ESTIMATION OF A COVARIANCE MATRIX USING THE REFERENCE PRIOR*

by

Ruoyong Yang and James O. Berger

Technical Report #93-13C

Department of Statistics
Purdue University

February 1993

Research was supported by NSF grant DMS-8923071 at Purdue University.

*

Estimation of a Covariance Matrix Using the Reference Prior *

Ruoyong Yang and James O. Berger

Department of Statistics, Purdue University

February, 1993

Abstract

Estimation of a covariance matrix, Σ, is a notoriously difficult problem; the standard unbiased estimator can be substantially suboptimal. We approach the problem from a noninformative prior Bayesian perspective, developing the *reference* noninformative prior for a covariance matrix, and obtaining expressions for the resulting Bayes estimators. These expressions involve the computation of high-dimensional posterior expectations, which is done using a recent Markov chain simulation tool, the *hit-and-run sampler*. Frequentist risk comparisons with previously suggested estimators are also given, and determination of the accuracy of the estimators is addressed.

AMS 1991 subject classifications. Primary 62C10; secondary 62F15, 62H12.
Key words and phrases. Jeffreys prior, reference prior, covariance matrix, information matrix, Markov chain simulation, hit-and-run sampler, entropy loss, quadratic loss, risk.

*This research was supported by NSF grant DMS-8923071 at Purdue University.

1 Introduction

Suppose that $\vec{X}_1, \ldots, \vec{X}_n$ are i.i.d. $N_p(0,\ \Sigma)$, and consider the problem of estimating the $p \times p$ positive definite Σ under the losses

$$L_1(\hat{\Sigma},\ \Sigma) = \mathrm{tr}(\hat{\Sigma}\Sigma^{-1}) - \log|\hat{\Sigma}\Sigma^{-1}| - p, \tag{1}$$

$$L_2(\hat{\Sigma},\ \Sigma) = \mathrm{tr}(\hat{\Sigma}\Sigma^{-1} - I)^2, \tag{2}$$

where $\hat{\Sigma}$ denotes an arbitrary estimator. The first loss was advocated by Stein (1956) and is usually called entropy loss while the second is typically called quadratic loss. The corresponding frequentist risk functions will be denoted by

$$R_i(\hat{\Sigma},\ \Sigma) = E_\Sigma L_i(\hat{\Sigma}, \Sigma), \quad i = 1, 2. \tag{3}$$

Analogous losses and risks can be defined for the problem of estimating Σ^{-1}; see Section 3.1.

The usual (unbiased) estimator of Σ is the sample covariance matrix

$$\frac{1}{n}S = \frac{1}{n}\sum_{i=1}^{n} \vec{X}_i \vec{X}_i^t \sim \frac{1}{n}W_p(\Sigma,\ n), \tag{4}$$

where $W_p(\Sigma,\ n)$ is the Wishart distribution with scale matrix Σ and n degrees of freedom. This estimator, and $S/(n+p+1)$, are the best scalar multiples of S for L_1 and L_2, respectively (see, e.g., Haff, 1980). It was, however, pointed out by Stein (1956, 1975) and Dempster (1969) that the eigenstructure of Σ tends to be systematically distorted by these estimators unless p/n is quite small. The problem is especially bad when $\Sigma \cong I$. Starting with Stein's Rietz lecture (1975), several major efforts have been made to overcome this distortion. The literature includes Stein (1975, 1977a, 1977b), Efron and Morris (1976), Haff (1977, 1979a, 1979b, 1980, 1991), Olkin and Selliah (1977), Sharma (1980), Sugiura and Fujimoto (1982), Sharma and Krishnamoorthy (1983, 1985a, 1985b), Takemura (1984), Dey and Srinivasan (1985, 1986), Lin and Perlman (1985), Dey (1988), Loh (1991a, 1991b) and Perron (1992). Note that dramatic gains in risk are achievable.

Simulation studies (cf., Lin and Perlman, 1985, and Haff, 1991) seem to suggest that the estimators of Stein (1975) and Haff (1991) are particularly successful in adequately "shrinking" the eigenstructure of S. Both estimators are approximately Bayes (especially that of Haff) but require incorporation of an isotonizing step in their computation to avoid overshrinkage of certain eigenvalues. Also, no approach to shrinkage estimation of Σ has produced reportable measures of the accuracy of $\hat{\Sigma}$. This is a serious limitation.

Because of the centrality in statistics of the covariance matrix estimation problem and because of the limitations of the existing estimation methods, it seemed desirable to attempt a fully Bayesian approach to the problem based on use of reference (noninformative) priors. These priors seem to be remarkably successful in many multivariate problems in producing estimators with simultaneously good Bayesian and frequentist properties (cf., Berger and Bernardo, 1992a, b, c, and Ye and Berger, 1991). Also, they tend to yield very satisfactory measures of accuracy, through the posterior covariance, or posterior expected loss.

Section 2 contains the development of the reference prior for this problem. Rather surprisingly, the reference prior turns out to be remarkably simple. Indeed, it is the prior proposed by Chang and Eaves (1990), which was based on the simpler (but less satisfactory) reference prior algorithm in Bernardo (1979). Not unexpectedly, however, computation with this prior is not possible in closed form; thus Section 3 develops an efficient computational scheme. Section 4 compares the reference prior Bayes estimator to the estimators of Stein (1975) and Haff (1991). Section 5 discusses determination of the accuracy of $\hat{\Sigma}$.

Note that there have been previous partial Bayesian approaches to estimation of Σ. These include the empirical Bayes analyses of Efron and Morris (1976) and Haff (1980). Conjugate priors have also been used (cf., Press, 1982), but these do not achieve the type of eigenvalue shrinkage that seems most desirable. A flexible and very appealing general class of prior distributions for Σ has recently been introduced by Leonard and Hsu (1992). Their approach allows for a wide variety of subjective shrinkage patterns, but it is not clear if the shrinkage pattern we seek can be reproduced in this way.

The common noninformative prior for the problem has been the Jeffreys prior

$$\pi(\Sigma) = (\det \Sigma)^{-(p+1)/2} d\Sigma. \tag{5}$$

This prior was developed by Jeffreys (1961), for $p = 1, 2$, and by Geisser and Cornfield (1963), Geisser (1965) and Villegas (1969) for arbitrary p. Use of the Jeffreys prior tends to simply reproduce classical answers, however, and hence also fails to appropriately shrink the eigenvalues.

The work most closely related to this study is that of Haff (1991), which proposes an estimator based on a variational form of the Bayes estimator. In the derivation of Haff's estimator, however, a term in the expression for the Bayes estimator is (purposely) ignored, so that it is unclear if the result actually corresponds to a Bayes rule or what the implied prior distribution might be. We do, however, observe considerable similarity between our estimator and that of Haff.

2 The Reference Prior for a Covariance Matrix

2.1 The Fisher Information for a Covariance Matrix

We will use the following notation. The entries of a matrix A will be denoted by $A_{[i,j]}$, and A^t, $|A|$ and $\mathrm{tr}(A)$ will denote the transpose, determinant and trace of a square matrix, A, respectively. Denote the matrix operator which arranges the columns of a matrix into one long column as vec(). The Kronecker product of two matrices, A and B, will be denoted by $A \otimes B$. The covariance matrix Σ can be decomposed as $\Sigma = O^t D O$ with O an orthogonal matrix with positive elements for the first row and D a diagonal matrix, $D = \mathrm{diag}(d_1, \ldots, d_p)$, with $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. Write $O = (O_{12}O_{13} \cdots O_{1p})(O_{23} \cdots O_{2p}) \cdots (O_{p-1,p})D_\epsilon$, with O_{ij} being a simple orthogonal matrix such as

$$O_{ij} = O_{ij}(o_{ij}) = \begin{array}{c|ccccc|} & & i & & j & \\ & I & 0 & 0 & 0 & 0 \\ i & 0 & \cos o_{ij} & 0 & -\sin o_{ij} & 0 \\ & 0 & 0 & I & 0 & 0 \\ j & 0 & \sin o_{ij} & 0 & \cos o_{ij} & 0 \\ & 0 & 0 & 0 & 0 & I \end{array}\,, \tag{6}$$

where $-\pi/2 < o_{ij} \leq \pi/2$, and D_ϵ being a diagonal matrix with diagonal elements ± 1 (see Anderson, Olkin, and Underhill, 1987). Let $(d\Sigma)$ denote $\prod_{i \leq j} d\sigma_{ij}$, (dD) denote $\prod_{i=1}^{p} dd_i$, (dO) denote $\prod_{i<j} do_{ij}$, and (dH) denote the *conditional invariant Haar measure* over the space of orthogonal matrices $\mathcal{O} = \{O : O^t O = I\}$ (see Anderson, 1984, for definition).

Lemma 1 . *The Fisher information matrix for Σ, w.r.t the reparameterization $(D,\ O)$, is of the form*

$$I(D,\ O) = \begin{pmatrix} I(D) & 0 \\ 0 & I(O) \end{pmatrix}, \tag{7}$$

with $I(D) = \mathrm{diag}(1/2d_1^2, \ldots, 1/2d_p^2)$. *(*Note that the explicit form of $I(O)$ will not be needed.*)*

Proof. See Appendix A. □

Lemma 2 . *(i) The determinant of the Fisher information matrix of Σ is*

$$|I(\Sigma)| \propto |\Sigma|^{-(p+1)}. \tag{8}$$

(ii) The relationship between the Fisher information matrix w.r.t the parameter Σ and $(D,\ O)$ is

$$I(D,\ O) = \left[\frac{\partial(\Sigma)}{\partial(D,\ O)}\right]^t I(\Sigma) \left[\frac{\partial(\Sigma)}{\partial(D,\ O)}\right] \tag{9}$$

and

$$\begin{aligned} (d\Sigma) &\propto \left[\prod_{i<j}(d_i - d_j)\right] \quad (dD)(dH) & (10) \\ &\propto \left[\prod_{i=1}^{p-1}\prod_{j=i+1}^{p} \cos^{j-i-1} o_{ij}\right] \left[\prod_{i<j}(d_i - d_j)\right] \quad (dD)(DO). & (11) \end{aligned}$$

(iii) The determinant of $I(O)$ in Lemma 1 is

$$|I(O)| \propto \left[\prod_{i=1}^{p-1}\prod_{j=i+1}^{p} \cos^{j-i-1} o_{ij}\right]^2 \left[\prod_{i<j}(d_i - d_j)\right]^2 \prod_{i=1}^{p} d_i^{-(p-1)}. \tag{12}$$

Proof. Equation (9) is trivial; for (8), see Press (1982, p.79); for (10), see Farrell (1985, p.74); and for (11), see Anderson, Olkin and Underhill (1987). Part (iii) follows from the representation $\Sigma = O^t DO$. □

2.2 The Reference Prior

Bernardo (1979) initiated an information based approach to development of noninformative priors, called the reference prior approach. A review and discussion of the current status of the approach can be found in Berger and Bernardo (1992c). The motivation for developing the approach was the acknowledged problems of the Jeffreys prior in higher dimensions. Even Jeffreys would often alter Jeffreys prior in multiparameter problems to remove perceived inadequacies. The reference prior approach seeks to overcome these difficulties by breaking up multiparameter problems into a series of conditional one-parameter problems, for which reasonable noninformative priors can be determined. The approach has proven to be remarkably successful in overcoming the inadequacies of Jeffreys prior in multiparameter problems (cf., Berger and Bernardo, 1989, 1992a,b,c, Ye and Berger, 1991).

In the following theorem, the reference prior for Σ is given. The Jeffreys prior is also given for comparison purposes. Note that the reference prior can depend on what is called the "group ordering," which is typically simply a listing of parameters according to perceived "importance." Note, also, that the reference prior here was first given in Chang and Eaves (1990), although their derivation utilized the early version of the reference prior algorithm in Bernardo (1979), which was improved in Berger and Bernardo (1992a, b, c).

Theorem 1 . *The reference prior for the parameter $(D,\ O)$ is as follows, providing the group ordering used lists D before O and the $\{d_i\}$ are ordered monotonically (either increasing or decreasing):*

$$\begin{aligned} \pi_R(D,\ O)\ \ (dD)(dO) &\propto \left[\prod_{i=1}^{p-1}\prod_{j=i+1}^{p} \cos^{j-i-1} o_{ij}\right] / |D|\ \ (dD)(dO) \\ &\propto 1/|D|\ \ (dD)(dH) \\ &\propto 1/\left[|\Sigma| \prod_{i<j}(d_i - d_j)\right]\ \ (d\Sigma). \end{aligned} \tag{13}$$

The Jeffreys prior is

$$\begin{aligned} \pi_J(D,\ O)\ \ (dD)(dO) &\propto \prod_{i<j}(d_i - d_j)/|D|^{\frac{p+1}{2}}\ \ (dD)(dH) \\ &\propto |\Sigma|^{-\frac{p+1}{2}}\ \ (d\Sigma). \end{aligned} \tag{14}$$

Proof. See Appendix B. □

Corollary 1 . *The resulting posterior distribution are*

$$\begin{aligned}
\pi_R(\Sigma|S)\ \ (d\Sigma) &\propto \frac{\mathrm{etr}(-\frac{1}{2}\Sigma^{-1}S)}{|\Sigma|^{\frac{n}{2}+1}\prod_{i<j}(d_i-d_j)}\ \ (d\Sigma) \\
&\propto \mathrm{etr}(-\frac{1}{2}OD^{-1}O^tS)/|D|^{\frac{n}{2}+1}\ \ (dD)(dH); \qquad (15) \\
\pi_J(\Sigma|S)\ \ (d\Sigma) &\propto \mathrm{etr}(-\frac{1}{2}\Sigma^{-1}S)/|\Sigma|^{\frac{n+p+1}{2}}\ \ (d\Sigma) \\
&\propto \prod_{i<j}(d_i-d_j)\ \mathrm{etr}(-\frac{1}{2}OD^{-1}O^tS)/|D|^{\frac{n+p+1}{2}}\ \ (dD)(dH), \qquad (16)
\end{aligned}$$

where etr *stands for* exp(tr()).

Proof. Simply multiply the prior and the likelihood. □

Note that the posterior in (15) is proper, having all moments of order less than $n/2$ (including negative moments), because it is bounded by an Inverse Gamma distribution. Compared to the Jeffreys prior, note that the reference prior seems to put considerably more mass near the region of equality of the eigenvalues; thus it is intuitively plausible that the reference prior would produce a covariance matrix estimator with better eigenstructure shrinkage.

Sometimes Σ^{-1}, rather than Σ itself, is of interest. Note, however, that the reference prior for Σ^{-1} will be the same as that for Σ. This follows from the fact that $\Sigma^{-1} = O^tD^{-1}O$, and that the reference prior for the group ordering that lists first the ordered $\{d_i^{-1}\}$, and then O, is the same as that listing $\{d_i\}$ followed by O. (It can be shown that a one-to-one transformation of an element of the group ordering does not change the reference prior.) Similarly, if it is the eigenvalues of Σ that are of interest, the reference prior again turns out to be given by (13). It is methodologically pleasant that this same reference prior emerges for any of the usual quantities of interest.

3 Computation of the Bayes Estimators

3.1 Bayes Estimators for Σ and Σ^{-1}

To find the Bayes estimators for Σ w.r.t the loss functions L_1 and L_2, one merely minimizes the associated posterior expected losses.

Lemma 3 . *The Bayes estimator for Σ w.r.t the posterior $\pi(\Sigma|S)$ and under L_1 is*

$$\delta_1^\pi = \left[E^{\pi(\Sigma|S)}\Sigma^{-1}\right]^{-1}; \tag{17}$$

the Bayes estimator under L_2 is

$$\mathrm{vec}(\delta_2^\pi) = \left[E^{\pi(\Sigma|S)}(\Sigma^{-1} \otimes \Sigma^{-1})\right]^{-1} \mathrm{vec}\left[E^{\pi(\Sigma|S)}\Sigma^{-1}\right]. \tag{18}$$

Both δ_1^π and δ_2^π are orthogonally invariant in the sense $\delta_i^{\pi(\Sigma|\Gamma S\Gamma^t)} = \Gamma\delta_i^{\pi(\Sigma|S)}\Gamma^t$, $i = 1, 2$, provided the prior is orthogonally invariant in the sense $\pi(\Gamma\Sigma\Gamma^t) = \pi(\Sigma)$, where Γ is an arbitrary orthogonal matrix. Also, for such priors, the Bayes estimators are diagonal when S is diagonal.

Proof. See Appendix C. □

Corollary 2 . *The Jeffreys prior Bayes estimator for the covariance matrix under L_1 is the usual unbiased estimator S/n.*

Proof. Straightforward computation. □

Often, estimation of Σ^{-1}, rather than Σ, is desired. The literature in this field includes Efron and Morris (1976), Haff (1977), Sharma and Krishnamoorthy (1985b), Sinha and Ghosh (1986), Krishnamoorthy and Gupta (1989) and Krishnamoorthy (1991). The commonly used loss functions are the natural analogues of (1) and (2), namely

$$L_1(\hat{\Sigma}^{-1},\ \Sigma^{-1}) = \mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) - \log|\hat{\Sigma}^{-1}\Sigma| - p,$$
$$L_2(\hat{\Sigma}^{-1},\ \Sigma^{-1}) = \mathrm{tr}(\hat{\Sigma}^{-1}\Sigma - I)^2.$$

As in Lemma 3, these two loss functions result in Bayes estimators of Σ^{-1} given by, respectively,

$$\left[E^{\pi(\Sigma|S)}\Sigma\right]^{-1} \quad \text{and} \quad \left[E^{\pi(\Sigma|S)}(\Sigma \otimes \Sigma)\right]^{-1} \mathrm{vec}\left[E^{\pi(\Sigma|S)}\Sigma\right].$$

Efron and Morris (1976) and Haff (1977) used a slightly different loss function,

$$L(\hat{\Sigma}^{-1},\ \Sigma^{-1};\ S) = \frac{\mathrm{tr}[(\hat{\Sigma}^{-1} - \Sigma^{-1})^2 S]}{k\mathrm{tr}(\Sigma^{-1})};$$

this would result in the Bayes estimator $E^{\pi(\Sigma|S)}[\Sigma^{-1}/\mathrm{tr}(\Sigma^{-1})]$. Haff (1977) also considered the loss function

$$L(\hat{\Sigma}^{-1},\ \Sigma^{-1}) = \mathrm{tr}[(\hat{\Sigma}^{-1} - \Sigma^{-1})^2 Q],$$

where Q is an arbitrary positive definite matrix; this would result in the simple Bayes estimator $E^{\pi(\Sigma|S)}[\Sigma^{-1}]$.

3.2 Exponential Matrix Transformation

In computing the expectations in (17) and (18), it will be convenient to transform from the space of positive definite matrices to all of Euclidean space. We do this, as in Leonard and Hsu (1992), by defining $\Sigma^* = \log\Sigma$, or $\Sigma = e^{\Sigma^*}$, in the sense that

$$\Sigma = \sum_{i=0}^{\infty}(\Sigma^*)^i/i!\,. \tag{19}$$

Writing $\Sigma^* = OD^*O^t$, with $D^* = \mathrm{diag}(d_1^*,\ldots,d_p^*)$, $d_1^* \geq d_2^* \geq,\cdots,\geq d_p^*$, and O orthogonal, it follows that $\Sigma = ODO^t$, with $D = \mathrm{diag}(d_1,\ldots,d_p)$, $d_i = e^{d_i^*}$, $1 \leq i \leq p$. By Lemma 2, the Jacobian of this transformation is

$$\begin{aligned}(d\Sigma) &\propto \prod_{i<j}(d_i - d_j)\ \ (dH)(dD)\\ &= |D|\prod_{i<j}(d_i - d_j)\ \ (dH)(dD^*)\\ &\propto \frac{|\Sigma|\prod_{i<j}(d_i - d_j)}{\prod_{i<j}(d_i^* - d_j^*)}\ \ (d\Sigma^*).\end{aligned} \tag{20}$$

Using (15), it follows that the reference posterior for Σ^* is

$$\pi_R^*(\Sigma^*|S)\ \ (d\Sigma^*) \propto \frac{\mathrm{etr}\{-\frac{n}{2}D^* - \frac{1}{2}Oe^{-D^*}O^tS\}}{\prod_{i<j}(d_i^* - d_j^*)}\ \ (d\Sigma^*). \tag{21}$$

The important feature of transforming to Σ^* is that Σ^* will be an unconstrained symmetric matrix, which is more efficient to simulate. One can simply transform back to Σ to get a simulated sample in the original space.

3.3 Hit-and-Run Sampler

For the reference posterior, analytical evaluation of the quantities in Lemma 3 appears to be quite difficult. Thus we turn to Monte Carlo integration to do the computation.

Recently, Monte Carlo methods for Bayesian integration have undergone extensive development. The methods that are commonly used are importance sampling, data augmentation, and the Gibbs sampler. Attempts to apply these methods encountered difficulties, so we turned to the less common Hit-and-Run sampler, which is another Markov chain sampler.

The Hit-and-Run sampler was first proposed by Smith (1980, 1984) and later generalized by Belisle, Romeijin and Smith (1993). The algorithm we used is a version that was developed by Chen and Schmeiser (1991, 1993), and is called the *Metropolisized Hit-and-Run Sampler.* This algorithm is particularly useful when the domain of the posterior along a random direction from a given point can be obtained without undue difficult.

Our actual sampling procedure proceeds as follows:

(i) Select a starting positive definite matrix Σ_0, set $\Sigma_0^* = \log \Sigma_0$ and $k = 0$. Here we choose $\Sigma_0 = \frac{1}{n} S$.

(ii) Select a random direction (symmetric) matrix

$$T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ t_{12} & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1p} & t_{2p} & \cdots & t_{pp} \end{pmatrix}, \tag{22}$$

defined by $T = Z / \sqrt{\sum_{i \leq j} z_{ij}^2}$, where $z_{ij} \overset{i.i.d}{\sim} N(0,\ 1)$, $i \leq j$, and Z is the symmetric matrix with $(i,\ j)$ element z_{ij}, $i \leq j$.

(iii) Generate $\lambda \sim N(0,\ 1)$.

(iv) Set $Y = \Sigma_k^* + \lambda T$. Then set

$$\Sigma_{k+1}^* = \begin{cases} Y, & \text{with probability } \min(1,\ \pi^*(Y|S)/\pi^*(\Sigma_k^*|S)) \\ \Sigma_k^*, & \text{otherwise.} \end{cases} \tag{23}$$

(v) Set $k = k + 1$ and go back to (ii).

Finally, after a sufficiently large sample $\Sigma_1^*, \Sigma_2^*, \ldots, \Sigma_N^*$ has been generated, one simply approximates a posterior expectation by $E^{\pi(\Sigma|S)} f(\Sigma) \approx \frac{1}{N}\sum_{k=1}^{N} f(e^{\Sigma_k^*})$, where f is the function of interest. As $N \to \infty$, the ergodic theorem asserts that the approximation converges to the true value (see Schmeiser and Chen, 1991). Of course, one should simultaneously evaluate $E^{\pi(\Sigma|S)}[f(\Sigma)]$ for all f of interest. In the simulation in Section 4, we set $p = 5$ (so that the integrals are 15 dimensional) and $N = 50,000$. This gave simulation accuracies (100 × simulation error / true value) of about 1.5% for the loss $L_1(\delta_1^\pi,\ \Sigma)$ and 0.75% for $L_2(\delta_2^\pi,\ \Sigma)$, the quantities needed in the risk evaluations. The individual elements of δ_1^π and δ_2^π were not quite so accurate, having simulation accuracies of about 5%.

4 Frequentist Risk Comparisons

4.1 Stein's and Haff's Estimators

Writing $S = VLV^t$, where V is a orthogonal matrix and $L = \mathrm{diag}(l_1, \ldots, l_p)$ with $l_1 \geq l_2 \geq \cdots \geq l_p$, Stein (1975) considered the orthogonal invariant estimator:

$$\hat{\Sigma} = V\Phi(L)V^t, \tag{24}$$

where $\Phi(L) = \mathrm{diag}(\phi_1, \ldots, \phi_p)$ with $\phi_i = l_i/\alpha_i$,

$$\alpha_i = (n - p + 1) + 2l_i \sum_{j \neq i} 1/(l_i - l_j), \qquad i = 1, \ldots, p. \tag{25}$$

This estimator has two problems. First, the intuitively compatible ordering $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_p$ is frequently violated. Second, and more seriously, some of the ϕ_i may even be negative. Stein suggests an isotonizing algorithm to avoid these problem. The idea of the algorithm is to pool the adjacent pairs $(l_i,\ \alpha_i)$. The resulting estimators of the eigenvalues are

$$\phi_i = \phi_{i+1} = \cdots = \phi_{i+s} = \frac{l_i + l_{i+1} + \cdots + l_{i+s}}{\alpha_i + \alpha_{i+1} + \cdots + \alpha_{i+s}}. \tag{26}$$

The details of this isotonizing algorithm can be found in Lin and Perlman (1985).

Haff's estimator (1991) is closely related to the above estimator. He minimizes the formal

Bayes risk for an orthogonally invariant prior by a variational technique. Assuming the prior yields $1/|S|$ as the marginal distribution of S, this technique reproduces Stein's unconstrained estimator. By imposing the constraint $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_p$ in the minimization under L_1, the formal Bayes estimator is of the form (24) with the eigenvalue estimators obtained by solving the equations

$$\epsilon_i \sum_{j=1}^{i} \left[(\phi_j^s)^{-1} - (\phi_j)^{-1} \right] = 0, \quad i = 1, 2, \ldots, p, \tag{27}$$

where $\phi_j^s = l_j/\alpha_j$, $j = 1, 2, \ldots, p$, and $\epsilon_1^2 = \phi_1 - \phi_2$, $\epsilon_2^2 = \phi_2 - \phi_3, \ldots,$ $\epsilon_p^2 = \phi_p$.

The two estimators discussed above are both obtained under L_1. Stein's and Haff's methods are difficult to apply under L_2. Thus it is common to take the L_1 estimators and simply rescale for L_2 (see Haff, 1991, and Lin and Perlman, 1985). That is, if $\hat{\Sigma}$ is derived under L_1, then one simply considers the estimator $n\hat{\Sigma}/(n+p+1)$ under L_2. This corresponds to the optimal rescaling for the unbiased estimator under L_1. Note that such adhoc adjustments are not required for the Bayes estimators.

4.2 Risk Simulations

The frequentist risks of the various estimators under L_1 and L_2 will be approximated by average losses in simulation. The simulation was designed as follows: Set $p = 5$ and $n = 10,\ 20,\ 40$. The test covariance matrices were chosen to be

$$\begin{aligned} \Sigma_1 &= \text{diag}(1,\ 1,\ 1,\ 1,\ 1), \\ \Sigma_2 &= \text{diag}(5,\ 4,\ 3,\ 2,\ 1), \\ \Sigma_3 &= \text{diag}(16,\ 8,\ 4,\ 2,\ 1). \end{aligned}$$

For fixed n and Σ_i, we do the following:

(i) Generate 50 random $Y_k \sim W_p(I,\ n)$, $1 \leq k \leq 50$, using Bartlett's decomposition, and then transform them into $W_p(\Sigma_i,\ n)$ random variables, S_k.

(ii) For each observation S_k, estimate the covariance matrix using the reference prior Bayes estimator, Stein's estimator, and Haff's estimator, under L_1 and L_2. Record the associated loss for each estimator.

(iii) Compute the mean and standard error of the *differences* in loss between the three different estimators.

(iv) Following the tradition of Lin and Perlman (1985), we also record the percentage reduction in average loss (PRIAL) of the three estimators relative to the usual estimator, defined by:

For L_1,

$$\text{PRIAL} = \frac{R(\frac{1}{n}S,\ \Sigma) - R(\hat{\Sigma},\ \Sigma)}{R(\frac{1}{n}S,\ \Sigma)} \times 100; \tag{28}$$

For L_2,

$$\text{PRIAL} = \frac{R(\frac{1}{n+p+1}S,\ \Sigma) - R(\hat{\Sigma},\ \Sigma)}{R(\frac{1}{n+p+1}S,\ \Sigma)} \times 100. \tag{29}$$

The simulation results for frequentist risk are given in Table 1, with the standard errors in parentheses. Table 2 presents the results for PRIAL.

Table 1. Risk Differences of Reference Prior, Stein, and Haff Estimators.

		L1		L2	
	n	R(Stein)−R(Ref.)	R(Haff)−R(Ref.)	R(Stein)−R(Ref.)	R(Haff)−R(Ref.)
Σ_1	10	−.14 (.016)	−.15 (.029)	.023 (.029)	−.13 (.023)
	20	−.056 (.0063)	−.059 (.010)	.0023 (.015)	−.061 (.012)
	40	−.031 (.0033)	−.029 (.0049)	−.025 (.0093)	−.036 (.0071)
Σ_2	10	.049 (.024)	.045 (.032)	.089 (.029)	.035 (.036)
	20	.050 (.011)	.054 (.011)	.069 (.019)	.066 (.019)
	40	.011 (.0037)	.012 (.0043)	.014 (.0070)	.016 (.0073)
Σ_3	10	.10 (.026)	.11 (.032)	.095 (.031)	.11 (.042)
	20	.044 (.0089)	.048 (.0089)	.045 (.015)	.051 (.014)
	40	.017 (.0033)	.017 (.0034)	.030 (.0066)	.030 (.0070)

Table 2. PRIAL Relative to the Usual Estimator.

		L1				L2			
	n	Ref.	Ref.*	Stein	Haff	Ref.	Ref.*	Stein	Haff
Σ_1	10	59.02	80.03	66.98	67.34	35.02	59.17	33.76	41.92
	20	64.06	84.19	71.15	71.50	50.46	75.36	50.25	55.89
	40	63.61	80.95	71.63	71.06	55.78	75.31	59.55	61.23
Σ_2	10	45.99	55.41	43.28	43.50	26.12	32.47	21.29	24.23
	20	34.71	31.58	28.35	27.84	23.69	16.50	17.52	17.82
	40	21.25	17.52	18.46	18.05	15.91	9.01	13.74	13.41
Σ_3	10	31.92	33.71	26.28	25.55	16.08	16.32	10.89	10.33
	20	14.08	11.61	8.45	8.01	8.80	6.10	4.82	4.26
	40	6.09	6.02	1.73	1.75	4.39	3.88	−1.66	−2.24

Table 3. Risk Differences of Modified Reference Prior, Stein, and Haff Estimators.

		L1		L2	
	n	R(Stein)−R(Ref.*)	R(Haff)−R(Ref.*)	R(Stein)−R(Ref.*)	R(Haff)−R(Ref.*)
Σ_1	10	.23 (.036)	.23 (.042)	.47 (.037)	.32 (.034)
	20	.10 (.015)	.10 (.016)	.28 (.024)	.22 (.023)
	40	.036 (.0076)	.039 (.0075)	.10 (.021)	.092 (.018)
Σ_2	10	.22 (.041)	.21 (.048)	.21 (.040)	.15 (.040)
	20	.025 (.019)	.029 (.018)	−.011 (.034)	−.014 (.032)
	40	−.0036 (.011)	−.0021 (.011)	−.031 (.032)	−.029 (.031)
Σ_3	10	.13 (.032)	.15 (.034)	.10 (.053)	.11 (.056)
	20	.025 (.017)	.028 (.016)	.014 (.026)	.021 (.024)
	40	.017 (.0051)	.017 (.0052)	.027 (.011)	.027 (.011)

The performance of the reference prior Bayes estimator is very comparable to that of the Stein and Haff estimators. It is somewhat worse when $\Sigma = I$, somewhat better otherwise. This behavior is indicative of an estimator that has more moderate shrinkage than that of the Stein or Haff estimators. This might well be desirable; indeed, for Σ far from the identity matrix, there is a suggestion in Table 2 that the Stein and Haff estimators overshrink, at least for L_2, where the PRIAL can become negative.

To investigate this further, we considered alternative reference priors that happened to yielded more shrinkage. For instance, if one applies the reference prior algorithm discussed in Appendix B to the ordered group $\{(d_1,\ d_p), (d_2,\ \ldots,\ d_{p-1}), (o_{12},\ \ldots,\ o_{p-1,p})\}$, the reference prior turns out to be

$$\pi_{R^*}(D,\ O)\ (dD)(dH)\ \propto\ |D|^{-1}[\log d_1 - \log d_p]^{-(p-2)}\ \ (dD)(dH), \tag{30}$$

which intuitively will induce more shrinkage than will (13). Risk differences and PRIALs for the associated Bayes estimators are given in Table 3 and the "Ref.*" columns of Table 2, respectively. The apparently more aggressive shrinkage of these estimators results in dramatically improved risk when $\Sigma \cong I$. And these high shrinkage Bayes estimators seem somewhat superior to the Stein and Haff estimators.

Note that there is nothing particular compelling about π_{R^*} from a reference prior perspective. Indeed, we would still probably recommend π_R in (13); it is likely to have better confidence properties, and in general one should be wary of overshrinking. It must be admitted, however, that all these conclusions are quite tentative, being based on only a very limited study. Indeed, we feel that all four estimators considered here are comparably effective in practice

5 Determination of Accuracy

It is something of a bonus that the reference prior Bayes estimators may actually have superior risk properties, compared with existing shrinkage estimators. The primary motivation and value of the Bayesian approach to estimation problems such as this is, rather, that the Bayesian approach readily allows determination of accuracy of estimation, and allows associated prediction and numerous other types of inference involving Σ. We illustrate this here by providing estimates of the loss of $\hat{\Sigma}$ under L_1 and L_2.

Lemma 4 . *The posterior expected loss of the Bayes estimator in Lemma 3 under L_1 is*

$$\rho_1(\pi(\Sigma|S),\ \delta_1^{\pi}) = E^{\pi(\Sigma|S)} \log|\Sigma| - \log|\delta_1^{\pi}|;$$

the posterior expected loss under L_2 is

$$\rho_2(\pi(\Sigma|S),\ \delta_2^\pi) = p - \mathrm{tr}[\delta_2^\pi(\delta_1^\pi)^{-1}].$$

Proof. See Appendix D. □

Example 1. Consider, as data, the following matrix S/n, where S is generated from a $W_5(\Sigma,\ 10)$ distribution, with $\Sigma = \mathrm{diag}(5, 4, 3, 2, 1)$:

$$S/n = \begin{pmatrix} 1.925 & 1.618 & 0.132 & -1.101 & 0.264 \\ 1.618 & 8.437 & 1.638 & -0.880 & -0.983 \\ 0.132 & 1.638 & 2.147 & -0.439 & -0.646 \\ -1.101 & -0.880 & -0.439 & 1.331 & -0.035 \\ 0.264 & -0.983 & -0.646 & -0.035 & 1.280 \end{pmatrix}.$$

The corresponding Bayes estimators are

$$\delta_1^\pi = \begin{pmatrix} 1.726 & 0.946 & 0.077 & -0.712 & 0.207 \\ 0.946 & 5.536 & 0.917 & -0.483 & -0.567 \\ 0.077 & 0.917 & 1.889 & -0.282 & -0.385 \\ -0.712 & -0.483 & -0.282 & 1.366 & -0.022 \\ 0.207 & -0.567 & -0.385 & -0.022 & 1.412 \end{pmatrix},$$

$$\delta_2^\pi = \begin{pmatrix} 1.233 & 0.723 & 0.057 & -0.567 & 0.165 \\ 0.723 & 4.120 & 0.682 & -0.364 & -0.413 \\ 0.057 & 0.682 & 1.371 & -0.215 & -0.307 \\ -0.567 & -0.364 & -0.215 & 0.936 & -0.018 \\ 0.165 & -0.413 & -0.307 & -0.018 & 0.993 \end{pmatrix}.$$

Using Lemma 4, the posterior expected losses of these estimators can be computed to be $\rho_1(\pi(\Sigma|S),\ \delta_1^\pi) = 1.152$ and $\rho_2(\pi(\Sigma|S),\ \delta_2^\pi) = 1.509$, respectively. Note that the actual losses (computable since we know Σ) are $L_1(\delta_1^\pi, \Sigma) = 1.270$ and $L_2(\delta_2^\pi, \Sigma) = 1.548$, respectively. (For comparison, observe that the actual losses for S/n and $S/(n+p+1)$ are $L_1(S/n, \Sigma) = 2.267$ and $L_2(S/(n+p+1), \Sigma) = 2.147$.)

Classical estimates of loss (the unbiased estimates of risk) are available here (cf., Haff, 1991), but they are unwieldy and potentially unreliable (for instance, they can even be negative).

Other estimates of accuracy could be found using the Bayesian approach, such as the posterior covariance matrix for Σ (a $[p(p+1)/2] \times [p(p+1)/2]$ matrix), or even credible intervals for components of Σ. This could also be done for functions of Σ that are of interest.

6 Comments and Generalizations

1. Though unquestionably computationally intensive, the rewards for adopting the Bayesian approach here are considerable. Resulting estimators have exceptional risk properties, and determination of accuracy and inference for functions of Σ is straightforward. Indeed, once the computation for $\hat{\Sigma}$ is set up, it is easy to compute the expectation of any function $g(\Sigma)$ that is of interest.

2. The situation in which the $\vec{X}_i$ are i.i.d. $N_p(\vec{\mu}, \Sigma)$ can be handled similarly; the reference prior will be constant over $\vec{\mu}$, so that $\vec{\mu}$ can simply be integrated out to reduce the problem (in a Bayesian sense) to consideration of

$$S = \sum_{i=1}^{n}(\vec{X}_i - \bar{X})(\vec{X}_i - \bar{X})^t \sim W_p(\Sigma, n-1).$$

Analysis then proceeds as before, with n replaced by $n-1$. Note, however, that, from a frequentist or a hierarchical Bayesian perspective, there might be advantages in utilizing "shrinkage priors" for μ, rather than the constant prior.

3. As with all scale problems, choice of the loss is a rather perplexing question. From Example 1 it is clear that the effect can be substantial. We have no firm recommendation here, except to note that, when p=1, the Jeffreys and reference priors both equal the usual invariant prior $1/\sigma^2$, and use of L_1 with this prior yields, as the Bayes estimator, the "standard" estimator S/n (see Corollary 2). Hence use of L_1, and the corresponding δ_1^π, has some appeal.

4. It can be argued that the reference prior should depend on the loss function. For instance, when p=1, the standard reference prior is $1/\sigma^2$, and this is completely satisfactory for invariant

losses such as L_1 or L_2, but it is not optimal for, say, squared error loss. Unfortunately, it is not easy, in general, to determine how the reference prior should depend on the loss (see Bernardo, 1981, and Bernardo and Smith, 1994, for discussion). In our situation, the problem is probably not severe, since we only utilized invariant losses.

5. A related problem can arise if one is interested in some function, $g(\Sigma)$, of Σ. Conceivably a better reference prior can be developed that recognizes the centrality of $g(\Sigma)$ (cf., Berger and Bernardo, 1992c). Use of the given reference prior, π_R, is likely to be quite satisfactory, however, especially because it arises from so many different group orderings that it will be the reference prior for "most" $g(\Sigma)$.

Appendix A: Proof of Lemma 1

Tracy and Jinadasa (1988) established that the Fisher information matrix for Σ is

$$I(\Sigma) = \frac{1}{2} G^t(\Sigma^{-1} \otimes \Sigma^{-1})G, \tag{31}$$

where G is defined as $G = \partial \text{vec}(\Sigma)/\partial \text{vecp}(\Sigma)$, with $\Sigma = O^t DO$ and $\text{vecp}(\Sigma) = (d_1, \ldots, d_p, o_{12}, \ldots, o_{1p}, o_{23}, \ldots, o_{2p}, \ldots, o_{p-1,p})$. Writing $\vec{a}_i = \text{vec}\partial(O^tDO)/\partial d_i$ and $\vec{b}_{ij} = \text{vec}\partial(O^tDO)/\partial o_{ij}$ yields $G = (\vec{a}_1, \ldots, \vec{a}_p, \vec{b}_{12}, \ldots, \vec{b}_{p-1,p})$, and thus the Fisher information matrix, w.r.t the reparameterization $(D,\ O)$, is

$$I(D,\ O) = \frac{1}{2} \begin{pmatrix} \vec{a}_1^t \\ \vdots \\ \vec{a}_p^t \\ \vec{b}_{12}^t \\ \vdots \\ \vec{b}_{p-1,p}^t \end{pmatrix} (\Sigma^{-1} \otimes \Sigma^{-1})(\vec{a}_1, \ldots, \vec{a}_p, \vec{b}_{12}, \ldots, \vec{b}_{p-1,p}). \tag{32}$$

The elements of $I(D,\ O)$ are of three types:

1. $\vec{a}_i^t(\Sigma^{-1} \otimes \Sigma^{-1})\vec{a}_j$,

2. $\vec{a}_i^t(\Sigma^{-1} \otimes \Sigma^{-1})\vec{b}_{rs}$, or $\vec{b}_{rs}^t(\Sigma^{-1} \otimes \Sigma^{-1})\vec{a}_i$,

3. $\vec{b}_{ij}^t(\Sigma^{-1} \otimes \Sigma^{-1})\vec{b}_{rs}$.

To finish the proof of Lemma 1, we need only evaluate the first two types. We will utilize the following matrix equality (from, e.g., Magnus and Neudecker, 1988) to do so:

$$\text{tr}(ABCD) = (\text{vec}D^t)^t(C^t \otimes A)(\text{vec}B). \tag{33}$$

(i) For the first type,

$$\begin{aligned}
\vec{a}_i^t(\Sigma^{-1}\otimes\Sigma^{-1})\vec{a}_j &= (\text{vec}\frac{\partial\Sigma}{\partial d_i})(\Sigma^{-1}\otimes\Sigma^{-1})(\text{vec}\frac{\partial\Sigma}{\partial d_j}) \\
&= \text{tr}(\Sigma^{-1}\frac{\partial\Sigma}{\partial d_j}\Sigma^{-1}\frac{\partial\Sigma}{\partial d_i}) \\
&= \text{tr}[(O^tD^{-1}O)(O^t\frac{\partial D}{\partial d_j}O)(O^tD^{-1}O)(O^t\frac{\partial D}{\partial d_i}O)] \\
&= \text{tr}[D^{-1}\frac{\partial D}{\partial d_j}D^{-1}\frac{\partial D}{\partial d_i}] \\
&= \begin{cases} 1/d_i^2 & \text{if } i=j \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{34}$$

(ii) For the second type,

$$\begin{aligned}
\vec{a}_i^t(\Sigma^{-1}\otimes\Sigma^{-1})\vec{b}_{rs} &= (\text{vec}\frac{\partial\Sigma}{\partial d_i})(\Sigma^{-1}\otimes\Sigma^{-1})(\text{vec}\frac{\partial\Sigma}{\partial o_{rs}}) \\
&= \text{tr}(\Sigma^{-1}\frac{\partial\Sigma}{\partial o_{rs}}\Sigma^{-1}\frac{\partial\Sigma}{\partial d_i}) \\
&= \text{tr}[(O^tD^{-1}O)\frac{\partial(O^tDO)}{\partial o_{rs}}(O^tD^{-1}O)(O^t\frac{\partial D}{\partial d_i}O)] \\
&= \text{tr}[D^{-1}\frac{\partial D}{\partial d_i}D^{-1}\cdot O\frac{\partial(O^tDO)}{\partial o_{rs}}O^t] \\
&= \text{tr}[\frac{\partial D}{\partial d_i}D^{-1}\cdot O\frac{\partial O^t}{\partial o_{rs}} + D^{-1}\frac{\partial D}{\partial d_i}\cdot\frac{\partial O}{\partial o_{rs}}O^t] \\
&= \text{tr}[\text{diag}(0,\ldots,0,1/d_i,0,\ldots,0)(O\frac{\partial O^t}{\partial o_{rs}} + \frac{\partial O}{\partial o_{rs}}O^t)] \\
&\overset{say}{=} \text{tr}[\text{diag}(0,\ldots,0,1/d_i,0,\ldots,0)(C+C^t)],
\end{aligned} \tag{35}$$

where $C = O\partial O^t/\partial o_{rs}$.

To complete the proof, it is enough to show that C is skew symmetric. Note that

$$\begin{aligned} C &= (O_{12}\cdots O_{1p})(O_{23}\cdots O_{2p})\cdots(O_{p-1,p})\frac{\partial[O^t_{p-1,p}\cdots(O^t_{1p}\cdots O^t_{12})]}{\partial o_{rs}} \\ &= O_{12}\cdots O_{rs}\frac{\partial O^t_{rs}}{\partial o_{rs}}\cdots O^t_{12}. \end{aligned} \tag{36}$$

Thus it suffices to show that $O_{rs}\frac{\partial O^t_{rs}}{\partial o_{rs}}$ is skew symmetric. This is true, because

$$\begin{aligned} O_{rs}\frac{\partial O^t_{rs}}{\partial o_{rs}} &= \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & \cos o_{ij} & 0 & -\sin o_{ij} & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & \sin o_{ij} & 0 & \cos o_{ij} & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -\sin o_{ij} & 0 & \cos o_{ij} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -\cos o_{ij} & 0 & -\sin o_{ij} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{37}$$

Appendix B: Proof of Theorem 1

First, we briefly sketch the algorithm from Berger and Bernardo (1992c) for computing ordered group reference priors. Let $\theta = (\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(m)})$ be the m groups of unknown parameters, where each $\theta_{(i)} = (\theta_{i_1}, \theta_{i_2}, \ldots, \theta_{i_{n_i}})$ has size n_i. Define

$$\theta_{[j]} = (\theta_{(1)}, \ldots, \theta_{(j)}), \text{ and } \theta_{[\sim j]} = (\theta_{(j+1)}, \ldots, \theta_{(m)}),$$

with the conventions that $\theta_{[\sim 0]} = \theta$ and $\theta_{[0]}$ is vacuous. Let $S(\theta) = (I(\theta))^{-1}$, where $I(\theta)$ is the

Fisher information matrix. Write $S(\theta)$ as

$$S(\theta) = \begin{pmatrix} A_{11} & A_{21}^t & \cdots & A_{m1}^t \\ A_{21} & A_{22} & \cdots & A_{m2}^t \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{pmatrix}$$

so that A_{ij} is $(n_i \times n_j)$. Denoting $N_j = \sum_{i=1}^{j} n_i$, define

$$\begin{aligned} S_j &\equiv \text{ upper left } (N_j \times N_j) \text{ corner of } S, \\ H_j &\equiv S_j^{-1}, \text{ and} \\ h_j &\equiv \text{ lower right } (n_j \times n_j) \text{ corner of } H_j,\ j = 1, \ldots, m. \end{aligned}$$

Suppose $\Theta^1 \subset \Theta^2 \subset \cdots$ are a sequence of compact subset of Θ such that $\cup_{i=1}^{\infty} \Theta^i = \Theta$, where Θ is the domain for θ. Define

$$\Theta^l(\theta_{[j]}) = \{\theta_{(j+1)} : (\theta_{[j]}, \theta_{(j+1)}, \theta^*_{[\sim(j+1)]}) \in \Theta^l \text{ for some } \theta^*_{[\sim(j+1)]}\}.$$

Also define

$$1_{\Omega}(y) \quad = \quad \begin{cases} 1 & \text{if } y \in \Omega \\ 0 & \text{otherwise.} \end{cases}$$

In the situation where $|h_j(\theta)|$ depends only on $\theta_{[j]}$, for $j = 1, \ldots, m$, the reference prior is given by

$$\pi(\theta) = \lim_{l \to \infty} \frac{\pi_1^l(\theta)}{\pi_1^l(\theta^*)}, \tag{38}$$

where θ^* is any fixed point in Θ with positive density for all π_1^l, and

$$\pi_1^l(\theta) = \left(\prod_{i=1}^{m} \frac{|h_i(\theta)|^{1/2}}{\int_{\Theta^l(\theta_{[i-1]})} |h_i(\theta)|^{1/2} d\theta_{(i)}} \right) 1_{\Theta^l}(\theta). \tag{39}$$

We will take the ordered group to be $\{d_1, \ldots, d_p, (o_{12}, \ldots, o_{p-1,p})\}$, as an example of the computation of the reference prior for Σ; all the other ordered groups give the same answer, providing the $\{d_i\}$ are listed before the $\{o_{ij}\}$ and the $\{d_i\}$ are ordered monotonically (either increasing or decreasing). Define the compact subsets of the parameter space to be

$$\Theta^l = \{(D,\, O) : 0 < a_l \le d_p \le \cdots \le d_1 \le b_l < \infty,\ -\pi/2 < o_{ij} \le \pi/2,\ \forall i \le j\},$$

where $a_l \to 0$ and $b_l \to \infty$. Using Lemmas 1 and 2, note that

$$\begin{aligned} h_i &= 1/(2d_i^2), \quad i = 1, \ldots, p, \\ |h_{p+1}| &\propto \left[\prod_{i=1}^{p-1} \prod_{j=i+1}^{p} \cos^{j-i-1} o_{ij}\right]^2 \left[\prod_{i<j}(d_i - d_j)\right]^2 \prod_{i=1}^{p} d_i^{-(p-1)}. \end{aligned}$$

Also,

$$\begin{aligned} \Theta^l(\theta_{[i-1]}) &= \{d_i : a_l \le d_i \le d_{i-1}\}, \quad i = 1, \ldots, p, \\ \Theta^l(\theta_{[(p+1)-1]}) &= \{0_{ij} : -\pi/2 < o_{ij} \le \pi/2,\ \forall i \le j\}, \end{aligned}$$

where d_0 is interpreted as b_l. Thus (38) becomes

$$\begin{aligned} \pi_1^l(\theta) &= \left(\prod_{i=1}^{p} \frac{1/d_i}{\int_{a_l \le d_i \le d_{i-1}} 1/d_i\, dd_i}\right) \left(\frac{|h_{p+1}|^{1/2}}{\int_{-\pi/2 < o_{ij} \le \pi/2} |h_{p+1}|^{1/2} dO}\right) 1_{\Theta^l}(\theta). \\ &\propto \frac{\prod_{i=1}^{p} 1/d_i}{(\log b_l - \log a_l) \prod_{i=2}^{p} (\log d_{i-1} - \log a_l)} \left[\prod_{i=1}^{p-1} \prod_{j=i+1}^{p} \cos^{j-i-1} o_{ij}\right]. \end{aligned}$$

From (37), the reference prior w.r.t this group ordering is thus given by

$$\begin{aligned} \pi(D,\, O) &\propto \lim_{l \to \infty} \frac{\pi_1^l(\theta)}{\pi_1^l(\theta^*)} \\ &\propto \left[\prod_{i=1}^{p-1} \prod_{j=i+1}^{p} \cos^{j-i-1} o_{ij}\right] / |D|. \end{aligned}$$

Appendix C: Proof of Lemma 3

Let δ denote an arbitrary estimator with $(i,\ j)$ element δ_{ij}. Bayes estimators are calculated as follows:

(i) For the loss function L_1, define

$$R_1^{\pi(\Sigma|S)}(\delta,\ \Sigma) \ = \ E^{\pi(\Sigma|S)}[\mathrm{tr}(\delta\Sigma^{-1}) - \log|\delta\Sigma^{-1}| - p].$$

Using the matrix identities (see, e.g., Magnus and Neudecker, 1988)

$$\frac{\partial \mathrm{tr}(\delta A)}{\partial\delta} = A^t, \text{ and } \frac{\partial|\delta|}{\partial\delta} = |\delta|(\delta^{-1})^t,$$

and setting $\partial R_1^{\pi(\Sigma|S)}(\delta,\ \Sigma)/\partial\delta = 0$, one obtains the Bayes estimator as $\left[E^{\pi(\Sigma|S)}\Sigma^{-1}\right]^{-1}$.

(ii) For the loss function L_2, define

$$\begin{aligned} R_2^{\pi(\Sigma|S)}(\delta,\ \Sigma) &= E^{\pi(\Sigma|S)}\mathrm{tr}(\delta\Sigma^{-1} - I)^2 \\ &= E^{\pi(\Sigma|S)}\mathrm{tr}(\delta\Sigma^{-1}\delta\Sigma^{-1} - 2\delta\Sigma^{-1} + I). \end{aligned}$$

By setting $\partial R_2^{\pi(\Sigma|S)}(\delta,\ \Sigma)/\partial\delta_{ij} = 0$, it follows that

$$E^{\pi(\Sigma|S)}(\Sigma^{-1}\delta\Sigma^{-1} - \Sigma^{-1})_{[j,i]} = 0,$$

$$E^{\pi(\Sigma|S)}\sum_{l=1}^{p}\sum_{k=1}^{p}(\Sigma^{-1})_{[j,l]}\delta_{[l,k]}(\Sigma^{-1})_{[k,i]} = E^{\pi(\Sigma|S)}(\Sigma^{-1})_{[j,i]},$$

$$[\text{the } (i-1)p+j\text{th row of } E^{\pi(\Sigma|S)}(\Sigma^{-1}\otimes\Sigma^{-1})]\mathrm{vec}(\delta) = E^{\pi(\Sigma|S)}(\Sigma^{-1})_{[j,i]}, \quad \forall i,\ j.$$

Therefore,

$$\mathrm{vec}(\delta_2^{\pi}) = \left[E^{\pi(\Sigma|S)}(\Sigma^{-1}\otimes\Sigma^{-1})\right]^{-1}\mathrm{vec}\left[E^{\pi(\Sigma|S)}\Sigma^{-1}\right].$$

If the prior is orthogonally invariant, orthogonal invariance of δ_1^{π} is trivial. For δ_2^{π},

$$\begin{aligned} \mathrm{vec}(\delta_2^{\pi(\Sigma|\Gamma S\Gamma^t)}) &= \left[E^{\pi(\Sigma|\Gamma S\Gamma^t)}(\Sigma^{-1}\otimes\Sigma^{-1})\right]^{-1}\mathrm{vec}\left[E^{\pi(\Sigma|\Gamma S\Gamma^t)}\Sigma^{-1}\right] \\ &= \left[E^{\pi(\Sigma|S)}(\Gamma\Sigma^{-1}\Gamma^t)\otimes(\Gamma\Sigma^{-1}\Gamma^t)\right]^{-1}\mathrm{vec}\left[E^{\pi(\Sigma|S)}\Gamma\Sigma^{-1}\Gamma^t\right] \\ &= \left[(\Gamma\otimes\Gamma)E^{\pi(\Sigma|S)}(\Sigma^{-1}\otimes\Sigma^{-1})(\Gamma^t\otimes\Gamma^t)\right]^{-1}\mathrm{vec}\left[\Gamma E^{\pi(\Sigma|S)}\Sigma^{-1}\Gamma^t\right] \\ &= (\Gamma\otimes\Gamma)\left[E^{\pi(\Sigma|S)}(\Sigma^{-1}\otimes\Sigma^{-1})\right]^{-1}(\Gamma^t\otimes\Gamma^t)\cdot(\Gamma\otimes\Gamma)\mathrm{vec}(E^{\pi(\Sigma|S)}\Sigma^{-1}) \\ &= (\Gamma\otimes\Gamma)\mathrm{vec}(\delta_2^{\pi(\Sigma|S)}) \\ &= \mathrm{vec}(\Gamma\delta_2^{\pi(\Sigma|S)}\Gamma^t), \end{aligned}$$

where the last three steps used the following matrix equality (from, e.g., Magnus and Neudecker, 1988):

$$\text{vec}(ABC) = (C^t \otimes A)(\text{vec}B).$$

Now suppose S is diagonal, define $U = \text{diag}(1,\ldots,1,-1,1,\ldots,1)$, and consider the transformation $\Lambda = U\Sigma U^t$. Then

$$\begin{aligned}
\delta_1^{\pi(\Sigma|S)} &= \left[E^{\pi(\Sigma|S)}\Sigma^{-1}\right]^{-1} \\
&\propto \left[\int_\Sigma \Sigma^{-1}\frac{\text{etr}(-\frac{1}{2}\Sigma^{-1}S)}{|\Sigma|^{n/2}}\pi(\Sigma)\,(d\Sigma)\right]^{-1} \\
&= \left[\int_\Lambda U^t\Lambda^{-1}U\frac{\text{etr}(-\frac{1}{2}U^t\Lambda^{-1}US)}{|U^t\Lambda U|^{n/2}}\pi(U^t\Lambda U)\,(d\Lambda)\right]^{-1} \\
&= \left[\int_\Lambda U^t\Lambda^{-1}U\frac{\text{etr}(-\frac{1}{2}\Lambda^{-1}S)}{|\Lambda|^{n/2}}\pi(\Lambda)\,(d\Lambda)\right]^{-1} \\
&\propto U^t\delta_1^{\pi(\Sigma|S)}U,
\end{aligned}$$

from which it follows that the off diagonal elements of $\delta_1^{\pi(\Sigma|S)}$ are 0. For $\delta_2^{\pi(\Sigma|S)}$, use of the same transformation as above yields

$$\begin{aligned}
E^{\pi(\Sigma|S)}(\Sigma^{-1}\otimes\Sigma^{-1}) &= E^{\pi(\Lambda|S)}[(U^t\Lambda^{-1}U)\otimes(U^t\Lambda^{-1}U)] \\
&= (U^t\otimes U^t)E^{\pi(\Lambda|S)}(\Lambda^{-1}\otimes\Lambda^{-1})(U\otimes U).
\end{aligned}$$

Letting i denote the coordinate of U equal to -1, note that $U\otimes U$ is a diagonal matrix with the $(i-1)p+j$th diagonal element equal to -1 $(\forall j\neq i)$. It follows that the $(i-1)p+j$th row and the $(i-1)p+j$th column $(\forall j\neq i)$ of $E^{\pi(\Sigma|S)}(\Sigma^{-1}\otimes\Sigma^{-1})$ are 0 except for the $[(i-1)p+j,\ (i-1)p+j]$ element. The same is thus true for $\left[E^{\pi(\Sigma|S)}(\Sigma^{-1}\otimes\Sigma^{-1})\right]^{-1}$. But, since $E^{\pi(\Sigma|S)}\Sigma^{-1}$ is diagonal, the $(i-1)p+j$th element $(\forall j\neq i)$ of vec $\left[E^{\pi(\Sigma|S)}\Sigma^{-1}\right]$ is 0. By (18), it follows that the $(i-1)p+j$th element $(\forall j\neq i)$ of $\text{vec}(\delta_2^{\pi(\Sigma|S)})$ is 0, and hence $\delta_2^{\pi(\Sigma|S)}$ is diagonal.

Appendix D: Proof of Lemma 4

Computation of $\rho(\pi(\Sigma|S),\, \delta_1^\pi)$ is trivial. For L_2,

$$\begin{aligned}
\rho(\pi(\Sigma|S),\, \delta_2^\pi) &= E^{\pi(\Sigma|S)}\mathrm{tr}(\delta_2^\pi\Sigma^{-1} - I)^2 \\
&= E^{\pi(\Sigma|S)}\mathrm{tr}(\delta_2^\pi\Sigma^{-1}\delta_2^\pi\Sigma^{-1} - 2\delta_2^\pi\Sigma^{-1} + I) \\
&= E^{\pi(\Sigma|S)}\mathrm{tr}(\delta_2^\pi\Sigma^{-1}\delta_2^\pi\Sigma^{-1}) - 2\mathrm{tr}(\delta_2^\pi E^{\pi(\Sigma|S)}\Sigma^{-1}) + p \\
&= p - \mathrm{tr}[\delta_2^\pi(\delta_1^\pi)^{-1}],
\end{aligned}$$

where the last step follows from

$$\begin{aligned}
E^{\pi(\Sigma|S)}\mathrm{tr}(\delta_2^\pi\Sigma^{-1}\delta_2^\pi\Sigma^{-1}) &= E^{\pi(\Sigma|S)}(\mathrm{vec}\delta_2^\pi)^t(\Sigma^{-1}\otimes\Sigma^{-1})(\mathrm{vec}\delta_2^\pi) \\
&= (\mathrm{vec}\delta_2^\pi)^t\mathrm{vec}(E^{\pi(\Sigma|S)}\Sigma^{-1}) \\
&= (\mathrm{vec}\delta_2^\pi)^t\mathrm{vec}[(\delta_1^\pi)^{-1}].
\end{aligned}$$

Acknowledgments. The authors would like to thank Ming-Hui Chen for his generous help with the computer calculations. We would also like to thank L.R. Haff for providing Fortran subroutines for computation of his estimators, and the referees for helpful suggestions.

References

Anderson, T.W. (1984). *Introduction to Multivariate Statistical Analysis.* New York: John Wiley & Sons.

Anderson, T.W., Olkin, I. and Underhill, L.G. (1987). Generation of random orthogonal matrices. SIAM *J. Sci. Stat. Comput.* **8**, 625-629.

Belisle, C.J.P., Romeijin, H.E. and Smith, R.L. (1993). Hit-and-Run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, **18, 2.**

Berger, J. and Bernardo, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200-207.

Berger, J. and Bernardo, J.M. (1992a). Reference priors in a variance components problem. In *Bayesian Analysis in Statistics and Econometrics*, P. Goel and N.S. Iyengar(eds.). New York: Springer Verlag.

Berger, J. and Bernardo, J.M. (1992b). Ordered group reference priors with application to a multinomial problem. *Biometrika* **79**, 25-37.

Berger, J. and Bernardo, J.M. (1992c). On the development of the reference prior method. In *Bayesian Statistics* 4, J.M.Bernardo, J.O.Berger, D.V.Lindley, and A.F.M.Smith (eds.). London: Oxford University Press.

Bernardo, J.M. (1979). Reference posterior distributions for Bayes inference. *J. Roy. Statist. Soc. Ser* **B 41**, 113-147 (with discussion).

Bernardo, J.M. (1981). Reference decisions. *Symposia Mathematica* **25**, 85-94.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* New York: Wiley.

Chang, T. and Eaves, D. (1990). Reference priors for the orbit in a group model. *Ann. Statist.* **18**, 1595-1614.

Chen, M.H. and Schmeiser, B.W. (1993). Performance of the Gibbs, Hit-and-Run, and Metropolis samplers. *J. Comput. Graph. Statist.* **2**, 251-272.

Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading, MA.

Dey, D.K. (1988). Simultaneous estimation of eigenvalues. *Ann. Inst. Statist. Math.* **40**, No. 1, 137-147.

Dey, D.K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13**, 1581-1591.

Dey, D.K. and Srinivasan, C. (1986). Trimmed minimax estimation of a covariance matrix. *Ann. Inst. Statist. Math.* **38 A** 47-54.

Efron, B. and Morris, C. (1976). Multivariate empirical Bayes estimation of covariance matrices. *Ann. Statist.* **4** 22-32.

Farrell, Roger H. (1985). *Multivariate Calculation.* New York: Springer-Verlag.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameter. *J. Roy. Statist. Soc. Ser* **B 25**, 368-376.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.* **36**, 150-159.

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 application. *J. Multivariate Anal.*

9, 531-542.

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

Haff, L.R. (1991). The variational form of certain Bayes estimators. *Ann. Statist.* **19**, 1163-1190.

Jeffreys, H. (1961). *Theory of Probability.* London: Oxford University Press.

Krishnamoorthy, K. and Gupta, A.K. (1989). Improved minimax estimation of a normal precision matrix. *Canad. J. Statist.* **17**, No. 1, 91-102.

Krishnamoorthy, K. (1991). Estimation of normal covariance and precision matrices with incomplete data. *Commun. Statist, Theory Meth.*, **20(3)**, 757-770.

Leonard, T. and Hsu, J.S.J. (1992). Bayesian inference for a covariance matrix. *Ann. Statist.* **20** 1669-1696.

Lin, S.P. and Perlman, M.D. (1985). A Monte Carlo comparison of four estimators for a covariance matrix. *Multivariate Analy. VI,* 411-429, ed. P.R. Krishnaiah, North Holland, Amsterdam.

Loh, W.L. (1991a). Estimating covariance matrices I. *Ann. Statist.* **19**, 283-296.

Loh, W.L. (1991b). Estimating covariance matrices II. *J. Multivariate Anal.* **36**, 163-174.

Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* New York: John Wiley & Sons.

Olkin, I. and Selliah, J.B. (1977). Estimating covariance in a multivariate normal distribution. *Statistical Decision Theory and Related Topics II,* 313-326, ed. S.S. Gupta & D. Moore. Academic Press, New York.

Perron, F. (1992). Minimax estimators of a covariance matrix. *J. Multivariate Anal.* **43**, 16-28.

Press, S.J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Measures of Inference* (2nd edn.). New York: Kreiger.

Schmeiser, B.W. and Chen, M.H. (1991). On Hit-and-Run Monte Carlo sampling for evaluating multidimensional integrals. Technical Report **#91-39**, Department of Statistics, Purdue University.

Sharma, D. (1980). An estimator of normal covariance matrix. *Calcutta Statist. Assoc. Bull.* **29**, 161-167.

Sharma, D. and Krishnamoorthy (1983). Orthogonal equivariant estimators of bivariate normal covariance matrix and precision matrix. *Calcutta Statist. Assoc. Bull.* **32**, 23-45.

Sharma, D. and Krishnamoorthy (1985a). Empirical Bayes estimators of normal covariance matrix. *Sankhya Ser* **B** **47**, No. 2, 247-254.

Sharma, D. and Krishnamoorthy (1985b). Improved minimax estimators of normal covariance and precision matrices from incomplete samples. *Calcutta Statist. Assoc. Bull.* **34**, 23-42.

Sinha, B.K. and Ghosh, M. (1986). Inadmissibility of the best equivariant estimators of the variance-covariance matrix, the precision matrix, and the generalized variance under entropy loss. *Statist. Decisions.* **5**, 201-227.

Smith, R.L. (1980). A Monte Carlo procedures for generating random feasible solutions to mathematical programs. A Bulletin of the ORSA/TIMS Joint National Meeting, Washington, D.C., 101.

Smith, R.L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research* **32**, 6, 1297-1308.

Stein, C. (1956). Some problems in multivariate analysis. Part I. Technical Report No.**6**, Department of Statistics, Stanford University.

Stein, C. (1975). "Estimation of a covariance matrix". Rietz lecture, 39th annual meeting IMS. Atlanta, Georgia.

Stein, C. (1977a). Unpublished notes on estimating the covariance matrix.

Stein, C. (1977b). Lectures on the theory of estimation of many parameters. (In Russian.) In *Studies in the Statistical Theory of Estimation, Part I* (Ibragimov, I.A. and Nikulin, M.S., eds.), *Proceedings of Scientific Seminars of the Steklov Institute, Leningrad Division* **74**, 4-65.

Sugiura, N. and Fujimoto, M. (1982). Asymptotic risk comparison of improved estimators for normal covariance matrix. *Tsukuba J. of Math.* **6**, 103-126.

Takemura, A. (1984). An orthogonally univariant minimax estimator of the covariance matrix of a multivariate normal population. *Tsukuba J. Math.* **8**, No. 2, 367-376.

Tracy, D.S. and Jinadasa, K.G. (1988). Patterned matrix derivatives. *Canad. J. of Statist.* **16**, 411-418.

Villegas, C. (1969). On the a priori distribution of the covariance matrix. *Ann. Math. Statist.* **40**, 1098-1099.

Ye, K.Y. and Berger, J. (1991). Noninformative priors for inference in exponential regression models. *Biometrika* **78**, 645-656.