IMPROVING THE USUAL ESTIMATOR OF
A NORMAL COVARIANCE MATRIX

BY

CHARLES STEIN, BRADLEY EFRON and CARL MORRIS

TECHNICAL REPORT NO. 37
MARCH 30, 1972

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GP-30711X

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
IMPROVING THE USUAL ESTIMATOR OF
A NORMAL COVARIANCE MATRIX

by

Charles Stein*,
Bradley Efron*,
and Carl Morris**

TECHNICAL REPORT NO. 37
March 30, 1972

Prepared Under the Auspices of
National Science Foundation Grant GP-30711X

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

* Stanford University, Stanford, California
** Santa Monica, California
IMPROVING THE USUAL ESTIMATOR OF
A NORMAL COVARIANCE MATRIX

0. Summary.

Given \( k \) independent \( p \)-dimensional multivariate normal vectors, \( X_1, X_2, \ldots, X_k \), each having mean \( 0 \) and covariance matrix \( \Sigma \),

\[ X_i \overset{\text{ind}}{\sim} N_p(0, \Sigma) \]  \hspace{1cm} (1)

a sufficient statistic for estimating \( \Sigma \) is \( S = XX' \), where \( X \) is the \( p \times k \) matrix \((X_1, X_2, \ldots, X_k)\). The \( p \times p \) matrix \( S \) has the Wishart distribution

\[ S \sim W(\Sigma, k, p) \]  \hspace{1cm} (2)

The usual estimator of \( \Sigma \) is some multiple of \( S \), most often the unbiased estimator \( S/k \). However if we know that \( \Sigma \) has the special form \( \Sigma = \sigma I \), \( \sigma \) some unknown constant, a better unbiased estimator is \((\text{tr } S/p) I\). Both of these estimators are of the
form \([aS^{-1} + (b/\text{tr } S)I]^{-1}\). In this note we will show that if \(k > p + 1\), and the constants \(a\) and \(b\) are chosen to be \(a^* = k - p - 1,\ \ b^* = p(p + 1) - 2\), then the resulting estimator dominates any multiple of \(S\) under a natural loss function for which the motivation is given in Section 1. (In Section 5 of [3] Stein proves this same result, however using loss functions which he describes as making the problem somewhat artificial.)
1. A Natural Loss Function

We will use the loss function

\[ L(\hat{\mathbf{Z}}, \mathbf{Z}) = \frac{\text{tr}(\hat{\mathbf{Z}}^{-1} - \mathbf{Z}^{-1})S(\hat{\mathbf{Z}}^{-1} - \mathbf{Z}^{-1})}{k \text{ tr } \hat{\mathbf{Z}}^{-1}} \] (3)

to express the loss incurred by estimating \( \mathbf{Z} \) to be \( \hat{\mathbf{Z}}(\mathbf{S}) \). (This requires \( \mathbf{Z} \) to be non-singular, which for convenience we will always assume to be the case.)

Expression (3) is motivated by a Bayesian estimation problem: suppose \( \theta_1, \theta_2, \ldots, \theta_k \) are independent \( p \)-dimensional vectors each distributed as \( N_p(0, A) \). For each value of \( i, \ i = 1, 2, \ldots, k \), we observe \( X_i \sim N_p(\theta_i, I) \), the \( X_i \) independent of each other, and we wish to estimate the \( p \times k \) matrix \( \theta = (\theta_1, \theta_2, \ldots, \theta_k) \), with normalized squared error loss function

\[ l(\theta, \hat{\theta}) = \frac{1}{pk} \sum_{i=1}^{k} \sum_{j=1}^{p} (\hat{\theta}_{ij} - \theta_{ij})^2 \] (4)

If we know the \( p \times p \) matrix \( A \) we can use the Bayes estimator \( \hat{\theta}_i = (I - (A + I)^{-1})X_i, \ i = 1, 2, \ldots, k \), which has a Bayes risk of \( R^* = E_l(\theta, \theta^*) = 1 - \frac{1}{p} \text{ tr}(A + I)^{-1} \). This should be compared with the risk of the maximum likelihood estimator \( \hat{\theta}_i^0 = X_i, \ i = 1, 2, \ldots, k \), \( R^0 = E_l(\theta, \theta^0) = 1 \). However, even if \( A \) is unknown, we can estimate
it from the $X_i$, which are marginally independent with distribution $\mathcal{N}_p(0, A + I)$. Let $\hat{z} = A + I$, and let $\hat{Z}(S)$ be any function of $S = XX'$. It is shown in [1] that the rule $\hat{\theta}_i = (I - \hat{Z}^{-1})X_i$, $i = 1, 2, \ldots, k$, has risk $R = EL(\theta, \hat{\theta})$ satisfying

$$\frac{\hat{R} - R^*}{R^0 - R^*} = EL(\hat{z}, \hat{z})$$

(5)

where $L$ is given by (3). The left-hand side of (5) is a convenient measure of how much it costs to estimate $\hat{z}$ if it is not known, and the equality justifies the choice of the loss function $L$. 
2. An Estimator Which Dominates Any Multiple of $S$

A simple calculation, given in [1], shows that choosing
\[ \hat{\theta} = S/(k - p - 1) \]
gives \( EL(\hat{\theta}, \hat{\theta}) = \frac{p+1}{k} \) for every non-singular \( \hat{\theta} \), and that this choice uniformly dominates any other multiple of \( S \).

We now consider estimators of the form
\[ \hat{\theta}(S) = [aS^{-1} + (b/\text{tr} S)I]^{-1} \quad (6) \]

**Theorem 1.** The estimator (6) with \( a = a^* = k - p - 1 \) and \( b = b^* = p(p + 1) - 2 \) has risk
\[ EL(\hat{\theta}, \hat{\theta}) = \frac{p+1}{k} \left[ \frac{p(p + 1) - 2}{k(pk - 2)} \right]^2 \frac{E\frac{\text{tr}^{-1}S}{\text{tr}^{-1}\text{tr} S}} {\text{tr}^{-1}\text{tr} S} \quad (7) \]

Notice that in the special case \( \hat{\theta} = cI \) expression (7) becomes
\[ EL(\hat{\theta}, \hat{\theta}) = \frac{p+1}{k} \left[ \frac{p(p + 1) - 2}{pk(pk - 2)} \right]^2 \quad (8) \]

**Proof.** Because the loss function (3) and the rules (6) are invariant under orthogonal transformations, there is no loss of generality in assuming that \( \hat{\theta} \) is diagonal, say with diagonal elements \( c_1, c_2, \ldots, c_p \). We need the relationship of two quantities for the calculations which follow, \( E\frac{1}{\text{tr} S} \) and \( E\frac{\text{tr}^{-1}S}{\text{tr} S} \). The choice
of \( \hat{\Sigma} \) diagonal gives

\[
E \frac{\text{tr} \hat{\Sigma}^{-1} S}{\text{tr} S} = E \sum_{j=1}^{p} \frac{W_j}{\sum_{j=1}^{p} \sigma_j W_j} = \tau \quad \text{(say)} \tag{9}
\]

where the \( W_j \) are independent each \( \chi_k^2 \), and also

\[
E \frac{1}{\text{tr} S} = E \frac{1}{\sum_{j=1}^{p} \sum_{j=1}^{p} \sigma_j W_j} = E \sum_{j=1}^{p} \frac{W_j}{\sum_{j=1}^{p} W_j} = \frac{\tau}{kp-2} \tag{10}
\]

the last equality following since \( \frac{\sum_{j=1}^{p} W_j}{\sum_{j=1}^{p} \sigma_j W_j} \) is independent of

\( \sum_{j=1}^{p} W_j \sim \chi_{kp}^2 \).

We can now evaluate the numerator of (3) for any choice of \( a \) and \( b \) in (6):

\[
E \text{tr}(\hat{\Sigma}^{-1} - \Sigma^{-1}) S(\hat{\Sigma}^{-1} - \Sigma^{-1}) = E \text{tr}[a^2\Sigma^{-1} + \frac{2ab}{\text{tr} S} I + \frac{b^2 S}{\text{tr}^2 S} - 2 \hat{\Sigma}^{-1}(aI + \frac{b S}{\text{tr} S}) + \Sigma^{-1} \Sigma^{-1}] \tag{11}
\]

Using (9) and (10), and also \( ES = k\hat{\Sigma}, \ ES^{-1} = \hat{\Sigma}^{-1}/(k-p-1) \), reduces (11) to

\[
\frac{a^2}{k-p-1} \frac{1}{\sum_{j=1}^{p} \sigma_j} + \frac{2ab}{kp-2} + \frac{b^2}{kp-2} - 2a \sum_{j=1}^{p} \frac{1}{\sigma_j} - 2bx + k \sum_{j=1}^{p} \frac{1}{\sigma_j} \tag{12}
\]
\[ a^2 \left[ \frac{a}{k-p-1} - 2a + k \right] \sum_{j=1}^{P} \frac{1}{\sigma_j} + \left[ \frac{2ap+b}{kp-2} - 2 \right] b \]  \hspace{1cm} (13)

Substituting in \( a = k - p - 1 \) and \( b = p(p + 1) - 2 \), and dividing by \( k \) \( \text{tr} \ \hat{\Sigma}^{-1} = k \sum_{j=1}^{P} \frac{1}{\sigma_j} \) gives the theorem.

The "best" values of \( a \) and \( b \) naturally depend on the parameter \( \xi \). Nevertheless it is interesting to calculate them by differentiating (13) with respect to \( a \) and \( b \) and setting the results equal to zero:

\[ \left( \sum_{j=1}^{P} \frac{1}{\sigma_j(k-p-1)} \right) \dot{a} + \left( \frac{p}{p(kp-2)} \right) \dot{b} = \sum_{j=1}^{P} \frac{1}{\sigma_j} \]  \hspace{1cm} (14)

\[ \dot{a} + \dot{b} = kp-2 \]  \hspace{1cm} (15)

Solving for \( \dot{a} \) gives

\[ \dot{a} = a^* \left[ 1 - \frac{pb^*}{(kp-2)(\sum_{j=1}^{P} \frac{1}{\sigma_j}) - p^2(k-p-1)} \right] \]  \hspace{1cm} (16)

In the special case where all the \( \sigma_j \) are equal, say to \( \sigma \), then \( \sum_{j=1}^{P} \frac{1}{\sigma_j} = p \) and formula (16) gives \( \dot{a} = 0 \), hence \( \dot{b} = kp - 2 \) by (15). At the other extreme, if one \( \sigma_j \) is finite and at least one \( \sigma_j \) is infinite then \( \dot{a} = a^* \), \( \dot{b} = b^* \). For all other cases \( \dot{a} \in (0, a^*), \dot{b} \in (b^*, kp - 2) \). The choice of the constants \( a = a^* \), \( b = b^* \) is conservative in the sense that for any \( \xi \) a smaller \( a \)
and larger $b$ are optimal.

(To show that $\mathbb{E}(0, a^*)$ for all $\mathcal{F}$ it is equivalent to show that $\tau / \sum_{j=1}^{P} \sigma_j e(0, \frac{1}{P})$. Writing $\tau = \mathbb{E} \frac{1}{\sum_{j=1}^{P} \sigma_j D_j}$ where $D_j = W_j / (\sum_{j=1}^{P} W_j)$,

Jensen's inequality gives $1 / \sum_{j=1}^{P} \sigma_j D_j \leq \sum_{j=1}^{P} D_j (1/\sigma_j)$ for every choice of the $D_j$ all positive and summing to one. Taking expectations on both sides gives the desired inequality.)
3. **Using the Estimator in a Simultaneous Estimation Problem**

Suppose we wish to estimate the \( p \times k \) matrix \( \theta = (\theta_1, \theta_2, \ldots, \theta_k) \) given the data \( X = (X_1, X_2, \ldots, X_k) \) using the loss function (4), where \( x_i \sim N(\theta_i, I) \) independently for \( i = 1, 2, \ldots, k \). This is the problem discussed in section 1, except that we have dropped the Bayesian assumption that \( \theta_i \sim N(0, A) \). Nevertheless we will consider the estimation rule motivated by the Bayesian assumption,

\[
\hat{\theta}_i = (I - [a^\# S^{-1} - (b^\# / \text{tr} \ S)I])X_i \quad i = 1, 2, \ldots, k \quad (17)
\]

where \( S = XX' \) as before.

The following theorem shows that this rule dominates the maximum likelihood estimator and also the estimation rule (0.10) considered in [1] in the strong sense that the risk function \( R(\theta, \hat{\theta}) \equiv E_{\theta}(\theta, \hat{\theta}) \) is smaller for every value of \( \theta \).

**Theorem 2.** The estimation rule (17) has risk function

\[
R(\theta, \hat{\theta}) = 1 - \frac{(k-p-1)^2}{pk} E_{\theta} \text{tr} \ S^{-1} - \frac{[p(p+1)-2]^2}{pk} E_{\theta} \frac{1}{\text{tr} \ S} \quad (18)
\]

We have written "\( E_{\theta} \)" to indicate that the expectation is taken with the \( \theta \) matrix fixed, rather than random as in section 1. We will show below that if we let \( \rho = \sqrt{\text{tr} \ \theta'^\theta} \) then

\[
\frac{1}{(pk-2)\rho^2} \leq E_{\theta} \frac{1}{\text{tr} \ S} \leq \frac{1}{(pk-2) + \frac{pk-2}{pk} \rho^2} \quad (19)
\]
This indicates the amount by which (18) is less than 
\[ 1 - \frac{(k-p-1)^2}{pk} \mathbb{E}_\theta \text{tr } S^{-1}, \] 
the risk function of the rule (0.10)
considered in [1].

The proof of theorem 2 proceeds exactly as in section 2 of [1],
and we will only sketch it here. The method is to take the expectation
of the expression for \( R(\hat{\theta}, \theta) \) in (18) with respect to the distribution
of \( \theta \) given by the assumption that

\[ \theta_i \sim N_p(0, A) \]  \hspace{0.5cm} (20)

independently for \( i = 1, 2, \ldots, k \), and to show that this quantity,
which is identical with what was called \( \bar{R} \) in section 1, gives

\[ \frac{\hat{R} - R^*}{R^0 - R^*} = \frac{p+1}{k} - \frac{[p(p+1)-2]^2}{k(pk-2)} \mathbb{E} \frac{\text{tr}(A+I)^{-1} S}{\text{tr}(A+I)^{-1} \text{tr } S} \]  \hspace{0.5cm} (21)

as must be the case by (5) and (7). Simple calculations show that
this will be the case if

\[ \mathbb{E} \frac{pk-2}{\text{tr } S} = \mathbb{E} \frac{\text{tr}(A+I)^{-1} S}{\text{tr } S} \]  \hspace{0.5cm} (22)

(The symbol "\( \mathbb{E} \)" without the subscript \( \theta \) indicates expectation over
\( \theta \), as distributed in (20), as well as over \( X \).) However (22)
is equivalent to (10), which proves the theorem.

The bound (19) follows by noticing that, given \( \theta \), \( \text{tr } S \) has a noncentral chi-square distribution, \( \text{tr } S \sim \chi_{pk}^2(\rho^2) \), and therefore can be written as a Poisson mixture of central chi-squares as in [5], say \( \text{tr } S \sim \chi_{pk+2J}^2 \), \( J \) Poisson with parameter \( \rho^2/2 \). Letting \( E_\rho \) indicate expectation with respect to the Poisson distribution,

\[
E_\theta \frac{1}{\text{tr } S} = E_\rho \frac{1}{pk+2J-2}
\]

and the left hand side of (19) is given by Jensen's inequality. To obtain the right hand inequality write \( E_\rho \frac{1}{pk+2J-2} \) as

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

and notice that this can also be expressed as \( \frac{1}{pk-2} [1 - \rho^2 E_\rho \frac{1}{pk+2J}] \). By Jensen's inequality \( E_\rho \frac{1}{pk+2J} \geq \frac{1}{pk+\rho^2} \), giving the result.

(Note: In the case \( p = 1 \) the last term in (16) is zero, the correct value since, in this case, the rule (17) coincides with that given in [1].)
4. A Wider Class of Estimators

All of the estimators considered in this paper are of the form

\[ \hat{z}(S) = G\hat{\sigma}G' \]  

(29)

when \( G \) is the matrix of eigenvectors of \( S \), say \( S = GDG' \), \( D \) diagonal, and \( \hat{\sigma} \) is a diagonal matrix whose entries are functions of the eigenvalues of \( S \), \( \hat{\sigma} = \hat{\sigma}(D) \). Explicitly the best linear multiple of \( S \), \( \hat{z}(S) = S/(k-p-1) \), estimates the \( i^{th} \) eigenvalue of \( z \) by \( \hat{\lambda}_i^{(1)} = d_i/(k-p-1) \), while the estimator (6) uses

\[ \hat{\lambda}_i^{(2)} = \frac{1}{1 + \frac{b^*}{a^*} \sum_{j=1}^{p} d_j} \frac{d_i}{k-p-1} \]  

(30)

We see that \( \hat{\lambda}_i^{(2)}/\hat{\lambda}_i^{(1)} = 1/[1 + (b^*/a^*)(d_i/\sum_{j=1}^{p} d_j)] \), so that the estimator (6) improves on the best linear estimator by shrinking all of the estimated eigenvalues toward zero, the larger eigenvalues being shrunk more than the smaller. This is reminiscent of the James-Stein estimator for the mean of a multivariate normal vector [3], and the basic phenomenon seems to be the same: the eigenvalues of \( S \), considered as an ensemble of \( p \) numbers, are distorted in a systematic nonlinear way from the eigenvalues of \( z \). By undoing this distortion with a given transformation we obtain a universally improved estimator.
The transformation (30) used in this paper is actually a very conservative choice, as remarked in the proof of Theorem 1. A more ambitious estimation scheme would estimate the parameter \[ \frac{1}{n} \sum_{j=1}^{P} \sigma_j^{1/\tau} \] from the data and then use \( \hat{a} \) and \( \hat{b} \) as given by (15) and (16) in (b). Of course there is no reason to restrict attention to the form (6). For example another approach would be to perform a preliminary hypothesis test and to estimate \( \sigma_j \) by zero whenever the corresponding \( d_j \) was not significantly larger than zero. (This approach is used by Mandel [4] and Collob [2] for the simultaneous estimation problem discussed in section 3. Writing the data matrix \( X \) in the singular decomposition \( X = GD^{1/2}H \), where \( G \) and \( H \) are respectively the eigenvectors of \( XX' \) and \( X'X \), and \( D \) is the diagonal matrix of non-zero eigenvalues of \( S \), they replace the maximum likelihood estimator \( \hat{\Theta} = X \) by \( \tilde{\Theta} = GD^{1/2}H \), where \( \tilde{D} \) is \( D \) with nonsignificant values of \( d_j \) set equal to zero.) Although calculations are difficult and have not been carried out by the authors, it seems likely that estimators such as these could yield considerable improvement over the standard estimators of \( \Sigma \).
REFERENCES


