The Annals of Statistics
1986, Vol. 14, No. 4, 1625–1633

MINIMAX ESTIMATORS OF A NORMAL MEAN VECTOR FOR ARBITRARY QUADRATIC LOSS AND UNKNOWN COVARIANCE MATRIX[1]

BY LEON JAY GLESER

Purdue University

The problem of finding classes of estimators which dominate the usual estimator X of the mean vector μ of a p-variate normal distribution ($p \geq 3$) under general quadratic loss is analytically difficult in cases where the covariance matrix is unknown. Estimators of μ in this case depend upon X and an independent Wishart matrix W. In the present paper, integration-by-parts methods for both the multivariate normal and Wishart distributions are combined to yield unbiased estimates of risk difference (versus X) for certain classes of estimators, defined indirectly through a "seed" function $h(X, W)$. An application of this technique produces a new class of minimax estimators of μ.

1. Introduction. Assume that a p-dimensional ($p \geq 3$) random vector $X = (X_1, \ldots, X_p)'$ is observed which is normally distributed with mean vector μ and positive definite covariance matrix Σ. It is desired to estimate μ by an estimator δ under the quadratic loss

$$L(\delta; \mu, \Sigma) = [\mathrm{tr}(Q\Sigma)]^{-1}(\delta - \mu)'Q(\delta - \mu), \tag{1.1}$$

where Q is a known positive definite matrix and $\mathrm{tr}(A)$ stands for the trace of the matrix A.

Since Σ is assumed unknown, a random matrix W is observed along with X. It is assumed that W is statistically independent of X and has a p-dimensional Wishart distribution with parameter Σ and degrees of freedom n, $n > p + 1$. Estimators $\delta = \delta(X, W)$ of μ are evaluated in terms of their risk

$$R(\delta; \mu, \Sigma) = E[L(\delta(X, W); \mu, \Sigma)].$$

The above situation can occur, for example, when i.i.d. observations $Y_1, \ldots, Y_n$ are taken from a p-dimensional normal distribution with mean vector μ and covariance matrix Ψ, and the data are reduced by sufficiency to

$$X = N^{-1}\sum_{i=1}^{N} Y_i, \qquad W = N^{-1}\sum_{i=1}^{N}(Y_i - X)(Y_i - X)',$$

in which case $\Sigma = N^{-1}\Psi$, $n = N - 1$.

Regardless of whether or not Σ is known, and for any Q, the optimal equivariant estimator $\delta_0(X, W) = X$ of μ is minimax. However, beginning with

Received October 1985; revised January 1986.
[1]Research supported by NSF Grant DMS-8501966.
AMS 1980 *subject classifications*. Primary 62C20; secondary 62F11, 62H99, 62J07.
Key words and phrases. Integration-by-parts identities, unbiased estimates of risk difference, Wishart distribution.

the landmark paper of Stein (1956), a large body of research has been devoted to establishing broader and broader classes of estimators which dominate δ_0 in risk, often substantially. For the most part such research has concentrated upon cases where Σ is known (in which case W is not needed), or known up to a positive scalar multiple σ^2. The most successful analytic technique in these papers has been Stein's (1981) integration-by-parts identities for the normal distribution, which permit construction of unbiased estimators of risk difference (versus $\delta_0 = X$) for competing estimators.

Some attention has also been given to the case $Q = \Sigma^{-1}$, Σ unknown. However, this is a quite special situation (invariant loss), and also violates the assumption made in this paper that Q is a known matrix. Under the assumption that Σ is diagonal (or has known eigenvectors), Berger and Bock (1976, 1977) and Shinozaki (1977) have found estimators which dominate δ_0 in risk under losses (1.1), Q arbitrary.

The case of completely unknown Σ has been the most resistant to solution, even though it would clearly be of practical importance to find estimators of μ which are superior to δ_0 (and thus minimax) in such situations. Berger et al. (1977), Gleser (1979) and Berger and Haff (1983) have succeeded in developing estimators which dominate δ_0 in risk when no restrictions on Σ (or Q) are made. However, in the first two papers proof of risk domination depended upon the results of a simulation—a somewhat unsatisfactory demonstration. Berger and Haff (1983) provide a completely analytic proof of dominance, but for a fairly narrow class of estimators. Their method of attack depends upon Haff's (1977, 1979a, b, 1980) and Stein's (unpublished) integration-by-parts techniques for the Wishart distribution, but they do not obtain unbiased estimates of risk difference.

The present paper uses the integration-by-parts techniques for the normal and Wishart distributions in a new way. In Section 2, it is shown how to start with a "seed" function $h(X, W) = (h_1(X, W), \ldots, h_p(X, W))'$, and use this function to construct estimators $\delta(X, W)$ of μ having an unbiased estimator of the (weighted) risk difference

$$\operatorname{tr}(Q\Sigma)\{R(\delta; \mu, \Sigma) - R(\delta_0; \mu, \Sigma)\}$$

versus $\delta_0 = X$. In Section 3, the new method is applied to provide a completely analytic proof that a certain intuitively appealing class of estimators dominates δ_0 in risk.

2. The general method. To reduce notational complexity, in the remainder of this paper it is assumed that

$$Q = I_p, \tag{2.1}$$

where I_p is the p-dimensional identity matrix. As is verified in greater detail in Berger and Haff (1983), estimators $\delta^*(X, W)$ for the case of general Q can be obtained from estimators $\delta(X, W)$ for the case (2.1) as follows:

$$\delta^*(X, W) = (T')^{-1}[\delta(T'X, T'WT)], \tag{2.2}$$

where T is any solution of $Q = TT'$. An estimator $\delta(X, W)$ dominates $\delta_0 = X$ in risk in the case (2.1) if and only if $\delta^*(X, W)$ defined by (2.2) dominates X in risk when the loss function (1.1) is defined by general Q.

For any (scalar, vector, matrix) function $F(X, W)$, the notations

$$E_X[F(X, W)], \qquad E_W[F(X, W)],$$

respectively, denote expectation of $F(X, W)$ taken over X (with W fixed), and over W (with X fixed). When expectation jointly over both X and W is meant, no subscripts on E will be used. Since X and W are assumed independent,

$$E[F(X, W)] = E_X\{E_W[F(X, W)]\} = E_W\{E_X[F(X, W)]\}$$

provided one of the above expectations (E, $E_X E_W$, $E_W E_X$) exists.

Let $T = T(W)$ be a $p \times p$ matrix function of $W = ((w_{ij}))$. If $T = ((t_{ij}))$, define

$$D^*T_{(r)} = \sum_{i=1}^{p} \frac{\partial t_{ii}}{\partial w_{ii}} + r \sum_{i \neq j} \frac{\partial t_{ij}}{\partial w_{ij}}.$$

Under conditions on $T(W)$ specified in Haff (1979b), it can be shown [see equation (2.4) in Haff (1980)] that the following identity holds:

$$E_W[\operatorname{tr}(T\Sigma^{-1})] = 2E_W[D^*T_{(1/2)}] + (n - p - 1)E_W[\operatorname{tr}(W^{-1}T)].$$

Rewriting this equation in the form to be used in this section,

$$E_W[\operatorname{tr}(W^{-1}T)] = \frac{1}{n - p - 1}\{E_W[\operatorname{tr}(T\Sigma^{-1})] - 2E_W[D^*T_{(1/2)}]\}. \tag{2.3}$$

Let

$$h(X, W) = (h_1(X, W), h_2(X, W), \ldots, h_p(X, W))'$$

be a p-dimensional vector-valued function of X and W. Define

$$T = T(X, W) = [(X - \mu)h'(X, W)]W = (X - \mu)[Wh(X, W)]'. \tag{2.4}$$

Using the second representation of $T(X, W)$ in (2.4), it is easy to show that

$$D^*T_{(1/2)} = r'(X, W)(X - \mu), \tag{2.5}$$

where $r(X, W) = (r_1(X, W), \ldots, r_p(X, W))'$ and

$$r_i(X, W) = \frac{\partial(Wh(X, W))_i}{\partial w_{ii}} + \frac{1}{2}\sum_{j \neq i} \frac{\partial(Wh(X, W))_j}{\partial w_{ij}}, \qquad i = 1, \ldots, p. \tag{2.6}$$

Here, for any vector $U = (u_1, \ldots, u_p)$ the notation $(U)_i$ denotes u_i, the ith component of U. Alternatively, for future applications, it may be useful to note that

$$r_i(X, W) = \frac{1}{2}(p + 1)h_i(X, W) + \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p}\left[\frac{\partial h_k(X, W)}{\partial w_{ij}}\right] w_{kj}(1 + \delta_{ij}), \tag{2.6'}$$

where δ_{ij} is the Kronecker delta. This can be shown using the first representation of $T(X, W)$ in (2.4).

It now follows from (2.3), (2.4), and (2.5) that when $h(X, W)$ allows the regularity conditions underlying Haff's identity (2.3) to be met for $T(X, W)$ defined by (2.4),

$$E[h'(X,W)(X-\mu)] = E_X\{E_W[\operatorname{tr}(W^{-1}T(X,W))]\}$$

$$= \frac{1}{n-p-1}E\{\operatorname{tr}(T(X,W)\Sigma^{-1}) - 2r'(X,W)(X-\mu)\}. \tag{2.7}$$

Let

$$t(X,W) = h(X,W) + \frac{2}{n-p-1}r(X,W). \tag{2.8}$$

From (2.4) and (2.7),

$$E[t'(X,W)(X-\mu)] = \frac{1}{n-p-1}E[\operatorname{tr}(T(X,W)\Sigma^{-1})]$$

$$= \frac{1}{n-p-1}E_W\{E_X[h'(X,W)W\Sigma^{-1}(X-\mu)]\}.$$

However, if $h(X, W)$ satisfies the regularity conditions for the integration-by-parts identity for the multivariate normal distribution [see Stein (1981) and Berger and Haff (1983)],

$$E_X[h'(X,W)W\Sigma^{-1}(X-\mu)] = E_X[\operatorname{tr}(W\nabla h(X,W))], \tag{2.9}$$

where

$$\nabla h(X,W) = \left(\left(\frac{\partial h_i(X,W)}{\partial X_j}\right)\right).$$

Consequently,

$$E[t'(X,W)(X-\mu)] = \frac{1}{n-p-1}E[\operatorname{tr}(W\nabla h(X,W))]. \tag{2.10}$$

Equation (2.10) is the key result needed to prove the following theorem.

THEOREM 1. *Let $h(X, W)$ satisfy the regularity conditions needed to establish the identities* (2.7) *and* (2.9), *and let $t(X, W)$ be defined from $h(X, W)$ by* (2.6) *and* (2.8). *Define the estimator*

$$\delta(X,W) = X - t(X,W).$$

Then if $\delta(X, W)$ has finite risk,

$$\operatorname{tr}(\Sigma)[R(\delta;\mu,\Sigma) - R(\delta_0;\mu,\Sigma)] = E[M(X,W)], \tag{2.11}$$

where

$$(2.12)\quad M(X,W) = t'(X,W)t(X,W) - \left(\frac{2}{n-p-1}\right)\mathrm{tr}(W\nabla h(X,W)).$$

PROOF. First, note that for $\delta(X,W)$ to have finite risk, it is sufficient that $E[t'(X,W)t(X,W)] < \infty$. By a standard argument (remember that $Q = I_p$),

$$\begin{aligned}&\mathrm{tr}(\Sigma)[R(\delta;\mu,\Sigma) - R(\delta_0;\mu,\Sigma)]\\ &\quad = E[t'(X,W)t(X,W)] - 2E[t'(X,W)(X-\mu)].\end{aligned}$$

The assertion of the theorem is now a direct consequence of (2.10). □

Theorem 1 describes an admittedly indirect way of arriving at an estimator $\delta(X,W)$ for which one can determine an unbiased estimate of risk difference. The big advantage of this approach is the unbiased estimate of risk difference, which can simplify verification of minimaxity. The disadvantage of the approach is that one starts with one possible adjustment $X - h(X,W)$ to X, but winds up with a different adjusted estimator $X - t(X,W)$. This complicates searching for good (minimax) estimators. To apply Theorem 1 to a given estimator $\delta(X,W) = X - t(X,W)$, one must solve the set of partial differential equations defined by (2.6) and (2.8) for $h(X,W)$, and then check that $h(X,W)$ satisfies the required regularity conditions.

3. A class of minimax estimators. Let

$$(3.1)\qquad h(X,W) = \left[\frac{b(W)}{X'W^{-1}X}\right]W^{-1}X,$$

where $b(W)$ is a positive scalar function of W which is continuously differentiable as a function of the $p(p+1)/2$ free elements of W. Define the matrix

$$U(W) = \left(\left(\frac{\partial^* \log b(W)}{\partial^* w_{ij}}\right)\right),$$

where ∂^* is the symmetric partial derivative:

$$\frac{\partial^* \log b(W)}{\partial^* w_{ij}} = \begin{cases}\dfrac{\partial \log b(W)}{\partial w_{ij}}, & i = j,\\ \dfrac{1}{2}\dfrac{\partial \log b(W)}{\partial w_{ij}}, & i \neq j.\end{cases}$$

Using the fact that when $W^{-1} = ((w^{km}))$

$$\frac{\partial w^{km}}{\partial w_{ij}} = \begin{cases} -w^{ki}w^{jm} - w^{mi}w^{jk}, & i \neq j,\\ -w^{ki}w^{im}, & i = j,\end{cases}$$

it can be seen that

$$\frac{\partial}{\partial w_{ij}}(X'W^{-1}X) = (X'W^{-1})_i(X'W^{-1})_j(2 - \delta_{ij}).$$

Consequently,

$$\begin{aligned}\frac{\partial[Wh(X,W)]_j}{\partial w_{ij}} &= \frac{X_j b(W)}{(X'W^{-1}X)^2}\left\{\frac{(X'W^{-1}X)}{b(W)}\frac{\partial b(W)}{\partial w_{ij}} - \frac{\partial}{\partial w_{ij}}(X'W^{-1}X)\right\} \\ &= \frac{X_j b(W)}{(X'W^{-1}X)} \\ &\quad \times\left\{\frac{\partial \log b(W)}{\partial w_{ij}} + \frac{(2 - \delta_{ij})(X'W^{-1})_i(X'W^{-1})_j}{(X'W^{-1}X)}\right\}.\end{aligned}$$

It then follows from (2.6) and the definition of $U(W)$ that

$$r(X,W) = h(X,W) + \frac{b(W)}{X'W^{-1}X}U(W)X.$$

It is also straightforward to show that

$$\text{(3.2)} \qquad \operatorname{tr}[W\nabla h(X,W)] = \operatorname{tr}[\nabla(Wh(X,W))] = \frac{(p-2)b(W)}{X'W^{-1}X}.$$

Finally, it is not difficult (see Berger and Haff, 1983) to show that $h(X,W)$ satisfies the regularity conditions assumed in proving Theorem 1.

Hence, consider the class of estimators

$$\text{(3.3)} \qquad \delta(X,W) = X - t(X,W),$$

where

$$\begin{aligned}\text{(3.4)} \qquad t(X,W) &= \left(1 + \frac{2}{n-p-1}\right)h(X,W) + \frac{2}{n-p-1}\left[\frac{b(W)}{X'W^{-1}X}\right]U(W)X \\ &= \frac{b(W)}{(n-p-1)(X'W^{-1}X)}\{(n-p+1)W^{-1}X + 2U(W)X\}.\end{aligned}$$

By Theorem 1,

$$\text{(3.5)} \qquad \operatorname{tr}(\Sigma)[R(\delta;\mu,\Sigma) - R(\delta_0;\mu,\Sigma)] = E[M(X,W)],$$

where

$$\text{(3.6)} \quad M(X,W) = t'(X,W)t(X,W) - \left(\frac{2}{n-p-1}\right)\operatorname{tr}[W\nabla h(X,W)].$$

Since for x, y any p-dimensional column vectors ($x \neq 0$) and c_1, c_2 any two

scalars, the Cauchy–Schwarz inequality yields

$$\begin{aligned}(c_1x + c_2y)'(c_1x + c_2y) &= c_1^2x'x + 2c_1c_2x'y + c_2^2y'y \\ &\le |c_1|^2x'x + 2|c_1|\,|c_2|(x'xy'y)^{1/2} + |c_2|^2y'y \\ &= (x'x)\left[|c_1| + |c_2|\left(\frac{y'y}{x'x}\right)^{1/2}\right]^2,\end{aligned}$$

it follows directly from (3.4) that

$$\begin{aligned}t'(X,W)t(X,W) \le{}& \frac{b^2(W)X'W^{-2}X}{(n-p-1)^2(X'W^{-1}X)^2} \\ &\times\left\{(n-p+1) + 2\left[\frac{X'U'(W)U(W)X}{X'W^{-2}X}\right]^{1/2}\right\}^2.\end{aligned}$$

Assume that

$$U'(W)U(W) \le W^{-2} \tag{3.7}$$

in the ordering of positive semi-definiteness for matrices. It thus follows that

$$t'(X,W)t(X,W) \le \left[\frac{b^2(W)(n-p+3)^2}{(n-p-1)^2(X'W^{-1}X)}\right]\left[\frac{X'W^{-2}X}{X'W^{-1}X}\right]. \tag{3.8}$$

However for all X,

$$\frac{X'W^{-2}X}{X'W^{-1}X} \le \lambda_{\max}(W^{-1}) = \frac{1}{\lambda_{\min}(W)}, \tag{3.9}$$

where $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ denote the largest and smallest eigenvalues, respectively, of a symmetric matrix A. Consequently, it follows from (3.2), (3.6), (3.8), and (3.9) that

$$\begin{aligned}M(X,W) \le{}& \frac{b(W)}{(n-p-1)X'W^{-1}X} \\ &\times\left\{\frac{(n-p+3)^2b(W)}{(n-p-1)\lambda_{\min}(W)} - 2(p-2)\right\}.\end{aligned} \tag{3.10}$$

THEOREM 2. *If $b(W)$ satisfies* (3.7) *and also*

$$b(W) \le \frac{2(p-2)(n-p-1)}{(n-p+3)^2}\lambda_{\min}(W),$$

then the estimator $\delta(X,W)$ defined by (3.3) *and* (3.4) *dominates $\delta_0(X,W) = X$ in risk.*

PROOF. This is a direct consequence of (3.5) and (3.10). □

To show that the conditions of Theorem 2 are not contradictory, so that the class of estimators $\delta(X, W)$ in Theorem 2 is not empty, consider choosing

$$b(W) = c\lambda_{\min}(W), \qquad c > 0.$$

It is shown in Berger and Haff (1983) that for this choice of $b(W)$,

$$U(W) = \left(\left(\frac{\partial^* \log b(W)}{\partial^* w_{ij}}\right)\right) = \frac{1}{\lambda_{\min}(W)} gg',$$

where g is the characteristic vector of W corresponding to $\lambda_{\min}(W)$, $g'g = 1$. Hence it is easily seen that

$$U'(W)U(W) = \frac{1}{\lambda^2_{\min}(W)} gg' \leq W^{-2}.$$

Thus, when $b(W) = c\lambda_{\min}(W)$, the conditions of Theorem 2 are met when

$$c \leq \frac{2(p-2)(n-p-1)}{(n-p+3)^2}.$$

Of course, other choices of $b(W)$ are possible. For example, we can use

$$b(W) = c\left[\operatorname{tr}(W^{-1})\right]^{-1}.$$

The class of estimators covered by Theorem 2 is closely related to (subsets of) the classes of estimators considered by Gleser (1979) and Berger and Haff (1983). Indeed, the estimators discussed in Theorem 2 can be regarded as adjustments to special cases of estimators considered by these authors. Results concerning minimaxity of adjustments to the more general estimators considered by Gleser (1979) and Berger and Haff (1983) can be established using the methods of Section 2. One can also consider the minimaxity of adjustments to estimators of the form

$$X - \frac{b(W)}{X'W^{-2}X} W^{-1}X,$$

although the analysis is more complicated, and the resulting adjusted estimators are less attractive in form. However, the purpose here has been to illustrate application of the methods of Section 2. A comparison of the analysis and results here to the arguments and results in Berger et al. (1977), Gleser (1979) or Berger and Haff (1983) should give convincing evidence of the usefulness and relative simplicity of the methods of Section 2.

Acknowledgments. I am grateful to the Associate Editor and a referee for helpful comments that strengthened the exposition and conclusions in this paper.

REFERENCES

BERGER, J. and BOCK, M. E. (1976). Combining independent normal mean estimation problems with unknown variances. *Ann. Statist.* 4 642–648.

BERGER, J. and BOCK, M. E. (1977). Improved minimax estimators of normal mean vectors for certain types of covariance matrices. In *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. S. Moore, eds.) 19–36. Academic, New York.

BERGER, J., BOCK, M. E., BROWN, L. D., CASELLA, G. and GLESER, L. J. (1977). Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Ann. Statist.* **5** 763–771.

BERGER, J. and HAFF, L. R. (1983). A class of minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Statistics and Decisions* **1** 105–129.

GLESER, L. J. (1979). Minimax estimation of a normal mean vector when the covariance matrix is unknown. *Ann. Statist.* **7** 838–846.

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

HAFF, L. R. (1979a). Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *Ann. Statist.* **7** 1264–1276.

HAFF, L. R. (1979b). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9** 531–542.

HAFF, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8** 586–597.

SHINOZAKI, N. (1977). Simultaneous estimation of the means of independent variables with unknown variances. *Keio Math. Sem. Rep.* **2** 75–79.

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

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

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907