } \boldsymbol{\Sigma}^{-1} \sim W_{p\times p}[(1/\gamma)C^{-1}, k'] \tag{3.2}$$

with γ unknown, C p.d. and known, and $k' > 0$ a known integer. It is seen that

$$\boldsymbol{\Sigma}^{-1}|S \sim W_{p\times p}[(S + \gamma C)^{-1}, k + k'] \tag{3.3}$$

$$\text{and} \quad \boldsymbol{\Sigma}|S \sim W^{-1}_{p\times p}[S + \gamma C, (k + k') + p + 1].$$

Here $W^{-1}_{p\times p}$ indicates the inverse Wishart distribution—see Press [9], pages 109–112 for a useful description. From (3.3), the posterior mean of Σ is

$$E(\boldsymbol{\Sigma}|S, \gamma) = a(S + \gamma C), \quad a = 1/(k + k' - p - 1). \tag{3.4}$$

Our estimator in (3.1) is a modification of the posterior mean—being "empirical Bayes" in the sense that $\mathbf{u}t(\mathbf{u})$ is an estimator of γ. We now derive $\hat{\gamma} = \mathbf{u}t(\mathbf{u})$ as a generalized maximum likelihood estimator. From (3.2), the marginal density of $\mathbf{S}$ is

$$f(S|\gamma) \propto \gamma^{pk'/2}|S|^{(k-p-1)/2}|S + \gamma C|^{-(k+k')/2} \text{ for } S > 0; \tag{3.5}$$

hence, the likelihood function is

$$l(\gamma|S) = \gamma^{pk'/2}[\det(I + \gamma CS^{-1})]^{-(k+k')/2}.$$

Let

$$\begin{aligned} l^* &= \log l(\gamma|S) \\ &= \frac{pk'}{2}\log\gamma - \frac{(k+k')}{2}\log\det(I + \gamma CS^{-1}) \\ &= \frac{pk'}{2}\log\gamma - \frac{(k+k')}{2}\log\det\left(I + \gamma C^{\frac{1}{2}}S^{-1}C^{\frac{1}{2}}\right) \\ &= \frac{pk'}{2}\log\gamma - \frac{(k+k')}{2}\Sigma_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}\gamma^n\text{tr}\left(C^{\frac{1}{2}}S^{-1}C^{\frac{1}{2}}\right)^n \end{aligned}$$

[see (2.1)]. We shall proceed on a formal basis and discuss convergence later on. The first order approximation

$$l^* \doteq \frac{pk'}{2}\log\gamma - \frac{(k+k')}{2}\gamma\,\text{tr}\left(C^{\frac{1}{2}}S^{-1}C^{\frac{1}{2}}\right) \tag{3.6}$$

has a maximum at

$$\begin{aligned} \hat{\gamma} &= c^*/\text{tr}\left(C^{\frac{1}{2}}S^{-1}C^{\frac{1}{2}}\right) \\ &= c^*u \end{aligned} \tag{3.7}$$

where $c^* = pk/(k + k')$ and $u = 1/\text{tr}\,(S^{-1}C)$. Because of the approximation in (3.6), it is plausible that the constant c^* may be improved upon. Actually, functions $t(u)$ are given for which $\hat{\Sigma}$ dominates the best scalar multiple of $\mathbf{S}$. Among these are constant functions which outperform $t(u) \equiv c^*$. (See Theorem 4.3 and 4.4.)

We note that the expansion of l^* converges at any $\hat{\gamma} = ut(u)$, $0 \leqslant t(u) < 1$; i.e.,

$$\Sigma_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} (\hat{\gamma})^n \text{tr}\,\left(C^{\frac{1}{2}} S^{-1} C^{\frac{1}{2}}\right)^n = \Sigma_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} t^n(u) \text{tr} \left[\frac{C^{\frac{1}{2}} S^{-1} C^{\frac{1}{2}}}{\text{tr}\,\left(C^{\frac{1}{2}} S^{-1} C^{\frac{1}{2}}\right)} \right]^n,$$

and the series converges since $t^n(u) \searrow 0$ and the matrix inside brackets has spectral radius less than 1.

4. The main results. We now state the main results of this paper. The proofs are given in Section 5.

THEOREM 4.1. *Under loss function L_1, the best estimator of the form $a\mathbf{S}$ is* $\hat{\Sigma}_1 = (1/k)\mathbf{S}$.

THEOREM 4.2. *Under loss function L_2, the best estimator of the form $a\mathbf{S}$ is* $\hat{\Sigma}_2 = [1/(k + p + 1)]\mathbf{S}$.

Our main results concern the estimators

$$\hat{\Sigma} = a[\mathbf{S} + ut(u)C], \quad a = 1/(k + k' - p - 1). \tag{4.1}$$

For the purpose of comparing $\hat{\Sigma}$ with $\hat{\Sigma}_i$, $i = 1, 2$, we shall set k' equal to $p + 1$ and $2(p + 1)$. Note that each problem $(\Sigma, \hat{\Sigma}, L_i)$, $i = 1, 2$, is invariant under the transformations $\mathbf{S} \to A\mathbf{S}A'$, $\Sigma \to A\Sigma A'$, $\hat{\Sigma} \to A\Sigma A'$, where A is an arbitrary nonsingular matrix. In particular, let $A = C^{-\frac{1}{2}}$. Since $R_i(\hat{\Sigma}, \Sigma) = R_i(A\hat{\Sigma}A', A\Sigma A)$, we assume without loss in generality that $C = I$.

THEOREM 4.3. *In* (4.1), *let* $\hat{\Sigma}$ *be given by*

(i) $k' = p + 1$;

(ii) C *an arbitrary* p.d. *matrix; and*

(iii) $t(\mathbf{u})$ *an absolutely continuous and nonincreasing function,* $0 \leqslant t(\mathbf{u}) \leqslant 2(p - 1)/k$.

Then $\hat{\Sigma}$ *dominates* $\hat{\Sigma}_1$ (mod L_1), i.e., $R_1(\hat{\Sigma}, \Sigma) \leqslant R_1(\hat{\Sigma}_1, \Sigma)$ $(\forall \Sigma)$.

If t is a constant, then an optimal value of t is $(p - 1)/k$ (as seen from the proof).

Corollary 4.4 is useful when we have an a priori upper bound on γ, say $\bar{\gamma}$.

COROLLARY 4.4. *Assume the conditions of Theorem* 4.3 *with*

$$t(\mathbf{u}) = \bar{\gamma}/\mathbf{u} \text{ for } \mathbf{u} \geqslant k\bar{\gamma}/(p - 1) \tag{4.2}$$
$$= (p - 1)/k \text{ otherwise.}$$

Then $R_1(\hat{\Sigma}, \Sigma) \leqslant R_1(\hat{\Sigma}_1, \Sigma)$ $(\forall \Sigma)$.

According to (4.2), the estimator $\mathbf{u}t(\mathbf{u})$ never exceeds $\bar{\gamma}$. Theorems 4.5 and 4.6 concern loss function L_2.

THEOREM 4.5. *Let* $\hat{\Sigma}$ (*see equation* 4.1) *be given by*
(i) $k' = 2(p + 1)$;
(ii) C *an arbitrary* p.d. *matrix*; *and*
(iii) $0 \leqslant t \leqslant 2(p - 1)/(k - p + 3)$, t *a constant.*
Then $\hat{\Sigma}$ *dominates* $\hat{\Sigma}_2$ (mod L_2), i.e., $R_2(\hat{\Sigma}, \Sigma) \leqslant R_2(\hat{\Sigma}_2, \Sigma)$ $(\forall\Sigma)$.

The choice $t = (p - 1)/(k - p + 3)$ is optimal (again, see the proof). As mentioned before, the R_2 calculations are relatively difficult, so our present Theorem 4.5 lacks the generality of Theorem 4.3. The univariate case of Theorem 4.5 is interesting. For $p = 1$, our estimator becomes $\hat{\Sigma} = \hat{\Sigma}_2 = \mathbf{S}/(k + 2)$, the unique admissable minimax estimator of Σ (see Lehmann [7], pages 4–15).

An analogue of Corollary 4.4 is

THEOREM 4.6. *Let* $\hat{\Sigma}$ *be given by*
(i) $k' = 2(p + 1)$, $k - p - 3 > 0$;
(ii) C *an arbitrary* p.d. *matrix*; *and*
(iii) $t(\mathbf{u}) = \bar{\gamma}/\mathbf{u}$ *for* $\mathbf{u} \geqslant \bar{\gamma}(k - p + 3)/(p - 1) = (p - 1)/(k - p + 3)$ *otherwise.*
Then $R_2(\hat{\Sigma}_2, \Sigma) \leqslant R_2(\overline{\Sigma}_2, \Sigma)$ $(\forall\Sigma)$.

5. Mathematical details.

PROOF OF THEOREM 4.1. Let $\hat{\Sigma}_{1c} = (1/k)(1 + c)\mathbf{S}$, $|c| < 1$, and $\hat{\Sigma}_1 = (1/k)\mathbf{S}$. We show that

$$R_1(\hat{\Sigma}_{1c}, \Sigma) - R_1(\hat{\Sigma}, \Sigma) = E\big[(c/k)\ \mathrm{tr}\ (\mathbf{S}\Sigma^{-1}) - p \log(1 + c)\big] \geqslant 0 \quad (\forall\Sigma).$$

This inequality is true if $c - \Sigma_{n=1}^{\infty}(-1)^{n+1}(c^n/n) = \Sigma_{n=2}^{\infty}(-1)^n(c^n/n) \geqslant 0$. The latter holds because each positive term dominates its successor; and the proof is complete. □

PROOF OF THEOREM 4.2. A scalar multiple of $\mathbf{S}$ has risk $R_2(a\mathbf{S}, \Sigma) = E\ \mathrm{tr}(a\mathbf{S}\Sigma^{-1} - I)^2 = a^2\ \mathrm{tr}E(\mathbf{V}^2) - 2kpa + p$ in which $\mathbf{V} = \Sigma^{-\frac{1}{2}}\mathbf{S}\Sigma^{-\frac{1}{2}} \sim W(I, k)$. Thus $R_2(a\mathbf{S}, \Sigma) = kp(k + p + 1)a^2 - 2kpa + 2$, and the latter is minimized at $a = 1/(k + p + 1)$. □

The remaining calculations depend on special cases of (2.4). In particular, we shall need

(5.1)

(i) $E[h(\mathbf{S})\ \mathrm{tr}(Q\Sigma^{-1})] = \mathrm{tr}\ E[(k - p - 1)h(\mathbf{S})\mathbf{S}^{-1}Q + 2(\partial h(\mathbf{S})/\partial S) \cdot Q_{(\frac{1}{2})}]$ where $Q_{p\times p}$ is a matrix of constants,

(ii) $E[h(\mathbf{S})\ \mathrm{tr}(\mathbf{S}\Sigma^{-1})] = E[pkh(\mathbf{S}) + 2\ \mathrm{tr}(\partial h(\mathbf{S})/\partial S \cdot \mathbf{S}_{(\frac{1}{2})})]$ and

(iii) $E[h(\mathbf{S})\ \mathrm{tr}(\mathbf{S}^{-1}Q\Sigma^{-1})] = E\{(k - p - 2)h(\mathbf{S})\ \mathrm{tr}(\mathbf{S}^{-2}Q) - h(\mathbf{S})\ (\mathrm{tr}\ \mathbf{S}^{-1})(\mathrm{tr}\ \mathbf{S}^{-1}Q) + 2\ \mathrm{tr}[\partial h(\mathbf{S})/\partial S \cdot (\mathbf{S}^{-1}Q)_{(\frac{1}{2})}]\}$ with Q as in part (i).

PROOF OF THEOREM 4.3. Write (4.1) as $\hat{\Sigma} = (1/k)\mathbf{S} + g(\mathbf{S})I$, $g(\mathbf{S}) = (1/k)\mathbf{u}t(\mathbf{u})$, and set $\alpha_1(\Sigma) = R_1(\hat{\Sigma}, \Sigma) - R_1(\hat{\Sigma}_1, \Sigma) = E[g(\mathbf{S})\text{tr}\Sigma^{-1} - \log\det(I + kg(\mathbf{S})\mathbf{S}^{-1})]$. We must show that $\alpha_1 \leqslant 0$ $(\forall\Sigma)$. From (5.1(i)) with $Q = I$, we have

$$\alpha_1(\Sigma) = E\Big[(k - p - 1)g(\mathbf{S})\text{tr}\,\mathbf{S}^{-1} + 2\,\text{tr}\,(\partial g(\mathbf{S})/\partial S) - \Sigma_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}t^n(\mathbf{u})\text{tr}\,(\mathbf{S}^{-1}/\text{tr}\,\mathbf{S}^{-1})^n\Big].$$

This series is bounded below by a quadratic; i.e.,

$$\Sigma_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}t^n(\mathbf{u})\text{tr}\,(\mathbf{S}^{-1}/\text{tr}\,\mathbf{S}^{-1})^n \geqslant t(\mathbf{u}) - 1/2t^2(\mathbf{u}),$$

because the terms are in decreasing magnitude. A sufficient condition for $\alpha_1(\Sigma) \leqslant 0$ $(\forall\Sigma)$ is

$$(5.2)\qquad (k - p - 1)g(\mathbf{S})\text{tr}\,\mathbf{S}^{-1} + 2\text{tr}\,(\partial g(\mathbf{S})/\partial S) - t(\mathbf{u}) + 1/2\,t^2(\mathbf{u}) = -(p + 1)t(\mathbf{u}) + 2\Big[\text{tr}\frac{\partial}{\partial S}\{\mathbf{u}t(\mathbf{u})\}\Big] + (k/2)t^2(\mathbf{u}) \leqslant 0.$$

From [4], it can be seen that

$$\text{tr}\frac{\partial}{\partial S}\{\mathbf{u}t(\mathbf{u})\} = [\mathbf{u}t'(\mathbf{u}) + t(\mathbf{u})]\big[\text{tr}\,\mathbf{S}^{-2}/(\text{tr}\,\mathbf{S}^{-1})^2\big].$$

Finally, since $t'(\mathbf{u}) \leqslant 0$ and $\text{tr}\,\mathbf{S}^{-2}/(\text{tr}\,\mathbf{S}^{-1})^2 \leqslant 1$, it is easily seen that $0 \leqslant t(\mathbf{u}) \leqslant 2(p - 1)/k$ is sufficient for (5.2). □

If t is a constant, then (5.2) is bounded above by

$$-(p - 1)t + (k/2)t^2,$$

and the latter is minimized at $t = (p - 1)/k$.

Theorems 4.5 and 4.6 concern comparisons between $\hat{\Sigma} = a[\mathbf{S} + \mathbf{u}t(\mathbf{u})I]$ and $\hat{\Sigma}_2 = a\mathbf{S}$, $a = 1/(k + p + 1)$. In particular, for

$$R_2(\hat{\Sigma}, \Sigma) = R_2(\hat{\Sigma}_2, \Sigma) + \alpha_2(\Sigma)$$

$$(5.3)\qquad \text{and}$$

$$\alpha_2(\Sigma) \equiv E\big[2ag(\mathbf{S})\text{tr}\,(\mathbf{S}\Sigma^{-2}) - 2g(\mathbf{S})\text{tr}\Sigma^{-1} + g^2(\mathbf{S})\text{tr}\Sigma^{-2}\big],$$

they give conditions under which $\alpha_2(\Sigma) \leqslant 0$ $(\forall\Sigma)$.

We need the following properties of D^*(see the definition above line (2.4)):

(5.4)

(i) For a matrix $F(S)_{p\times p}$ and a scalar $\varphi(S)$,
$D^*(\varphi F) = \varphi D^*F + \text{tr}\,[(\partial\varphi(S)/\partial S)\cdot F]$;

(ii) For a matrix Q of constants,
$D^*(QS)_{(\frac{1}{2})} = \left(\frac{p+1}{2}\right)\text{tr}\,Q$;

(iii) $D^*S^{-1}_{(\frac{1}{2})} = -(\frac{1}{2})\,\text{tr}\,S^{-2} - (\frac{1}{2})(\text{tr}\,S^{-1})^2$;

(iv) $D^*S^{-2}_{(\frac{1}{2})} = -\text{tr}\,S^{-3} - (\text{tr}\,S^{-1})(\text{tr}\,S^{-2})$.

The calculation (5.4(i)) is routine, hence omitted. See the Appendix for verification of (ii) and (iv).

Identities for the terms in (5.3) are given by

LEMMA 5.1. *Assume that* $g_{ij}(\mathbf{S}) = \partial g(\mathbf{S})/\partial s_{ij}$ $(i, j, \cdots, p)$ *satisfy the conditions of Stoke's theorem—see Haff* [3] *and* [4].

(i) *For* $h(\mathbf{S}) = 2ag(\mathbf{S})$, *we have*

$$E\big[h(\mathbf{S})\ \mathrm{tr}(\mathbf{S}\Sigma^{-2})\big] = E\Big\{k(k-p-1)h(\mathbf{S})\ \mathrm{tr}\ \mathbf{S}^{-1} + 2(2k-p-1)\ \mathrm{tr}(\partial h(\mathbf{S})/\partial S) + 4D^*\big[(\partial h(\mathbf{S})/\partial S)_{(\frac{1}{2})}\cdot \mathbf{S}\big]_{(\frac{1}{2})}\Big\}.$$

(ii) *For* $h(\mathbf{S}) = -2g(\mathbf{S})$, *we have*

$$E\big[h(\mathbf{S})\ \mathrm{tr}\ \Sigma^{-1}\big] = E\big[(k-p-1)h(\mathbf{S})\ \mathrm{tr}\ \mathbf{S}^{-1} + 2\mathrm{tr}\partial h(\mathbf{S})/\partial S\big].$$

(iii) *For* $h(\mathbf{S}) = g^2(\mathbf{S})$, *we have*

$$E\big[h(\mathbf{S})\ \mathrm{tr}\Sigma^{-2}\big] = E\Big\{4D^*\big[\partial h(\mathbf{S})/\partial S\big]_{(\frac{1}{4})} + 4(k-p-1)\ \mathrm{tr}\Big[\partial h(\mathbf{S})/\partial S\cdot \mathbf{S}^{-1}_{(\frac{1}{2})}\Big] + 2(k-p-1)h(\mathbf{S})D^*\mathbf{S}^{-1}_{(\frac{1}{2})} + (k-p-1)^2h(\mathbf{S})\ \mathrm{tr}\mathbf{S}^{-2}\Big\}.$$

PROOF OF (i). Set $\mathbf{T} = \mathbf{S}\Sigma^{-1}$. From (2.4) we have

$$\begin{aligned} E\big[h(\mathbf{S})\ \mathrm{tr}(\mathbf{T}\Sigma^{-1})\big] &= E\Big\{2h(\mathbf{S})D^*T_{(\frac{1}{2})} + 2\ \mathrm{tr}\Big[\partial h(\mathbf{S})/\partial S\cdot \mathbf{T}_{(\frac{1}{2})}\Big] \\ &\qquad + (k-p-1)h(\mathbf{S})\ \mathrm{tr}\ \Sigma^{-1}\Big\} \\ &= E\Big\{kh(\mathbf{S})\ \mathrm{tr}\ \Sigma^{-1} + 2\ \mathrm{tr}\big[(\partial h(\mathbf{S})/\partial S)_{(\frac{1}{2})}\mathbf{S}\Sigma^{-1}\big]\Big\}. \end{aligned}$$

(Use (5.4(ii)) and combine terms.) From the last equation, the result follows by applying (5.1(i)) to the first term and then Lemma 5.1 to the second ($\mathbf{T} = (\partial h(\mathbf{S})/\partial S)_{(\frac{1}{2})}\cdot \mathbf{S}$).

PROOF OF (ii). In (5.1(i)), set $Q = I$.

PROOF OF (iii). Set $T = \Sigma^{-1}$. From (2.4) we have

$$E\big[h(\mathbf{S})\ \mathrm{tr}(T\Sigma^{-1})\big] = E\Big\{2\ \mathrm{tr}\big[\partial h(\mathbf{S})/\partial S\big]_{(\frac{1}{2})}\Sigma^{-1} + (k-p-1)h(\mathbf{S})\ \mathrm{tr}(\mathbf{S}^{-1}\Sigma^{-1})\Big\}.$$

Our result follows by applying (2.4) to the first term under the expectation with $\mathbf{T} = [\partial h(\mathbf{S})/\partial S]_{(\frac{1}{2})}$ and applying it to the second with $\mathbf{T} = \mathbf{S}^{-1}$.

Now we specialize Lemma 5.1 by taking $g(\mathbf{S}) = a\mathbf{u}t(\mathbf{u})$, $\mathbf{u} = 1/\ \mathrm{tr}\ (\mathbf{S}^{-1})$. The matrix

$$\partial u/\partial S = u^2S^{-2}_{(2)}$$

follows from standard perturbation results—see [4]. Thus we obtain

$$\partial g(S)/\partial S = a\big[\mathbf{u}^3t'(\mathbf{u}) + \mathbf{u}^2t(\mathbf{u})\big]S^{-2}_{(2)}. \tag{5.5}$$

Let us introduce the notation

$$\rho_m \equiv (\operatorname{tr} S^{-m})/(\operatorname{tr} S^{-1})^m, \qquad m = 1, 2, 3, \cdots. \tag{5.6}$$

Our specialization of Lemma 5.1 is

LEMMA 5.1*. *For* $g(\mathbf{S}) = a\mathbf{u}t(\mathbf{u})$ *and* $\tau(\mathbf{u}) = \mathbf{u}^3 t'(\mathbf{u}) + \mathbf{u}^2 t(\mathbf{u})$,

(i) $E[2a^2\mathbf{u}t(\mathbf{u}) \operatorname{tr}(\mathbf{S}\Sigma^{-2})] = E[2a^2k(k-p-1)\mathbf{t} - 4a^2(\tau/\mathbf{u}^2) + 4a^2(2k-p-2)(\tau/\mathbf{u}^2)\boldsymbol{\rho}_2 + 8a^2(\tau'/\mathbf{u})\boldsymbol{\rho}_3$,

(ii) $E[-2a\mathbf{u}t(\mathbf{u}) \operatorname{tr}\Sigma^{-1}] = E[-2a(k-p-1)t - 4a(\tau/\mathbf{u}^2)\boldsymbol{\rho}_2]$,

and

(iii) $Ea^2\mathbf{u}^2t^2(\mathbf{u}) = E\{[(k-p-1)^2\mathbf{t} - (k-p-1)\mathbf{t} - 8(\tau/\mathbf{u}^2)]a^2t\boldsymbol{\rho}_2 + 8(k-p-2)a^2t(\tau/\mathbf{u}^2)\boldsymbol{\rho}_3 + 8[(\tau/\mathbf{u}^2)^2 + \mathbf{t}(\tau'/\mathbf{u})]a^2\boldsymbol{\rho}_4 - (k-p-1)a^2\mathbf{t}^2\}$.

PROOF. In part (i), $h(S) = 2ag(S) = 2a^2ut(u)$. We shall compute $D^*[(\partial h(S)/\partial S)_{(\frac{1}{2})} \cdot S]_{(\frac{1}{2})}$ and leave the remaining details for the reader. Part (ii) is immediate. Also, we omit the details of part (iii). The latter calculation is lengthy; nevertheless, it is straightforward from (5.4), (5.5), and (5.6). We have

$$\begin{aligned}
D^*\big[(\partial h(S)/\partial S)_{(\frac{1}{2})} \cdot S\big]_{(\frac{1}{2})} &= 2a^2 D^*\big[(\partial ut(u)/\partial S)_{(\frac{1}{2})} \cdot S\big]_{(\frac{1}{2})} \\
&= 2a^2 D^*\Big[\big\{\tau(u)S^{-1}_{(\frac{1}{2})}\big\} \qquad \text{(from (5.5))} \\
&= 2a^2\Big\{\tau(u)D^*S^{-1}_{(\frac{1}{2})} + \operatorname{tr}\Big[\partial\tau(u)/\partial S \cdot S^{-1}_{(\frac{1}{2})}\Big]\Big\} \qquad \text{(from (5.4(i)))} \\
&= 2a^2\Big\{\tau(u)\Big[-\big(\tfrac{1}{2}\big)\operatorname{tr}S^{-2} - \big(\tfrac{1}{2}\big)(\operatorname{tr}S^{-1})^2 \\
&\quad + \tau'(u)\operatorname{tr}\Big[(\partial u/\partial S)\cdot S^{-1}_{(\frac{1}{2})}\Big]\Big\} \qquad \text{(from (5.4(iii)))} \\
&= 2a^2\Big\{\big(-\tfrac{1}{2}\big)\big[\tau(u)/u^2\big]\big[\rho_2 + 1\big] \\
&\quad + \tau'(u)\operatorname{tr}\Big[u^2 S^{-2}_{(2)}\cdot S^{-1}_{(\frac{1}{2})}\Big]\Big\} \\
&= 2a^2\Big\{\big(-\tfrac{1}{2}\big)\big[\tau(u)/u^2\big]\big[\rho_2 + 1\big] + \big[\tau'(u)/u\big]\rho_3\Big\}.
\end{aligned}$$

□

Finally, the proofs of Theorems 4.5 and 4.6 depend on a result from Bellman [1], page 137, namely,

LEMMA 5.2. *We have the inequality*

$$(1/p)^{m-1} \leqslant \rho_m \leqslant 1, \qquad m = 0, 1, 2, \cdots.$$

PROOF. Omitted. □

In addition, it is convenient to note that $\rho_m = \operatorname{tr}\,[S^{-1}/(\operatorname{tr} S^{-1})]^m$ decreases in m.

PROOF OF THEOREM 4.5. The function $\alpha_2(\Sigma)$ is given by Lemma 5.1*, being the sum of (i), (ii) and (iii). Let $t(u)$ be a constant, say t. Then $\tau(u)/u^2 = t$ and

$\tau'(u)/u = 2t$. The unbiased estimator of $\alpha_2(\Sigma)$ is

$$\begin{aligned}\hat{\alpha}_2(\Sigma) = &\{2ak[k-p-1+2\rho_2] \\ &+4a[(k-p-2)\rho_2+4\rho_3-1]-2[k-p-1+2\rho_2]\}at \\ &+\{[(k-p-1)^2-(k-p-1)-8]\rho_2+8[k-p-2]\rho_3+24\rho_4 \\ &-(k-p-1)\}a^2t^2.\end{aligned}$$

Recall that $a = 1/(k+p+1)$. The above coefficient of $at(u)$ can be written as

$$-2a[2+k(p+1)-(p+1)^2]+4a(k-2p-3)\rho_2+16a\rho_3. \tag{5.7}$$

In (5.7), the first term is negative because $k > p+1$. Thus the entire quantity is bounded above by

$$\begin{aligned}&-2a\rho_2[2+k(p+1)-(p+1)^2]+4a\rho_2(k-2p+3)+16a\rho_2 \\ &\qquad = -2a\rho_2(k-p+1)(p-1). \quad (\text{since } 0 \leqslant \rho_m \leqslant 1 \text{ and } \rho_m \searrow)\end{aligned} \tag{5.8}$$

The coefficient of a^2t^2 is bounded above by

$$\begin{aligned}&[(k-p-1)^2-(k-p-1)-8]\rho_2+8(k-p-2)\rho_2+24\rho_2 \\ &\qquad = \rho_2(k-p+3)(k-p+1).\end{aligned} \tag{5.9}$$

Finally, from (5.8) and (5.9), a sufficient condition for $\alpha_2(\Sigma) \leqslant 0$ $(\forall\Sigma)$ is

$$-2a^2\rho_2(k-p+1)(p-1)t+a^2\rho_2(k-p+3)(k-p+1)t^2 \leqslant 0 \tag{5.10}$$

which is equivalent to

$$0 \leqslant t \leqslant 2(p-1)/(k-p+3). \qquad \square$$

With respect to (5.10), the optimal t is $t = (p-1)/(k-p+3)$.

PROOF OF THEOREM 4.6. We have assumed that $0 \leqslant t \leqslant (p-1)/(k-p+3)$ and

$$\begin{aligned}t(u) &= \bar{\gamma}/u \text{ for } u \geqslant \bar{\gamma}(k-p+3)/(p-1) \\ &= (p-1)/(k-p+3) \quad \text{otherwise.}\end{aligned}$$

Recall that $\tau(u) = u^3t'(u)+u^2t(u)$. (Again, see Lemma 5.1*.) Let $\mathfrak{M} = \{u: u \geqslant \gamma(k-p+3)/(p-1)\}$. For $u \in \mathfrak{M}$, $\tau(u) = \tau'(u) = 0$. In this case, simple algebra shows that $\hat{\alpha}_2(\Sigma) \leqslant 0$ if $0 < t \leqslant (p+1)/(k-p-3)$. The latter inequality is satisfied by hypothesis. For $u \notin M$, $t = (p-1)/(k-p+3)$, and again $\hat{\alpha}_2(\Sigma) \leqslant 0$. The proof is complete. $\square$

APPENDIX

THE CALCULATIONS IN LINE (5.4).

Calculation (5.4(ii)). We show that

$$D^*(QS)_{(\frac{1}{2})} = \frac{p+1}{2}\operatorname{tr} Q.$$

Let $S = (S_1, \cdots, S_p)$ and $Q = (Q_1, \cdots, Q_p)$; thus $QS = [Q_i^t S_j]$. We have

$$\begin{aligned} D^*(QS)_{(\frac{1}{2})} &= \Sigma_i(\partial/\partial s_{ii})Q_i^t S_i + (\tfrac{1}{2})\Sigma_{i\neq j}(\partial/\partial s_{ij})Q_i^t S_j \\ &= \Sigma_i q_{ii} + (\tfrac{1}{2})\Sigma_{i\neq j} q_{ij} \\ &= (\tfrac{1}{2})\Sigma_i q_{ii} + (\tfrac{1}{2})\Sigma_{(i,j)} q_{ii} \\ &= (p+1)/2 \operatorname{tr} Q. \end{aligned}$$

Calculation (5.4(iv)). We show that

$$D^* S_{(\frac{1}{2})}^{-2} = -\operatorname{tr}(S^{-3}) - (\operatorname{tr} S^{-1})(\operatorname{tr} S^{-2}).$$

Let $S^{-1} = A = (a_{ij})$. We have

$$\begin{aligned} D^* A^2_{(\frac{1}{2})} &= \Sigma_i \Sigma_t \frac{\partial}{\partial s_{ii}} a_{it} a_{ti} + (\tfrac{1}{2})\Sigma_{i\neq j}\Sigma_t \frac{\partial}{\partial s_{ij}} a_{it} a_{tj} \\ &= \Sigma_i \Sigma_t 2a_{it}(\partial a_{it}/\partial s_{ii}) + (\tfrac{1}{2})\Sigma_{i\neq j}\Sigma_t \left[a_{it}(\partial a_{tj}/\partial s_{ij}) + a_{tj}(\partial a_{ti}/\partial s_{ij}) \right] \\ &= -\Sigma_i \Sigma_t 2a_{it} a_{ii} a_{ti} - (\tfrac{1}{2})\Sigma_{i\neq j}\Sigma_t a_{it}(a_{ti}a_{jj} + a_{tj}a_{ij}) \\ &\quad - (\tfrac{1}{2})\Sigma_{i\neq j}\Sigma_t a_{tj}(a_{ti}a_{ij} + a_{tj}a_{ii}) \\ &= -\left(\Sigma_i \Sigma_t a_{ii} a_{it}^2 + \Sigma_{i\neq j}\Sigma_t a_{ij} a_{it} a_{tj}\right) \\ &\quad - \left(\Sigma_i \Sigma_t a_{ii} a_{it}^2 + (\tfrac{1}{2})\Sigma_{i\neq j}\Sigma_t a_{ti}^2 a_{jj} + (\tfrac{1}{2})\Sigma_{i\neq j}\Sigma_t a_{tj}^2 a_{ii}\right) \\ &= -\operatorname{tr}(S^{-3}) - (\operatorname{tr} S^{-1})(\operatorname{tr} S^{-2}). \end{aligned}$$

□

REFERENCES

[1] BELLMAN, R. (1970). *Introduction to Matrix Analysis*. McGraw–Hill, New York.

[2] EFRON, B. and MORRIS, C. (1976). Multivariate Empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4** 22–32.

[3] HAFF, L. R. (1977a). Minimax estimators for a multinormal precision matrix. *J. Multivariate Anal.* **7** 374–385.

[4] HAFF, L. R. (1977b). Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *Ann. Statist.* **7** 1264–1276.

[5] HAFF, L. R. (1978). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9** (in press).

[6] JAMES W. and STEIN, C. (1961). Estimation with quadratic loss. In *Fourth Berkeley Symp. Math. Statist. Probability* Univ. California Press, Berkeley.

[7] LEHMANN, E. L. (1951). Notes on estimation. Unpublished manuscript.

[8] PERLMAN, M. D. (1972). Reduced mean square error estimation for several parameters. *Sankhyā* **34** 89–92.

[9] PRESS, S. J. (1972). *Applied Multivariate Analysis.* Holt, Rinehart and Winston, New York.

[10] SELLIAH, J. (1964). Estimation and testing problems in a Wishart distribution. Ph.D. thesis, Depart. Statist. Stanford, Univ.

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

[12] STEIN, C. (1975). Reitz lecture. 38th annual meeting IMS. Atlanta, Georgia.

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

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093