Bayesian Inference for a Covariance Matrix

Tom Leonard; John S. J. Hsu

The Annals of Statistics, Vol. 20, No. 4. (Dec., 1992), pp. 1669-1696.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199212%2920%3A4%3C1669%3ABIFACM%3E2.0.CO%3B2-H

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

The Annals of Statistics
1992, Vol. 20, No. 4, 1669–1696

BAYESIAN INFERENCE FOR A COVARIANCE MATRIX

BY TOM LEONARD AND JOHN S. J. HSU

University of Wisconsin, Madison, and University of California, Santa Barbara

A flexible class of prior distributions is proposed, for the covariance matrix of a multivariate normal distribution, yielding much more general hierarchical and empirical Bayes smoothing and inference, when compared with a conjugate analysis involving an inverted Wishart distribution. A likelihood approximation is obtained for the matrix logarithm of the covariance matrix, via Bellman's iterative solution to a Volterra integral equation. Exact and approximate Bayesian, empirical and hierarchical Bayesian estimation and finite sample inference techniques are developed. Some risk and asymptotic frequency properties are investigated. A subset of the Project Talent American High School data is analyzed. Applications and extensions to multivariate analysis, including a generalized linear model for covariance matrices, are indicated.

1. Sampling and prior assumptions. Initially consider n observation vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$, which given their common covariance matrix $\mathbf{C}$, are a random sample from a p-dimensional multivariate normal distribution, with zero mean vector and covariance matrix $\mathbf{C}$. In Sections 3 and 4, the statistical problem is addressed of obtaining estimators for $\mathbf{C}$ which sensibly smooth the sample covariance matrix $\mathbf{S} = \Sigma \mathbf{y}_i \mathbf{y}_i^T / n$. Assume that $p < n$ and assume occurrence of the almost sure event that $\mathbf{S}$ is observed to be positive definite. In the current section, broad families of probability measures are developed which may be associated with $(\mathscr{D}_p, \mathscr{A}_p)$, where $\mathscr{D}_p$ is the class of all positive definite (symmetric) $p \times p$ matrices, and $\mathscr{A}_p$ is the σ field of subsets of $\mathscr{D}_p$ defined below. These provide both flexible prior distributions for $\mathbf{C}$, which constrain $\mathbf{C}$ to lie in $\mathscr{D}_p$, and alternative sampling distributions for $\mathbf{S}$ when it is required to broaden the assumption of multivariate normality of the observation vectors. In the remainder of the paper it is demonstrated that these probability measures yield a broad spectrum of new applications for situations, where it is required to accurately investigate linear relationships between p variables. In particular, if $p \geq 3$, scalar multiples of the sample covariance matrix $\mathbf{S}$ should not always be regarded as the most reasonable estimators of $\mathbf{C}$, either in terms of their finite sample frequency properties [e.g., Efron and Morris (1976) and Haff (1980)] or in terms of smoothing the random fluctua-

Received January 1991; revised April 1992.

AMS 1980 *subject classifications*. Primary 62F15; secondary 62C12, 62E20, 62E25, 62F11, 62G05, 62J10, 62J12.

Key words and phrases. Multivariate normal distribution, covariance matrix, hierarchical prior, inverted Wishart prior, matrix exponential, intraclass hypothesis, Bayesian marginalization, generalized linear model, exchangeable distribution for a positive definite matrix.

tions of the data in a sensible manner; the corresponding correlation analysis is similarly open to improvement.

Evans (1965), Chen (1979), Dickey, Lindley and Press (1985) and Press (1992), consider prior distributions for $\mathbf{R} = \mathbf{C}^{-1}$ which, conditional on ν and $\mathbf{R}_0$, take $\mathbf{R}$ to belong to a conjugate Wishart family with ν degrees of freedom, mean matrix $\mathbf{R}_0 \in \mathscr{D}_p$ and density of the form

$$\pi(\mathbf{R}) \propto |\mathbf{R}|^{(\nu-p-1)/2} \exp\left\{-\tfrac{1}{2}\nu \operatorname{tr}\left(\mathbf{R}_0^{-1}\mathbf{R}\right)\right\} \qquad \mathbf{R} \in \mathscr{D}_p. \tag{1.1}$$

If ν and $\mathbf{R}_0$ are specified, this family may be much too restrictive, as there are just $q + 1$ distinct prior parameters, including $q = p(p + 1)/2$ prior estimates of the variances and covariances, just a single degree of belief ν for each of these estimates and no further parameters for modeling prior dependencies between the elements of $\mathbf{C}$. Mixtures of Wishart distributions are proposed by Dickey, Lindley and Press, and these assign further distributions to ν and simple parametric forms for $\mathbf{R}_0$ without substantially broadening the prior covariance structure. Under the conjugate prior distribution (1.1), the posterior distribution of $\mathbf{R}$ is Wishart with $\nu + n$ degrees of freedom, and the inverse of the posterior mean matrix of $\mathbf{R}$ assumes the simple weighted average from

$$\mathbf{C}^* = \left(n\mathbf{S} + \nu\mathbf{R}_0^{-1}\right)/(n + \nu) \tag{1.2}$$

providing shrinkages depending upon a single scalar weight $\nu/(\nu + n)$ with only very simple smoothing of the elements of $\mathbf{S}$. The posterior mean matrix of $\mathbf{C}$ is similarly simple.

These restrictions motivate consideration of the matrix logarithm $\mathbf{A} = \log \mathbf{C}$ of the covariance matrix $\mathbf{C}$. Consider the spectral decompositions

$$\mathbf{C} = \mathbf{E}\mathbf{D}\mathbf{E}^T \tag{1.3}$$

and

$$\mathbf{S} = \mathbf{E}_0\mathbf{D}_0\mathbf{E}_0^T, \tag{1.4}$$

where $\mathbf{D}$ and $\mathbf{D}_0$ are diagonal matrices of eigenvalues of $\mathbf{C}$ and $\mathbf{S}$, and the columns of the orthonormal matrices $\mathbf{E}$ and $\mathbf{E}_0$ are corresponding normalized eigenvectors. Then $\mathbf{A} = \log \mathbf{C}$ satisfies

$$\mathbf{C} = \exp\{\mathbf{A}\} = \sum_{r=0}^{\infty} \mathbf{A}^r/r!, \tag{1.5}$$

where $\exp\{\mathbf{A}\}$ denotes the matrix exponential of $\mathbf{A}$, or equivalently

$$\mathbf{A} = \mathbf{E}(\log \mathbf{D})\mathbf{E}^T, \tag{1.6}$$

where $\log \mathbf{D}$ denotes the diagonal matrix of log-eigenvalues of $\mathbf{C}$. Note that the upper triangular elements of the symmetric matrix $\mathbf{A}$ are unconstrained in q-dimensional real space R^P, and that $\mathbf{C} = \exp(\mathbf{A})$ exponentiates the eigenvalues of $\mathbf{A}$ leaving the normalized eigenvectors unchanged.

Let $\alpha = \text{vec}(\mathbf{A})$ denote the $q \times 1$ vector consisting of the upper triangular elements of $\mathbf{A}$, ordered by diagonal and subsequent off-diagonal, and taking the form

$$\begin{aligned}\boldsymbol{\alpha} = (&a_{11}, a_{22}, \ldots, a_{pp}; a_{12}, a_{23}, \ldots, a_{p-1,p};\\ &a_{13}, a_{14}, \ldots, a_{p-2,p}; \ldots; a_{1,p-1}, a_{2,p}; a_{1,p})^T.\end{aligned} \tag{1.7}$$

Consider any probability distribution for $\boldsymbol{\alpha}$ on (R^q, B^q), where B^q is the Borel field of subsets of R^q. Then this creates a probability distribution for $\mathbf{C}$ on $(\mathscr{D}_p, \mathscr{A}_p)$, which constrains $\mathbf{C}$ to fall in the class $\mathscr{D}_p$ of positive definite matrices, where the σ-field $\mathscr{A}_p$ may be conveniently defined as the collection of subsets of $\mathscr{D}_p$, which appropriately transform the elements of B^q via the matrix exponential transformation (1.5).

One possibility is to assume that $\boldsymbol{\alpha}$ has a multivariate normal distribution, say with mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Delta}$. This generalizes a class of priors proposed by Leonard (1975), when $\mathbf{C}$ is diagonal, which provided alternatives to the mixtures of inverted chi-squared distributions for the variances considered by Lindley (1971). For general $\mathbf{C}$, this specification provides a very flexible class of (nonconjugate) prior distributions, permitting the choice of $q(q+1)/2$ prior means, variances and covariances. It therefore permits subjective Bayesians to provide prior estimates for each element of $\boldsymbol{\alpha}$, a separate degree of belief corresponding to each estimate and a separate prior correlation between every pair of distinct elements of $\boldsymbol{\alpha}$. The Bayesian implications of this transformation to prior multivariate normality are discussed more fully in Section 3. In particular, a possible multivariate normal approximation to the posterior distribution of $\boldsymbol{\alpha}$ possesses mean vector $\boldsymbol{\xi}^*$ and covariance matrix $\boldsymbol{\Delta}^*$ satisfying (3.1) and (3.2).

Note that any rotations $\mathbf{Gy}_1, \ldots, \mathbf{Gy}_n$ of the observation vector will possess a common covariance matrix with matrix logarithm $\mathbf{GAG}^T$, which is a linear transformation of $\mathbf{A}$. The above prior specification transforming to multivariate normality is therefore closed under rotations of the observation vectors. This specification does not involve multivariate normality of the log-eigenvalues, the log-variances (except in the diagonal case) or the eigenvectors.

The presentation and discussion of results in Sections 2–9 is intended to demonstrate how a variety of statistical procedures can be developed, which depend upon properties of the transformation $\mathbf{A} = \log \mathbf{C}$, together with the choices of the prior mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Delta}$ of the multivariate normal prior. In Section 7, it is shown that all approximate Bayesian inference procedures discussed in Sections 2–6 can be replaced by the corresponding exact procedures. Asymptotics and frequency properties of the Bayesian procedures are discussed in Appendices 1 and 2. Broad classes of exchangeable distributions for positive definite matrices, which modify the exchangeable distributions for arrays considered by Aldous (1981), are proposed in Appendix 3. The co-authors and several colleagues have developed numerous useful extensions of the results obtained; these would appear to initiate a new research area in multivariate analysis.

2. Finite sample likelihood approximations. Since the log of the generalized variance $|\mathbf{C}|$ is equal to the trace of $\mathbf{A}$, the exact likelihood of $\mathbf{A} = \log \mathbf{C}$ given $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is

$$l(\mathbf{A}|\mathbf{y}) = (2\pi)^{-np/2} \exp\{-\tfrac{1}{2}n \operatorname{tr} \mathbf{A} - \tfrac{1}{2}n \operatorname{tr}[\mathbf{S} \exp\{-\mathbf{A}\}]\} \quad \mathbf{A} \in \mathscr{D}_p. \tag{2.1}$$

As the unique maximum likelihood estimate $\mathbf{S}$ of $\mathbf{C}$ is by assumption positive definite, the likelihood (2.1) is uniquely maximized when $\mathbf{A} = \mathbf{\Lambda}$, where

$$\mathbf{\Lambda} = \log \mathbf{S} = \mathbf{E}_0(\log \mathbf{D}_0)\mathbf{E}_0^T \tag{2.2}$$

with $\mathbf{S}$ defined in (1.4). A possible multivariate normal approximation to the likelihood of $\boldsymbol{\alpha} = \operatorname{vec}(\mathbf{A})$, is summarized by (2.12), together with (2.13), (2.14) and (2.11). This neglects the cubic and higher terms of a Taylor series expansion, about $\boldsymbol{\lambda} = \operatorname{vec}(\mathbf{\Lambda})$, of the log-likelihood of $\boldsymbol{\alpha}$. The Taylor series expansion is quite nontrivial, owing to the complicated nature of the second term within the exponential of (2.1). For example, Bellman [(1970), page 170] remarks that is it untrue that

$$\operatorname{tr}(\mathbf{S} \exp\{-\mathbf{A}\}) = \operatorname{tr}(\exp\{\mathbf{\Lambda} - \mathbf{A}\}) = \operatorname{tr}\left\{ \sum_{r=0}^{\infty} (\mathbf{\Lambda} - \mathbf{A})^r / r! \right\}, \tag{2.3}$$

unless $\mathbf{A}$ and $\mathbf{\Lambda}$ commute. However, Bellman [(1970), page 175] shows that

$$\exp\{-\mathbf{A}t\} = \mathbf{X}(t),$$

where $\mathbf{X}(t)$ satisfies the Volterra integral equation

$$\mathbf{X}(t) = \mathbf{S}^{-t} - \int_0^t \mathbf{S}^{s-t}(\mathbf{A} - \mathbf{\Lambda})\mathbf{X}(s)\, ds \qquad 0 < t < \infty. \tag{2.4}$$

A Taylor series expansion of $\mathbf{X}(t)$, about $\mathbf{A} = \mathbf{\Lambda}$, is now available by successively substituting the right-hand side of (2.4) for the $\mathbf{X}$ function in the integrand, yielding

$$\begin{aligned} \mathbf{C}^{-1} = \exp\{-\mathbf{A}\} &= \mathbf{X}(1) \\ &= \mathbf{S}^{-1} - \int_0^1 \mathbf{S}^{s-1}(\mathbf{A} - \mathbf{\Lambda})\mathbf{S}^{-s}\, ds \\ &\quad + \int_0^1 \int_0^s \mathbf{S}^{s-1}(\mathbf{A} - \mathbf{\Lambda})\mathbf{S}^{u-s}(\mathbf{A} - \mathbf{\Lambda})\mathbf{S}^{-u}\, du\, ds \\ &\quad + \text{cubic and higher terms} \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} \operatorname{tr}(\mathbf{S} \exp\{-\mathbf{A}\}) &= p - \operatorname{tr}(\mathbf{A} - \mathbf{\Lambda}) \\ &\quad + \int_0^1 \int_0^s \operatorname{tr}\left[(\mathbf{A} - \mathbf{\Lambda})\mathbf{S}^{u-s}(\mathbf{A} - \mathbf{\Lambda})\mathbf{S}^{-(u-s)}\right] du\, ds \\ &\quad + \text{cubic and higher terms.} \end{aligned} \tag{2.6}$$

The integrations in (2.6) can be completed analytically, using the spectral decomposition (1.4), since

$$\text{(2.7)} \qquad \operatorname{tr}\left[(\mathbf{A}-\boldsymbol{\Lambda})\mathbf{S}^{u-s}(\mathbf{A}-\boldsymbol{\Lambda})\mathbf{S}^{-(u-s)}\right] = \operatorname{tr}\left[\mathbf{B}\mathbf{D}_0^{u-s}\mathbf{B}\mathbf{D}_0^{-(u-s)}\right],$$

with

$$\text{(2.8)} \qquad \mathbf{B} = \mathbf{E}_0^T(\mathbf{A}-\boldsymbol{\Lambda})\mathbf{E}_0 = \mathbf{E}_0^T\mathbf{A}\mathbf{E}_0 - \log \mathbf{D}_0.$$

If the cubic and higher terms in (2.6) are neglected, then substitution of the remaining terms for trace $[\mathbf{S}\exp\{-\mathbf{A}\}]$ in (2.1) yields the expression

$$\text{(2.9)} \qquad l^*(\mathbf{A}|\mathbf{y}) = (2\pi)^{-np/2}e^{-np}|\mathbf{S}|^{-n/2}\exp\left\{-\tfrac{1}{4}n\sum_i b_{ii}^2 - \tfrac{1}{2}n\sum_{i<j}\xi_{ij}b_{ij}^2\right\},$$

which is open to justification as an approximation to the likelihood (2.1), where

$$\text{(2.10)} \qquad b_{ij} = \mathbf{e}_i^T(\mathbf{A}-\boldsymbol{\Lambda})\mathbf{e}_j$$

is the (i, j)th element of the matrix $\mathbf{B}$ in (2.8), and

$$\text{(2.11)} \qquad \xi_{ij} = \frac{d_{ii}/d_{jj} + d_{jj}/d_{ii} - 2}{(\log d_{ii} - \log d_{jj})^2},$$

with d_{ii} denoting the ith eigenvalue, and $\mathbf{e}_i$ denoting the ith normalized eigenvector of the sample covariance matrix $\mathbf{S}$ in (1.4).

The expression in (2.9) is related to the scalar result ($p = 1$), that the likelihood of a normal (zero mean) log-variance is approximately normal with location equal to the log of the sample variance and dispersion $2n^{-1}$. The scalar result relates to the normal approximation to the density of the log of a chi-squared variate reported by Bartlett and Kendall (1946), which they show to be surprisingly accurate even in the tails if $n \geq 10$. Note that (2.9) is exactly equivalent to the possible approximation

$$\text{(2.12)} \qquad l^*(\boldsymbol{\alpha}|\mathbf{y}) = (2\pi)^{-n/2}e^{-np}|\mathbf{S}|^{-n/2}\exp\left\{-\tfrac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\lambda})^T\mathbf{Q}(\boldsymbol{\alpha}-\boldsymbol{\lambda})\right\} \qquad \boldsymbol{\alpha} \in R^q$$

to the likelihood of $\boldsymbol{\alpha} = \operatorname{vec}(\mathbf{A})$, where

$$\text{(2.13)} \qquad \mathbf{Q} = \tfrac{1}{2}n\sum_i \mathbf{f}_{ii}\mathbf{f}_{ii}^T + n\sum_{i,j:\,i<j}\xi_{ij}\mathbf{f}_{ij}\mathbf{f}_{ij}^T,$$

with

$$\text{(2.14)} \qquad \mathbf{f}_{ij} = \mathbf{e}_i * \mathbf{e}_j$$

and the product $\mathbf{e}_i * \mathbf{e}_j$ denoting the $q \times 1$ vector satisfying

$$\text{(2.15)} \qquad \boldsymbol{\alpha}^T(\mathbf{e}_i * \mathbf{e}_j) = \mathbf{e}_i^T\mathbf{A}\mathbf{e}_j$$

for all possible realizations of $\mathbf{A}$. The positive definite matrix $\mathbf{Q}$ is the likelihood information matrix of $\boldsymbol{\alpha}$ (see Appendix 1).

It is important to justify the expression in (2.12) as a *likelihood approximation* when n is finite. Note that, the integral over $\mathbf{A} \in \mathscr{D}_p$ of the right-hand

side of (2.1) is finite. (See Appendix 1.) Therefore (2.1) also provides the exact posterior density of $\boldsymbol{\alpha} = \text{vec}(\mathbf{A})$ when $\boldsymbol{\alpha}$ possesses an improper prior distribution, which is uniform over R^q. Under the possible approximation (2.12), this posterior density is replaced by a multivariate normal density with mean vector $\boldsymbol{\lambda}$ and covariance matrix $\mathbf{Q}^{-1}$. The possible likelihood approximation (2.12) can therefore be validated by computing the exact posterior density of any linear combination $\eta = \mathbf{h}^T\boldsymbol{\alpha}$ of $\boldsymbol{\alpha}$, and comparing this with a normal density with mean $\mathbf{h}^T\boldsymbol{\lambda}$ and variance $\mathbf{h}^T\mathbf{Q}^{-1}\mathbf{h}$. Computational methods for this exact density are described in Section 7. It is now assumed for presentation purposes that (2.12) provides an adequate likelihood approximation for the particular data set under analysis. (See, e.g., the data example in Section 9.) The above methods also permit the computation of the exact posterior density of any continuous function $\eta = g(\mathbf{C})$ of $\mathbf{C}$ under a uniform prior for $\boldsymbol{\alpha}$, for example, correlations, eigenvalues, $\text{tr}(\mathbf{C})$ and the generalized variance $|\mathbf{C}| = \exp\{\text{tr}(\mathbf{A})\}$, where $\text{tr}(\mathbf{A})$ is the sum of the first p elements of $\boldsymbol{\alpha}$. The exact posterior density for $|\mathbf{C}|$ yields alternative inferences to the classical procedures relating to the normal approximation to the distribution of $|\mathbf{S}|$ described by Anderson [(1984), page 173]. Under approximation (2.12), the posterior density of $|\mathbf{C}|$ is log-normal.

As a finite sample refinement to (2.12), consider the expression

$$\tilde{l}(\boldsymbol{\alpha}|\mathbf{y}) = l(\tilde{\boldsymbol{\lambda}}|\mathbf{y})\exp\left\{-\tfrac{1}{2}(\boldsymbol{\alpha}-\tilde{\boldsymbol{\lambda}})^T\tilde{\mathbf{Q}}(\boldsymbol{\alpha}-\tilde{\boldsymbol{\lambda}})\right\}, \tag{2.16}$$

where the exact likelihood $l(\boldsymbol{\alpha}|\mathbf{y})$ may be obtained from (2.1), and where $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\mathbf{Q}}^{-1}$ denote the exact unconditional posterior mean vector and covariance matrix of $\boldsymbol{\alpha}$ (assuming these exist) under a uniform prior, calculated via the procedures described in Section 7. Then $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\mathbf{Q}}$ minimize, for fixed $\mathbf{y}_1, \ldots, \mathbf{y}_n$, the entropy distance

$$\text{ED}(\pi, \tilde{\pi}) = \int_{R^q} \log[\pi(\boldsymbol{\alpha}|\mathbf{y})/\tilde{\pi}(\boldsymbol{\alpha}|\mathbf{y})]\pi(\boldsymbol{\alpha}|\mathbf{y})\, d\boldsymbol{\alpha}, \tag{2.17}$$

whenever this is finite, between the exact posterior density $\pi(\boldsymbol{\alpha}|\mathbf{y})$ under a uniform prior for $\boldsymbol{\alpha}$ and a multivariate normal density $\tilde{\pi}(\boldsymbol{\alpha}|\mathbf{y})$ with mean vector $\tilde{\boldsymbol{\lambda}}$ and covariance matrix $\tilde{\mathbf{Q}}^{-1}$. Note that (2.17) yields a rigorous finite sample justification of (2.16) as an approximation to the exact likelihood of $\boldsymbol{\alpha}$. Asymptotic properties, as $n \to \infty$, related to likelihood approximation (2.12) are described in Appendix 1. In particular, a multivariate normal approximation is described to the distribution of $\boldsymbol{\lambda} = \text{vec}(\boldsymbol{\Lambda})$ when $\boldsymbol{\Lambda}$ is the matrix logarithm of a random Wishart matrix $\mathbf{S}$.

3. The flexibility of the prior assumptions.

3.1. *Duality with the posterior smoothing process.* Suppose that $\boldsymbol{\alpha} = \text{vec}(\mathbf{A})$ possesses a multivariate normal prior distribution with mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Delta}$. Under the likelihood approximation (2.12), the posterior density of $\boldsymbol{\alpha}$ is $\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}^*, \boldsymbol{\Delta}^*)$, a multivariate normal density with mean

vector

$$\boldsymbol{\xi}^* = (\mathbf{Q} + \boldsymbol{\Delta}^{-1})^{-1}(\mathbf{Q}\boldsymbol{\lambda} + \boldsymbol{\Delta}^{-1}\boldsymbol{\xi}) \tag{3.1}$$

and covariance matrix

$$\boldsymbol{\Delta}^* = (\mathbf{Q} + \boldsymbol{\Delta}^{-1})^{-1}, \tag{3.2}$$

where $\boldsymbol{\lambda}$ and $\mathbf{Q}$ are, respectively, the maximum likelihood vector and observed information matrix of $\boldsymbol{\alpha}$. Some frequency risk properties of (3.1) and related estimators are discussed in Appendix 2. As the posterior vector (3.1) assumes the form of a matrix weighted average of $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$, it can provide much more complex smoothing of λ when compared with the simple scalar shrinkages of $\mathbf{S}$ suggested by the estimator (1.2) for $\mathbf{C}$ under a conjugate inverted Wishart prior distribution. As discussed by Chamberlain and Leamer (1976), the complex posterior smoothing is very dependent upon the prior choices of $\boldsymbol{\xi}$ and $\boldsymbol{\Delta}$. Note, for example, that the sum of the first p elements of (3.1) provides a Bayesian estimate for $|\mathbf{C}|$ which is not a simple weighted average of $\log|\mathbf{S}|$ and the prior mean of $\log|\mathbf{C}|$.

3.2. *An exchangeable distribution for a positive definite matrix.* Suppose that $\mathbf{C}$ is a priori *exchangeable*, that is, its prior distribution is invariant under any permutation of the rows of $\mathbf{C}$ together with the same permutation of the columns. Then take $\boldsymbol{\alpha}$ to possess, conditionally upon $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, a multivariate normal distribution with mean vector $\mathbf{X}\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Delta}$, where $\mathbf{X}$ denotes a $q \times 2$ matrix with unit entries in the first p rows of the first column, and the last $q - p$ rows of the second column and zeros elsewhere. Furthermore, let $\boldsymbol{\Delta}$ be diagonal with the first p diagonal elements equal to σ_1^2, and the remaining diagonal elements equal to σ_2^2. This specification assigns a common prior estimate μ_1 to each diagonal element of $\mathbf{A}$, together with a common prior estimate μ_2 for each nondiagonal element. It is possible to express different degrees of belief σ_1^{-2} and σ_2^{-2} in these two prior estimates, therefore providing a more general prior specification for $\mathbf{C}$ than the exchangeable prior proposed by Chen (1979) and Dickey, Lindley and Press (1985), who take $\mathbf{C}^{-1}$ to possess a Wishart distribution, conditional on a mean matrix assuming intraclass form and the degrees of freedom ν. Very general formulations of an exchangeable distribution for a covariance matrix are indicated in Appendix 3.

If $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ is unknown, then the above assumptions may be extended by taking $\boldsymbol{\mu}$ to be uniformly distributed over R^2. Under the likelihood approximation (2.12), the posterior distribution of $\boldsymbol{\mu}$ given σ_1^2 and σ_2^2 is now bivariate normal with mean vector

$$\boldsymbol{\mu}^* = \left[\mathbf{X}^T(\mathbf{Q}^{-1} + \boldsymbol{\Delta})^{-1}\mathbf{X}\right]^{-1}\mathbf{X}^T(\mathbf{Q}^{-1} + \boldsymbol{\Delta})^{-1}\boldsymbol{\lambda}, \tag{3.3}$$

and covariance matrix

$$\mathbf{G} = \left[\mathbf{X}^T(\mathbf{Q}^{-1} + \boldsymbol{\Delta})^{-1}\mathbf{X}\right]^{-1}. \tag{3.4}$$

The corresponding distribution of $\boldsymbol{\alpha}$ is multivariate normal with mean vector

$$\boldsymbol{\alpha}^* = (\mathbf{Q} + \boldsymbol{\Delta}^{-1})^{-1}(\mathbf{Q}\boldsymbol{\lambda} + \boldsymbol{\Delta}^{-1}\mathbf{X}\boldsymbol{\mu}^*) \tag{3.5}$$

and covariance matrix

$$\tilde{\boldsymbol{\Delta}} = \boldsymbol{\Delta}^* + \boldsymbol{\Delta}^*\boldsymbol{\Delta}^{-1}\mathbf{X}\mathbf{G}\mathbf{X}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}^*, \tag{3.6}$$

where $\boldsymbol{\Delta}^*$ is defined in (3.2). The posterior smoothing of $\boldsymbol{\lambda}$ provided by (3.5) now depends upon the specification of two "smoothing parameters" σ_1^2 and σ_2^2. However, these prior parameters may themselves be identified from the current data (see Section 4). They measure the closeness of the corresponding estimate for $\mathbf{A}$ to intraclass form. When σ_1^2 and σ_2^2 are specified, the exact posterior distribution will always remain proper, because the posterior density of $\boldsymbol{\alpha}$ is proportional to the product of the likelihood and a nonnegative function whose maximum value is finite and because the posterior distribution under a uniform prior for $\boldsymbol{\alpha}$ is proper.

3.3. *Representing uncertainty in a hypothesis of diagonality.* Instead of assuming a uniform distribution for $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ in the assessment of the preceding exchangeable prior distribution, it might sometimes be appropriate to let $\sigma_1^2 \to \infty$ and set $\mu_2 = 0$, in which case μ_1 need not be specified. This represents prior information that the statistician thinks that $\mathbf{A}$ and $\mathbf{C}$ may be diagonal, but that he is unsure regarding his hypothesis. The remaining prior parameter $\sigma_2^2 = \sigma^2$ measures the uncertainty in the diagonal hypothesis. Under likelihood approximation (2.12), the posterior distribution of $\boldsymbol{\alpha}$ given σ^2 is now multivariate normal with mean vector

$$\boldsymbol{\alpha}^* = (\mathbf{Q} + \mathbf{H})^{-1}\mathbf{Q}\boldsymbol{\lambda}, \tag{3.7}$$

and covariance matrix

$$\boldsymbol{\Delta}^* = (\mathbf{Q} + \mathbf{H})^{-1}, \tag{3.8}$$

where $\mathbf{H}$ is the $q \times q$ matrix with last $q - p$ diagonal elements equal to σ^{-2} and all other elements equal to zero.

3.4. *Some useful features.* The above prior formulations possess the possible disadvantage that $\mathbf{A} = \log \mathbf{C}$ does not contain elements which possess obvious meaning; thus creating obstacles to the formulation of prior opinions regarding $\mathbf{A}$. They however provide Bayesians with the following three useful features:

1. The exchangeable prior distribution, previously described, which should frequently be completely meaningful [either for $\mathbf{C}$ or some rotation $\mathbf{G}\mathbf{C}\mathbf{G}^T$ of $\mathbf{C}$ with $\log(\mathbf{G}\mathbf{C}\mathbf{G}^T) = \mathbf{G}\mathbf{A}\mathbf{G}^T$] unless it is necessary to subjectively assess the prior variances σ_1^2 and σ_2^2.
2. For a general multivariate normal prior for $\boldsymbol{\alpha}$, with mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Delta}$, availability of exact calculations via Monte Carlo

simulations for the prior density and moments of any element of $\mathbf{C}$, for the bivariate prior density of any two elements or for the prior density of any scalar or vector function $\eta = g(\mathbf{C})$. A subjective Bayesian can therefore construct and consider different choices of ξ and Δ by comparing corresponding densities and joint densities for elements of $\mathbf{C}$, using modern software for computer graphics. Alternatively, a specified prior distribution for $\mathbf{C}$ can be converted, via Monte Carlo simulations, into a prior distribution for α.

3. A prior dependency structure, which may be more useful than the specification of prior correlations between elements of $\mathbf{C}$, since the latter cannot always be regarded as linearly related. The normalizing transformation $\mathbf{A} = \log \mathbf{C}$ permits the representation of relationships between the parameters by prior correlations between elements of $\mathbf{A}$. The elements of $\mathbf{A}$ would appear (consider for example the special case when $\mathbf{C}$ is diagonal) to be quite amenable to linear prior relationships. This parallels ideas discussed by Leonard (1973) who considered prior correlations between multivariate logits, transforming the probabilities in a histogram.

3.5. *Several parallel time series.* Suppose that the elements of $\mathbf{y}_1, \ldots, \mathbf{y}_n$ represent n parallel time series, each observed at p time points. Then the prior mean vector $\xi = \xi(\mu)$ might be specified by $\xi = \text{vec}(\mathbf{A}_0)$ with $\mathbf{A}_0 = \log \mathbf{C}_0$, where $\mathbf{C}_0$ represents a common hypothesized covariance structure for the $\mathbf{y}_i$ (e.g., the covariance matrix of a first order autoregressive process) and the elements of μ are unknown parameters appearing in this covariance structure, for example, the correlation and variance parameters appearing in the first order autoregressive covariance function. The prior covariance matrix $\Delta = \Delta(\zeta)$ of α can be specified by assigning sensible prior correlations (e.g., based upon autocorrelations from a spatial process) to pairs of elements of α (i.e., the upper triangular elements of $\mathbf{A}$), depending upon parameters ζ. This example illustrates the immense flexibility of the prior assumptions.

3.6. *Prior structures with unknown prior parameters.* As a generalization of the preceding multivariate normal prior distribution for α, for situations where it is difficult to fully specify the mean vector and covariance matrix of this distribution, the user can consider the hierarchical prior distribution for α described in the following two stages:

STAGE 1. Given μ and ξ, α possesses a multivariate normal prior distribution with mean vector

$$\xi = \xi(\mu)$$

and covariance matrix

$$\Delta = \Delta(\zeta),$$

where μ and ζ are unknown $l_1 \times 1$ and $l_2 \times 1$ vectors and $\xi(\cdot)$ and $\Delta(\cdot)$ are specified "prior structures." The dimensions l_1 and l_2 should be taken to be small when respectively compared with p and q.

STAGE 2. $\boldsymbol{\mu}$ and $\boldsymbol{\zeta}$ possess density $\pi(\boldsymbol{\mu}, \boldsymbol{\zeta})$ for $\boldsymbol{\mu} \in R^{l_1}$ and $\boldsymbol{\zeta} \in R^{l_2}$.

This hierarchical distribution provides a very broad paradigm. Some further special cases of prior structures, which can be employed at the first stage of the distribution, are:

1. The generalized linear model

$$\boldsymbol{\xi} = \mathbf{X}\boldsymbol{\mu} \tag{3.9}$$

and

$$\log \boldsymbol{\Delta} = \zeta_1 \mathbf{W}_1 + \cdots + \zeta_{l_2} \mathbf{W}_{l_2}, \tag{3.10}$$

where $\mathbf{X}$ is a specified $q \times l_1$ matrix, and the $\mathbf{W}_i$ are specified $q \times q$ matrices.

2. With $\boldsymbol{\xi} = \operatorname{vec}(\mathbf{A}_0)$, with $\mathbf{A}_0 = \log \mathbf{C}_0$ and $\mathbf{C}_0$, several meaningful choices of $\mathbf{C}_0 = \mathbf{C}_0(\boldsymbol{\zeta})$ are described by Chen (1979). In particular, special forms for $\mathbf{C}$ occurring in factor analysis, or structural equation models or relating to an assumption of equal eigenvalues, can be incorporated into $\mathbf{C}_0$ as prior structures or "prior hypotheses" for $\mathbf{C}$.

4. Hierarchical and empirical Bayes procedures. Consider, for simplicity, the special case of the hierarchical prior assumptions of Section 3, where $\boldsymbol{\xi}$ satisfies (3.11), $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\zeta})$, $\boldsymbol{\mu}$ is uniformly distributed over R^{l_1} and independent of $\boldsymbol{\zeta}$, and $\boldsymbol{\zeta}$ possesses prior density $\pi(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta} \in R^{l_2}$. Under the likelihood approximation (2.12), the posterior distribution of $\boldsymbol{\alpha}$, given $\boldsymbol{\Delta}$ is now multivariate normal with mean vector $\boldsymbol{\alpha}^*$ and covariance matrix $\tilde{\boldsymbol{\Delta}}$ defined in (3.7) and (3.8), respectively. Moreover, the posterior density of $\boldsymbol{\zeta}$ is

$$\pi^*(\boldsymbol{\zeta}|\mathbf{y}) \propto \pi(\boldsymbol{\zeta}) l^*(\boldsymbol{\zeta}|\mathbf{y}) \qquad \boldsymbol{\zeta} \in R^{l_2}, \tag{4.1}$$

where the "integrated likelihood" contribution to (4.1) is

$$l^*(\boldsymbol{\zeta}|\mathbf{y}) \propto |\mathbf{Q}^{-1} + \boldsymbol{\Delta}|^{-1/2} |\mathbf{G}|^{1/2} \exp\left\{-\tfrac{1}{2}\boldsymbol{\lambda}^T \mathbf{Q}^* \boldsymbol{\lambda}\right\} \qquad \boldsymbol{\zeta} \in R^{l_2}, \tag{4.2}$$

where $\mathbf{G}$ is defined in (3.6), $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\zeta})$ and

$$\mathbf{Q}^* = (\mathbf{Q}^{-1} + \boldsymbol{\Delta})^{-1} \mathbf{X}\mathbf{G}\mathbf{X}^T (\mathbf{Q}^{-1} + \boldsymbol{\Delta})^{-1}. \tag{4.3}$$

Hence, the posterior distribution and moments of $\boldsymbol{\alpha}$, unconditional upon $\boldsymbol{\zeta}$, may be computed by appropriate numerical integrations of this distribution, conditional on $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\zeta})$ with respect to the posterior density (4.1). Note that, if $\pi_1(\boldsymbol{\zeta}|\mathbf{y})$ and $\pi_2(\boldsymbol{\zeta}|\mathbf{y})$ denote the posterior densities of $\boldsymbol{\zeta}$ under two different prior assessments for $\boldsymbol{\zeta}$, then the unconditional posterior densities $\pi_1(\eta|\mathbf{y})$ and $\pi_2(\eta|\mathbf{y})$ of any parameter of interest η, which is a function of $\boldsymbol{\alpha}$, satisfy

$$\int |\pi_1(\eta|\mathbf{y}) - \pi_2(\eta|\mathbf{y})| \, d\eta \leq \int |\pi_1(\boldsymbol{\zeta}|\mathbf{y}) - \pi_2(\boldsymbol{\zeta}|\mathbf{y})| \, d\boldsymbol{\zeta}, \tag{4.4}$$

where the integrations should be taken over all possible realizations of η and $\boldsymbol{\zeta}$. The inequality (4.4) indicates that the hierarchical Bayes procedures are reasonably robust under modest changes in the prior assessment for $\boldsymbol{\zeta}$.

If, however, an improper distribution is chosen for ζ in the prior assessment, it should be carefully checked, for example, using the importance sampling techniques of Section 7, that the integral, over $\zeta \in R^l$ of the expression on the right-hand side of (4.2) is finite. For example, when Δ depends upon two prior parameters σ_1^2 and σ_2^2, and follows the structure assumed for the exchangeable prior distribution in Section 4, log-uniform distributions for σ_1^2 and σ_2^2 would invariably lead to an improper posterior distribution. However, in particular numerical examples with $p \geq 4$, uniform distributions for σ_1^2 and σ_2^2 in the prior assessment, can lead to a proper posterior distribution. Strawderman (1971) and Efron and Morris (1973) show that a uniform distribution for a first stage variance of a prior distribution can lead to estimators for normal means with excellent mean squared error properties.

If, following Dempster, Laird and Rubin (1977), Chen (1979), Haff (1980) and Morris (1983), an empirical Bayes approach is required, avoiding some of the complications of a full hierarchical Bayes approach, the integrated likelihood (4.2) can be maximized numerically providing an (approximate) integrated likelihood estimate $\hat{\zeta}$ for ζ. Then $\hat{\zeta}$ can be substituted for ζ in the approximate multivariate normal posterior distribution for α, given ζ, providing empirical Bayes estimates and empirical Bayes intervals for elements of α. For further discussion of this device involving marginal modes, but in a Bayesian context, see O'Hagan (1976).

5. Investigating the intraclass hypothesis. It is possible to investigate the hypothesis that $\mathbf{A}$, and hence $\mathbf{C}$, follows intraclass form by considering the posterior distribution of the "parametric residual" vector

$$\rho = \alpha - \mathbf{X}\mu, \tag{5.1}$$

where $\mu = (\mu_1, \mu_2)^T$, and $\mathbf{X}$ assumes the special structure introduced in the fourth paragraph of Section 3. Note that $\mathbf{X}\mu$ represents our hypothesized model, and that the elements of ρ measure individual deviations from this hypothesis. It is similarly possible to represent deviations from other hypothesized structures for $\mathbf{C}$, for example, the diagonal hypothesis introduced in Section 3.

Given σ_1^2 and σ_2^2, the likelihood approximation (2.12) and the prior assumptions of Section 3, the posterior distribution of ρ is multivariate normal with mean vector

$$\mu_\rho = (\mathbf{Q} + \Delta^{-1})^{-1}\mathbf{Q}(\lambda - \mathbf{X}\mu^*) \tag{5.2}$$

with μ^* defined in (3.5), and covariance matrix

$$\Delta_\rho = \Delta^* + \Delta^*\mathbf{Q}\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{Q}\Delta^* \tag{5.3}$$

with $\mathbf{G}$ and Δ^* described in (3.6) and (3.10). Therefore, the unconditional posterior density, of any element of ρ, can be approximated by averaging its univariate normal density, given σ_1^2 and σ_2^2, with respect to the density (4.1), but with $\zeta = (\sigma_1^2, \sigma_2^2)^T$. This permits the statistician to infer whether any individual residual is substantially different from zero. See Leonard and Novick (1986) and Albert (1988) for fuller discussions of this type of procedure.

Following Cohen (1974), Leonard and Ord (1976) and Leonard (1977), an overall model check may be developed by considering a preliminary test estimator for $\boldsymbol{\alpha}$ of the form

$$\tilde{\boldsymbol{\alpha}} = \begin{cases} \boldsymbol{\lambda}, & \text{for } \boldsymbol{\lambda} \in C_R, \\ \mathbf{X}\tilde{\boldsymbol{\mu}}, & \text{for } \boldsymbol{\lambda} \notin C_R, \end{cases} \tag{5.4}$$

for some 2×1 vector $\tilde{\boldsymbol{\mu}}$ and critical region C_R. For $\boldsymbol{\lambda} \notin C_R$, $\tilde{\boldsymbol{\alpha}}$ corresponds to choices of $\mathbf{A}$ and $\mathbf{C}$ assuming intraclass form. If $\lambda \in \mathbf{C}_R$, then $\tilde{\boldsymbol{\alpha}}$ represents a general unconstrained alternative hypothesis, corresponding to the choice of $\mathbf{S}$ to estimate $\mathbf{C}$. The Bayes choice of the critical region C_R under the subclass (5.4) of estimators, and the quadratic loss function

$$L(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) = (\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^T(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}), \tag{5.5}$$

for estimators $\tilde{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is straightforward to evaluate. Let $\boldsymbol{\alpha}^*$ denote the unconditional posterior mean vector of $\boldsymbol{\alpha}$, and assume that the posterior covariance matrix of $\boldsymbol{\alpha}$ exists. Then the optimal choices of $\tilde{\boldsymbol{\mu}}$ and C_R are

$$\tilde{\boldsymbol{\mu}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\alpha}^* \tag{5.6}$$

and

$$C_R = \{\boldsymbol{\lambda} : T(\boldsymbol{\lambda}) \geq 1\}, \tag{5.7}$$

where

$$T(\boldsymbol{\lambda}) = \boldsymbol{\alpha}^{*T}\left[\mathbf{I}_q - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]\boldsymbol{\alpha}^*/(\boldsymbol{\alpha}^* - \boldsymbol{\lambda})^T(\boldsymbol{\alpha}^* - \boldsymbol{\lambda}). \tag{5.8}$$

In particular, a high value for the statistic in (5.8) discredits the null hypothesis. Similar methods may be used to investigate the fit of the generalized linear model proposed in Section 9.

6. Incorporating prior information for the mean vector. Suppose now that $\mathbf{y}_1, \ldots, \mathbf{y}_n$ constitute a random sample from a multivariate normal distribution with unknown mean vector $\boldsymbol{\theta}$ and covariance matrix $\mathbf{C}$, so that the likelihood of $\boldsymbol{\theta}$ and $\mathbf{C}$ is

$$l(\boldsymbol{\theta}, \mathbf{C}|\mathbf{y}) = (2\pi)^{-n/2}|\mathbf{C}|^{-n/2}\exp\left\{-\tfrac{1}{2}\operatorname{tr}\left[\left\{\mathbf{U} + n(\boldsymbol{\theta} - \bar{\mathbf{y}})(\boldsymbol{\theta} - \bar{\mathbf{y}})^T\right\}\mathbf{C}^{-1}\right]\right\} \tag{6.1}$$
$$\boldsymbol{\theta} \in \mathbf{R}^P, \mathbf{C} \in \mathscr{D}^P,$$

where $\mathbf{U} = \Sigma(\mathbf{y}_i - \bar{\mathbf{y}})(y_i - \bar{\mathbf{y}})^T$ and $\bar{\mathbf{y}}$ is sample mean vector. Assume that $\mathbf{U}$ is observed to be positive definite.

The usual conjugate prior distribution for $\boldsymbol{\theta}$ and $\mathbf{C}$ involves a rather restrictive assumption that the conditional distribution of $\boldsymbol{\theta}$, given $\mathbf{C}$, is multivariate normal with covariance matrix equal to a scalar multiple of $\mathbf{C}$. Let $\boldsymbol{\mu}_0$ denote the prior mean vector of $\boldsymbol{\theta}$, and $\tau^{-1}\mathbf{C}$ denote the conditional covariance matrix of $\boldsymbol{\theta}$, given $\mathbf{C}$. If $\mathbf{R} = \mathbf{C}^{-1}$ has a Wishart distribution with ν degrees of freedom, in the prior assessment, with mean matrix $\mathbf{R}_0$, then the corresponding posterior distribution of $\mathbf{R}$ is Wishart with $\nu + n$ degrees of

freedom, and the inverse of the posterior mean matrix of $\mathbf{R}$ is

$$\mathbf{C}^* = \left[\nu\mathbf{R}_0^{-1} + \mathbf{U}^*\right]/(\nu + n), \tag{6.2}$$

where

$$n^{-1}\mathbf{U}^* = n^{-1}\mathbf{U} + \rho(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^T, \tag{6.3}$$

with $\rho = \tau/(\tau + n)$. The estimator in (6.2) again involves scalar, rather than matrix, shrinkage factors $\nu/(\nu + n)$, and $\rho = \tau/(\tau + n)$. It depends upon $\bar{\mathbf{y}}$ together with $\mathbf{U}$, following recommendations by Stein (1975). The second term on the right-hand side of (6.3) causes $n^{-1}\mathbf{U}^*$ to expand the maximum likelihood estimate $n^{-1}\mathbf{U}$ of $\mathbf{C}$, owing to prior information about $\boldsymbol{\theta}$. Then (6.2) shrinks the expanded maximum likelihood estimator $n^{-1}\mathbf{U}^*$ towards $\mathbf{R}_0^{-1}$, to take account of the prior information about $\mathbf{C}$. This apparently disconcerting *phenomenon of the expansion of the Bayes estimate of the covariance matrix* can be readily explained by noticing that

$$E(\boldsymbol{\theta}|\mathbf{y}) = (1 - \rho)\bar{\mathbf{y}} + \rho\boldsymbol{\mu}_0 \tag{6.4}$$

and

$$\mathbf{U}^* = (1 - \rho)\mathbf{U} + \rho\mathbf{U}_\theta, \tag{6.5}$$

where

$$\mathbf{U}_\theta = \sum_i (\mathbf{y}_i - \boldsymbol{\mu}_\theta)(\mathbf{y}_i - \boldsymbol{\mu}_\theta)^T. \tag{6.6}$$

Therefore, $n^{-1}\mathbf{U}^*$ compromises between the maximum likelihood estimate of $\mathbf{C}$ when $\boldsymbol{\theta}$ is unknown, and the maximum likelihood estimate of $\mathbf{C}$ when $\boldsymbol{\theta}$ is known to be $\boldsymbol{\mu}_\theta$. The shrinkage proportion $\rho = \tau/(\tau + n)$ is the same as the corresponding proportion in (6.4) for the posterior mean vector of $\boldsymbol{\theta}$.

Consider instead the following general prior formulation:

1. Given $\mathbf{C}$, $\boldsymbol{\theta}$ possesses a multivariate normal prior distribution with mean vector $\boldsymbol{\mu}_\theta$ and covariance matrix $\tau^{-1}\mathbf{F}^T\mathbf{C}\mathbf{F}$, where $\mathbf{F}$ is a $p \times p$ orthonormal matrix, and $\boldsymbol{\mu}_\theta$, τ and $\mathbf{F}$ are specified.
2. With $\mathbf{A} = \log\mathbf{C}$, $\boldsymbol{\alpha} = \mathrm{vec}(\mathbf{A})$ possesses a multivariate normal distribution, with mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Delta}$.

Then the joint distribution of $\mathbf{y}_1, \ldots, \mathbf{y}_n$, conditional upon $\mathbf{A}$ yields an "integrated likelihood" for $\mathbf{A}$, which may be expressed in the form

$$\begin{aligned} l(\mathbf{A}|\mathbf{y}) = p(\mathbf{y}|\mathbf{A}) &= (2\pi)^{-np/2}\tau^{p/2}\left|n\exp\{-\mathbf{A}\} + \tau\mathbf{F}^T\exp(-\mathbf{A})\mathbf{F}\right|^{-1/2} \\ &\quad \times \exp\left\{-\tfrac{1}{2}(n+1)\mathrm{tr}\,\mathbf{A} - \tfrac{1}{2}\mathrm{tr}(\mathbf{U}\exp\{-\mathbf{A}\})\right\} \\ &\quad \times \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left[(\bar{\mathbf{y}} - \boldsymbol{\mu}_\theta)(\bar{\mathbf{y}} - \boldsymbol{\mu}_\theta)^T\right.\right. \\ &\quad \left.\left. \times (n^{-1}\exp(\mathbf{A}) + \tau^{-1}\mathbf{F}^T\exp(\mathbf{A})\mathbf{F})^{-1}\right]\right\} \end{aligned} \tag{6.7}$$

Using the general computational techniques described in Section 7, it is possible to calculate the posterior distribution and moments of any function

$\eta = g(\mathbf{C})$ of $\mathbf{C}$ which is of interest. However, in the special case when $\mathbf{F} = \mathbf{I}_p$, (6.7) conveniently reduces to

$$(6.8) \quad l(\mathbf{A}|\mathbf{y}) = (2\pi)^{-np/2}\rho^{p/2}\exp\{-\tfrac{1}{2}n\ \mathrm{tr}(\mathbf{A}) - \tfrac{1}{2}\mathrm{tr}(\mathbf{U}^*\exp\{-\mathbf{A}\})\},$$

where $\mathbf{U}^*$ satisfies (6.3) and (6.5). The integrated likelihood (6.8) assumes a form proportional to (2.1), but with $n\mathbf{S}$ replaced by $\mathbf{U}^*$. Hence the likelihood approximations (2.12) and (2.16) are also applicable in this situation, where $\boldsymbol{\theta}$ is unknown. Therefore, the analyses of Sections 3 and 4, which incorporate prior information about $\mathbf{A} = \log \mathbf{C}$ may be repeated as before. In particular, under the exchangeable distribution for $\mathbf{C}$ introduced in Section 3.2, the distribution for the elements of $\boldsymbol{\theta}$ will also be exchangeable in the prior assessment. The Bayes estimates for $\mathbf{A}$ will smooth the matrix logarithm of the expanded term (6.3). Under a uniform prior for $\boldsymbol{\theta}$ on R^P, which is independent of the prior for $\mathbf{A}$, the analysis of Sections 3 and 4 may again be employed, but with $\mathbf{S}$ replaced by $\mathbf{U}/(n-1)$ and n reduced to $n-1$.

7. Bayesian computational methods. Suppose that, in general, a $q \times 1$ vector $\boldsymbol{\alpha}$ has posterior density $\pi(\boldsymbol{\alpha}|\mathbf{y})$ with positive support on R^q. Then the method of importance sampling [e.g., Geweke (1989), Leonard, Hsu and Tsui (1989) and Hsu, Leonard and Tsui (1991)] computes the exact posterior expectation h_E of any function $h(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha}$ (whenever this exists) by simulating realizations $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \ldots$, from any other density $\pi^*(\boldsymbol{\alpha}|\mathbf{y})$ with positive support on R^P, and calculating

$$(7.1) \qquad h_M = \sum_{k=1}^{M} h(\boldsymbol{\alpha}_k)W(\boldsymbol{\alpha}_k) \Big/ \sum_{k=1}^{M} W(\boldsymbol{\alpha}_k) \qquad M = 1, 2, 3, \ldots,$$

with

$$(7.2) \qquad W(\boldsymbol{\alpha}) = \pi(\boldsymbol{\alpha}|\mathbf{y})/\pi^*(\boldsymbol{\alpha}|\mathbf{y}) \qquad \boldsymbol{\alpha} \in R^q.$$

Note that the weight function (7.2) need only be specified up to a multiplicative constant, not depending upon $\boldsymbol{\alpha}$. The first theorem described by Geweke proves, via the strong law of large numbers, that h_M, satisfying (7.1), strongly converges as $M \to \infty$, to h_E, whenever h_E exists. His second theorem proves that $M^{1/2}[h_M - h_E]$ converges in distribution as $M \to \infty$, to a zero mean normal variate, whenever the conditions

$$(7.3) \qquad \int_{R^P} W^2(\boldsymbol{\alpha})\pi^*(\boldsymbol{\alpha}|\mathbf{y}) < \infty$$

and

$$(7.4) \qquad \int_{R^P} h^2(\boldsymbol{\alpha})W^2(\boldsymbol{\alpha})\pi^*(\boldsymbol{\alpha}|\mathbf{y}) < \infty$$

are satisfied. If conditions (7.3) and (7.4) are not satisfied, then Geweke remarks that h_M will still strongly converge to h_E, but that the practical convergence can (but need not) be both extremely slow and difficult to check,

as the value of h_M can occasionally change substantially as M increases. Computations by Hsu, Leonard and Tsui (1991) in fact suggest that closeness of h_M to h_E for finite M can still be good in practical terms, unless a very unlucky realization of $\boldsymbol{\alpha}$ occurs, if a reasonable preliminary transformation of the parameters (in our case the logarithmic matrix transformation) has been chosen to ensure that a convenient choice of the importance function $\pi^*(\boldsymbol{\alpha}|\mathbf{y})$ is available which is reasonably close to $\pi(\boldsymbol{\alpha}|\mathbf{y})$. If there are differences in the tails of these densities, it is of course better for the tails of $\pi^*(\boldsymbol{\alpha}|\mathbf{y})$ to be thicker than the tails of $\pi(\boldsymbol{\alpha}|\mathbf{y})$, so that the weight function (7.2) is less likely to occasionally assume extremely high values.

First consider the problem of computing the posterior distribution function (c.d.f.) $\Lambda_y(\eta)$ of any parameter $\eta = g(\mathbf{C}) = g^*(\boldsymbol{\alpha})$ of interest, when $\boldsymbol{\alpha} = \text{vec}(\mathbf{A})$ has a prior distribution which is uniform over R^q, and $\pi(\boldsymbol{\alpha}|\mathbf{y})$ is proportional to the likelihood function in (2.1). In this case, set $h(\boldsymbol{\alpha}) = I[g^*(\boldsymbol{\alpha}) \leq \eta]$, where $I(\mathbf{A})$ is the indicator function for the set $\mathbf{A}$. The importance function,

$$\pi^*(\boldsymbol{\alpha}|\mathbf{y}) = t_{\boldsymbol{\alpha}}[\omega, \boldsymbol{\lambda}, \omega\mathbf{Q}/(\omega + q)] \qquad \boldsymbol{\alpha} \in R^q \tag{7.5}$$

is recommended, that is a multivariate t density with ω degrees of freedom, mean vector $\boldsymbol{\lambda}$ and precision matrix $\omega\mathbf{Q}/(\omega + q)$, where $\boldsymbol{\lambda}$ and $\mathbf{Q}$, respectively, denote the maximum likelihood vector of $\boldsymbol{\alpha}$ and the likelihood information matrix (2.13). Reducing the value of ω can, in practical terms, reduce the fluctuations of the series (7.1). However, too small a value of ω can slow down the convergence. A sensible pragmatic choice of ω, which ensures, if possible, fast practical convergence of the series (7.1), is therefore recommended.

The importance function (7.5) may be developed from the exact posterior density $\pi(\boldsymbol{\alpha}|\mathbf{y})$, by exponentiating $\pi(\boldsymbol{\alpha}|\mathbf{y})$ to the power $-2/(\omega + q)$, expanding in a Taylor Series about $\boldsymbol{\alpha} = \boldsymbol{\lambda}$, neglecting cubic and higher terms and exponentiating to the power $-(\omega + q)/2$. This is a special case of a more general method described by Leonard, Hsu and Ritter (1993). For the numerical examples of Section 8, reasonably fast convergence is available with $\omega = 30$, and very similar results were obtained with $\omega = \infty$, though this choice is not in general recommended. Among many other applications of this procedure, it permits the computation of the exact posterior mean vector $\tilde{\boldsymbol{\lambda}}$ and covariance matrix $\tilde{\mathbf{Q}}^{-1}$, under a uniform prior for $\boldsymbol{\alpha}$, leading to the finite sample refinement (2.16) to the likelihood function (2.12). The computations in Section 8 are based upon $M = 40{,}000$ simulations, providing practical convergence for the data in Section 8, within a degree of tolerance of about 0.005 for all posterior c.d.f.'s, and convergence to about three significant figures for the posterior means and variances (if they exist) of many continuous functions $h(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha}$. The simulations were replicated several times with very similar results; the results reported assume $\omega = \infty$.

Next consider the prior informative analysis discussed in Section 4, where $\boldsymbol{\alpha}$ possesses a multivariate normal prior distribution with specified mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Delta}$, so that $\pi(\boldsymbol{\alpha}|\mathbf{y})$ multiplies the likelihood contribution by a factor proportional to $\exp\{-(\boldsymbol{\alpha} - \boldsymbol{\xi})^T\boldsymbol{\Delta}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\xi})/2\}$. In this case, the

importance function

$$\pi^*(\boldsymbol{\alpha}|\mathbf{y}) = t_{\boldsymbol{\alpha}}\big(\omega, \tilde{\boldsymbol{\xi}}, \omega\tilde{\boldsymbol{\Delta}}^{-1}/(\omega + q)\big) \qquad \boldsymbol{\alpha} \in R^q \tag{7.6}$$

is recommended, where $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\Delta}}$ replace $\boldsymbol{\lambda}$ and $\mathbf{Q}$ in the expressions for $\boldsymbol{\xi}^*$ and $\boldsymbol{\Delta}^*$ in (3.1) and (3.2) by the previously computed quantities $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\mathbf{Q}}$. The density in (7.6) may be developed from the current $\pi(\boldsymbol{\alpha}|\mathbf{y})$, by noting that, under the likelihood approximation (2.15) $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\Delta}}^{-1}$ are the posterior mode vector and information matrix of $\boldsymbol{\alpha}$, and obtaining a second order Taylor series approximation of the type considered in the previous paragraph of the present section. Again, $\omega = 30$ or $\omega = \infty$ can suffice when computing the exact posterior distribution of any parameter of interest. Very similar procedures may be employed when $\boldsymbol{\xi}^*$ and $\boldsymbol{\Delta}^*$ are replaced by the vector and matrix in (3.7) and (3.8) and for the exact version of the empirical Bayes procedures of Section 4, once the vector $\boldsymbol{\zeta}$ of prior parameters has been estimated.

Next, consider the problem of computing the exact version of the integrated likelihood for $\boldsymbol{\zeta}$ in (4.2). The exact integrated likelihood may be represented in the form

$$l(\boldsymbol{\zeta}|\mathbf{y}) \propto E\pi(\boldsymbol{\alpha}|\boldsymbol{\zeta}) \qquad \boldsymbol{\zeta} \in R^{l_2}, \tag{7.7}$$

where the expectation is with respect to the posterior distribution for $\boldsymbol{\alpha}$, under a uniform prior for $\boldsymbol{\alpha}$, and

$$\pi(\boldsymbol{\alpha}|\boldsymbol{\zeta}) \propto |\boldsymbol{\Delta}|^{-1/2}|\mathbf{X}^T\boldsymbol{\Delta}^{-1}\mathbf{X}|^{-1/2} \exp\Big\{-\tfrac{1}{2}\boldsymbol{\alpha}^T\overline{\mathbf{R}}\boldsymbol{\alpha}\Big\} \qquad \boldsymbol{\alpha} \in R^q, \tag{7.8}$$

with $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\zeta})$ and

$$\overline{\mathbf{R}} = \tilde{\mathbf{R}}\boldsymbol{\Delta}^{-1}\tilde{\mathbf{R}}, \tag{7.9}$$

with $\tilde{\mathbf{R}} = \mathbf{I}_q - \mathbf{X}(\mathbf{X}^T\boldsymbol{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$. In this case, the importance function in (7.5) is again recommended. The general importance sampling procedure may now be applied for any fixed $\boldsymbol{\zeta}$, but with $h(\boldsymbol{\alpha}) \propto \pi(\boldsymbol{\alpha}|\boldsymbol{\zeta})$, and $W(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}|\mathbf{y})/\pi^*(\boldsymbol{\alpha}|\mathbf{y})$.

Hierarchical Bayes solutions, which integrate the posterior density of $\boldsymbol{\alpha}$ given $\boldsymbol{\zeta}$, with respect to a posterior density of $\boldsymbol{\zeta}$, are typically overtedious to compute exactly. However, if the likelihood approximation (2.16) can be shown to be accurate, then all the procedures described in this paper based on (2.16) should also be reasonably accurate. Importance sampling techniques can be used to compare different components of these approximations with the exact results. The integrations with respect to $\boldsymbol{\zeta}$ can then be performed, under approximation (2.16), either using numerical integrations or further importance sampling techniques. The procedures recommended here may be contrasted with the alternative methodologies recommended by Kass and Steffey (1989), which appear to be difficult to apply in the present context.

8. Project Talent American High School data. Consider a subset of the Project Talent data reported by Flanagan, Davis, Dailey, Shaycott, Orr, Goldberg and Neyman (1964) and previously analyzed by Cooley and Loynes

(1971). The data concern $n = 78$ eighteen year old female twelfth grade students and their results on the following $p = 8$ tests: (a) information, part 1 (b) information, part 2 (c) English (d) reading comprehension (e) creativity (f) mechanical reading (g) abstract reading and (h) mathematics. The scores were calculated as proportions of the maximum scores possible. The 8×1 sample mean vector of scores was $\bar{\mathbf{y}} = (0.514, 0.496, 0.784, 0.677, 0.555, 0.446, 0.602, 0.410)^T$, with respective sample standard deviations (0.118, 0.136, 0.085, 0.181, 0.183, 0.168, 0.190, 0.165).

The eigenvalues of the sample covariance matrix $\mathbf{S} = \mathbf{U}/(n-1)$ are 0.0024, 0.0028, 0.0074, 0.0117, 0.0141, 0.179, 0.0225 and 0.1189, and the sample correlation matrix is reported as matrix A1 of Appendix 4, where matrix A2 is $\mathbf{\Lambda} = \log \mathbf{S}$. Note that $\mathbf{\Lambda}$ maximizes the integrated likelihood of $\mathbf{A} = \log \mathbf{C}$ when the mean vector $\boldsymbol{\theta}$ is also unknown, but a priori uniformly distributed over R^P. Theorem 2 of Appendix 1 tells us that the square roots of the diagonal terms of the inverse of the matrix $\mathbf{Q}$ in (2.13) provide approximate estimated standard errors for the elements of $\mathbf{\Lambda}$. These are summarized by matrix A3.

The calculations for $\mathbf{\Lambda}$ and $\mathbf{Q}$ lead to an approximation of the form (2.12) for the integrated likelihood of $\boldsymbol{\alpha}$ under a uniform prior for $\boldsymbol{\theta}$. Using the techniques of Section 7, the exact posterior mean matrix $\mathbf{\Lambda}$ of $\mathbf{A}$ (see matrix A4), and the exact posterior covariance matrix $\tilde{\mathbf{Q}}^{-1}$ of $\boldsymbol{\alpha} = \text{vec}(\mathbf{A})$ were calculated under a uniform prior for $\boldsymbol{\alpha}$. Matrix A5 reports the exact posterior standard errors for the corresponding elements of $\mathbf{A}$. The remarkable closeness between matrices A2 and A4 and between A3 and A5 suggests reasonable adequacy of the likelihood approximation (2.12). However, there are slight differences between the diagonal elements of A2 and the corresponding diagonal elements of A4. The likelihood approximation (2.16) with $\tilde{\boldsymbol{\lambda}} = \text{vec}\, \tilde{\mathbf{\Lambda}}$ and $\tilde{\mathbf{Q}}$ defined above, is therefore preferable.

Histogram (a) of Figure 1 represents the exact posterior density of the $(1,1)$th element a_{11} of $\mathbf{A}$, curve (b) represents the approximate normal curve based on (2.16) and curve (c) represents a similar curve based on (2.12). The result for likelihood approximation (2.16) is extremely accurate. Similar results were obtained for all diagonal elements of $\mathbf{A}$. For the off-diagonal elements, both approximations (2.12) and (2.16) were similarly accurate, when compared with the exact result to curve (b) of Figure 1. Curve (a) of Figure 2 describes the exact posterior c.d.f. of the $(1,2)$th correlation ρ_{12}, and curve (b) was computed via approximation (2.16). Results of similar accuracy were obtained for many parameters of interest, including the generalized variance $|\mathbf{C}|$. This provides an empirical finite sample size validation for the likelihood approximation (2.16).

Next consider the exchangeable prior formulation of Section 3.2, but where the prior parameters μ_1, μ_2, σ_1^2 and σ_2^2 are a priori independent and uniformly distributed over their ranges of possible values. The integrated likelihood (5.2) was maximized with $\boldsymbol{\zeta} = (\kappa_1, \kappa_2)$, $\kappa_1 = \log \sigma_1^2$ and $\kappa_2 = \log \sigma_2^2$, yielding the maximizing values $\hat{\kappa}_1 = -0.265$ and $\hat{\kappa}_2 = -3.804$. Less smoothing of the diagonal terms of $\mathbf{\Lambda}$ is therefore recommended when compared with

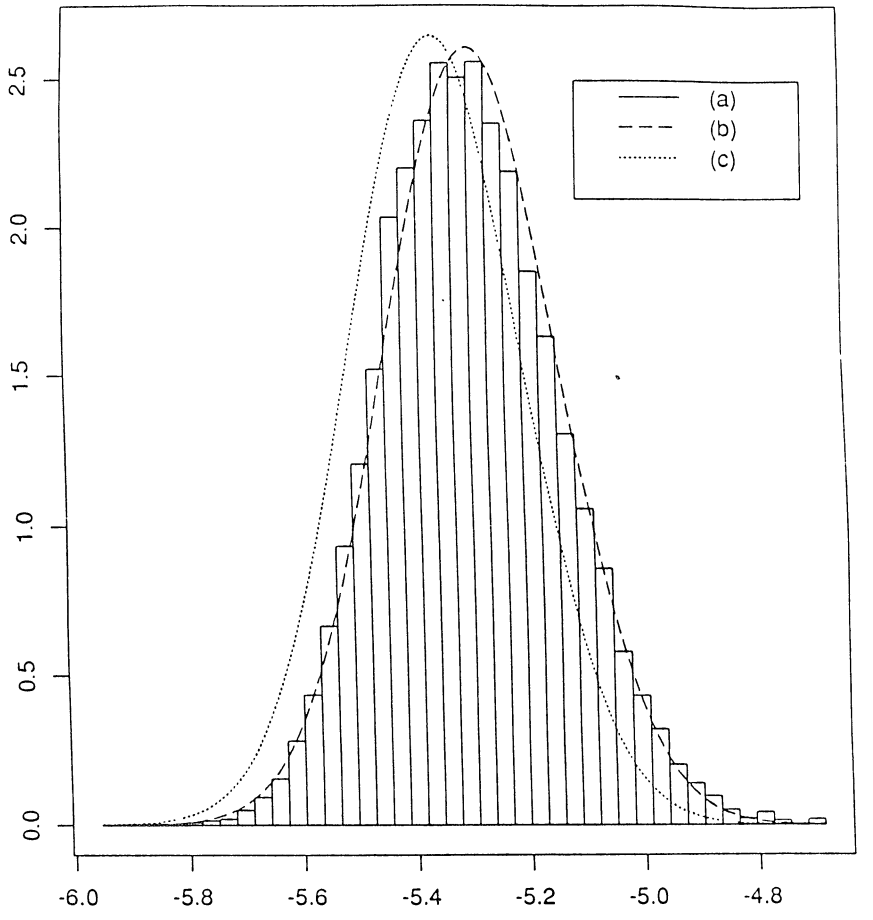

FIG. 1. *Posterior density of* a_{11} (a) *histogram, exact,* (b) *using likelihood approximation* (2.16), *and* (c) *using likelihood approximation* (2.12).

the diagonal terms. The above maximization was completed by inspection of a bivariate contour plot of the integrated likelihood of κ_1 and κ_2 (not reported here) which also permits more detailed inference regarding κ_1 and κ_2.

Curves (c) and (d) of Figure 2 provide our exact and approximate [i.e., based upon (2.16)] posterior c.d.f.'s for the correlation ρ_{12} under the above informative prior when $\kappa_1 = -0.265$ and $\kappa_2 = -3.804$, and these are again reassuring close. Curve (c) provides the corresponding exact "empirical Bayes" posterior density of ρ_{12}, and it is possible to calculate a similar density for any parameter of interest. Under these choices of κ_1 and κ_2, the exact posterior mean matrix of $\mathbb{A}$ and the matrix of corresponding exact posterior deviations are recorded as matrices A6 and A7. Comparison of A6 with A4 and A5 with A3, demonstrates the effect of the exchangeable prior assumptions, when compared with the vague prior situation. The diagonal elements of A6 tend to be

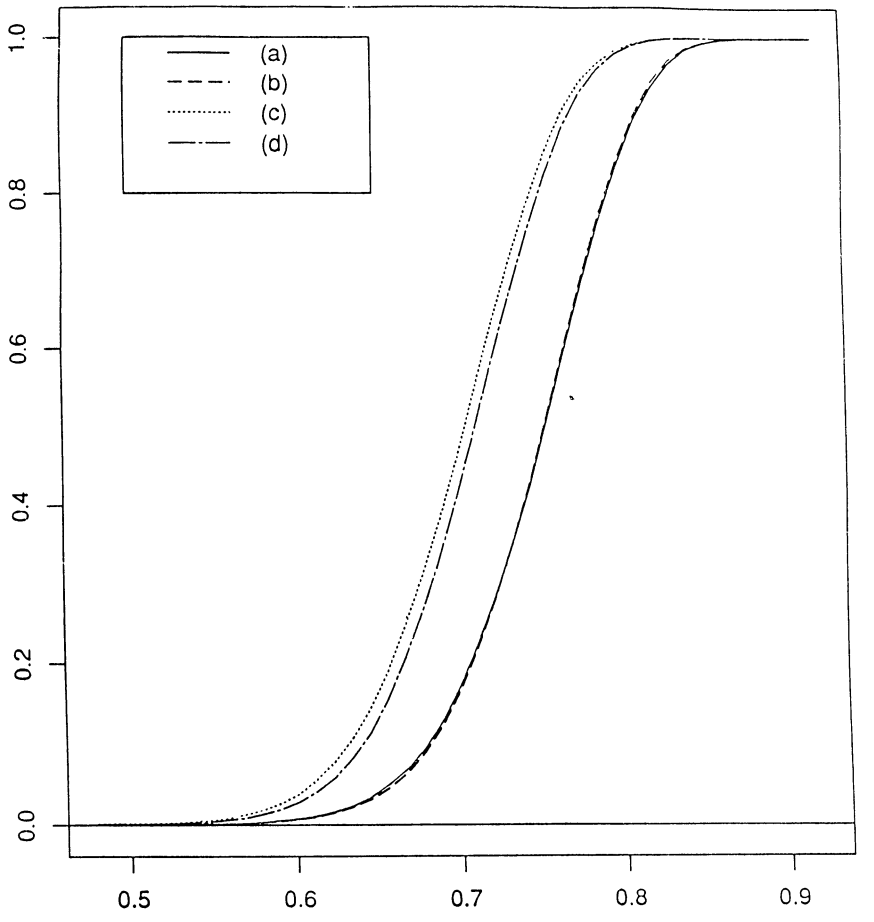

FIG. 2. *Posterior* c.d.f. *of correlation* ρ_{12} (a) *exact* c.d.f. *under uniform prior*, (b) *using uniform prior and likelihood approximation* (2.16), (c) *exact* c.d.f. *using empirical exchangeable prior, and* (d) *using exchangeable prior and likelihood approximation* (2.16).

closer together than the diagonal elements of A4 and similarly the off-diagonal elements. Most of the standard deviations in A7 are smaller than the corresponding elements of A6.

The posterior means of the population standard deviation were, respectively, 0.118, 0.137, 0.098, 0.180, 0.186, 0.171, 0.195 and 0.167, very close to the sample standard deviation previously described. However, matrix A8, which describes the exact posterior means of the population correlations, substantially smooths the sample correlation matrix A1. Matrix A9 of Appendix 4 denotes the posterior mean matrix for $\mathbf{A}$ based upon the full hierarchical prior, described above, likelihood approximation (2.16) and numerical integrations involving the corresponding posterior density for σ_1^2 and σ_2^2.

9. Applications and extensions in multivariate analysis. The posterior smoothing of $\mathbf{C}$ and finite sample inference techniques yield numerous potential applications and possible extensions in multivariate analysis, which are currently being considered by the co-authors, for example, multiple linear regression: (a) The prediction of one variable, via its conditional distribution, given the other $p - 1$ variables [see Anderson (1984), page 28]. (b) Smoothing a quadratic discrimination function, with all interactions present (Anderson, page 142). (c) Smoothing the eigenvalues in principal components analysis (Anderson, page 272). (d) The development of a generalized linear model of the form

$$\log \mathbf{C}_i = \alpha_1 \mathbf{U}_{i1} + \cdots + \alpha_r \mathbf{U}_{ir} \qquad i = 1, \ldots, n, \tag{9.1}$$

for several unequal covariance matrices $\mathbf{C}_1, \ldots, \mathbf{C}_n$, which may be incorporated with a linear model for the corresponding mean vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, of observation vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$. The co-authors, Tom Chiu and Kam-Wah Tsui, are currently developing this important model in detail.

APPENDIX 1

Asymptotic results. Owing to the uniqueness of the Taylor series expansion leading to (2.12) and (2.13), the matrix $\mathbf{Q}$ satisfying (2.13) is also the exact likelihood information matrix

$$\mathbf{Q} = -\partial^2 \log l(\boldsymbol{\alpha}|\mathbf{y}) / \partial(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\big|_{\boldsymbol{\alpha}=\boldsymbol{\lambda}} \tag{1}$$

of $\boldsymbol{\alpha}$. As the likelihood of $\boldsymbol{\alpha}$ is twice differentiable with a unique maximum at $\boldsymbol{\lambda} = \log \mathbf{S}$, the matrix $\mathbf{Q}$ is positive definite. The following theorem is a direct consequence of the Taylor series expansion, and a general result for asymptotic posterior normality described by Johnson (1970). Note that both $\mathbf{E}_0^T \mathbf{A} \mathbf{E}_0$ and $\mathbf{E}_0^T \boldsymbol{\Lambda} \mathbf{E}_0$ in the expression for $\mathbf{B}$ in (2.8) estimate the matrix logarithm of the covariance matrix of the rotated vectors $\mathbf{E}^T \mathbf{y}_i$.

THEOREM 1. *Suppose that $\boldsymbol{\alpha}$ possesses a prior density with positive support on R^q. Then:*

1. *As $n \to \infty$, $\mathbf{Q}^{1/2}(\boldsymbol{\alpha} - \boldsymbol{\lambda})$ converges in posterior distribution with sampling probability one, to a q-dimensional standardized spherical normal random vector.*
2. *Rotations to asymptotic posterior independence. As $n \to \infty$, the q upper triangular elements of the matrix $n^{1/2}\mathbf{B}$, where $\mathbf{B}$ satisfies (2.8), converge in posterior distribution with sampling probability 1, to q independent zero mean normal variables. For $i = 1, \ldots, p$, the ith diagonal element b_{ii} has limiting variance equal to 2. For $i \neq j$, b_{ij} has limiting variance equal to the limit of ξ_{ij}^{-1} satisfying (2.11), which exists with sampling probability 1.*

Note that Fisher's expected information matrix for $\boldsymbol{\alpha}$ is

$$\tilde{\mathbf{Q}} = \tfrac{1}{2}n\sum_i \tilde{\mathbf{f}}_{ii}\tilde{\mathbf{f}}_{ii}^T + n\sum_{i,j:i<j} \tilde{\xi}_{ij}\tilde{\mathbf{f}}_{ij}\tilde{\mathbf{f}}_{ij}^T, \tag{2}$$

where the $\tilde{\xi}_{ij}$ and $\tilde{\mathbf{f}}_{ij}$ replace the d_{ii} and $\mathbf{e}_i$ in the expression for ξ_{ij} and $\mathbf{f}_{ij}$ in (2.11) and (2.14) by the corresponding eigenvalues and normalized eigenvectors of the true covariance matrix $\mathbf{C}$. This result may be obtained by using (2.5) to expand the log of the likelihood (2.1) in a Taylor series about an arbitrary symmetric matrix $\mathbf{A} = \mathbf{A}_0$. This provides complex expressions for all first and second derivatives of the log-likelihood. Then $\mathbf{S}$ should be replaced by its expectation $\mathbf{C}$ leading to simplifications along the lines recommended in (2.7)–(2.14). [The first derivatives obtained by this procedure, can be used to obtain the Jacobian of the transformation $\boldsymbol{\alpha} = \text{vec}(\log \mathbf{C})$. By transforming back to $\mathbf{C}$, and referring to the expectation of the Jacobian term with respect to an inverted Wishart distribution, it is therefore straightforward to show that the exact likelihood (2.1) integrates over $\mathscr{D}_p$ to a finite quantity, as required in Sections 2 and 3.2.]

Note that, as $n \to \infty$ the elements of $n^{-1}\mathbf{Q}$ are strongly consistent for the corresponding elements of $n^{-1}\tilde{\mathbf{Q}}$. Therefore part (b) of the following theorem is a consequence of Slutsky's theorem, and part (a), which just comprises standard maximum likelihood asymptotics (e.g., based on asymptotic normality of the first derivative with respect to $\boldsymbol{\alpha}$ of the log-likelihood of $\boldsymbol{\alpha}$).

THEOREM 2. (a) *As* $n \to \infty$, $\tilde{\mathbf{Q}}^{1/2}(\boldsymbol{\lambda} - \boldsymbol{\alpha})$ *converges in sampling distribution to a q-dimensional standardized spherical normal random vector.*

(b) *As* $n \to \infty$, $\mathbf{Q}^{1/2}(\boldsymbol{\lambda} - \boldsymbol{\alpha})$ *converges in the same manner.*

COROLLARY 1. *As* $n \to \infty$,

$$n^{1/2}(\log \mathbf{S} - \log \mathbf{C}) \to_d \mathbf{EWE}^T, \tag{3}$$

where $\mathbf{E}$ *satisfied* (1.3), *and* $\mathbf{W}$ *is a symmetric matrix whose upper triangular elements are independent zero mean normal variates. The diagonal elements of* $\mathbf{W}$ *have variance equal to two, and the off-diagonal element* w_{ij} *has variance equal to* $\tilde{\xi}_{ij}$, *where* $\tilde{\xi}_{ij}$ *is defined above.*

Corollary 1 is just a restatement of part (a) of Theorem 2, but with the elements of the vector $\boldsymbol{\lambda} - \boldsymbol{\alpha}$ rearranged into matrix form. The theorem provides a multivariate normal approximation for the vectorization $\boldsymbol{\alpha} = \text{vec}(\log \mathbf{C})$ of the matrix logarithm of a random Wishart matrix $\mathbf{S}$. It may also be used to obtain a similar approximation to the posterior density of $\log \mathbf{C}$ when $\mathbf{C}^{-1}$ possesses a conjugate Wishart prior density. This would facilitate the comparison of several covariance matrices [see Press (1992)].

APPENDIX 2

Risk properties. Consider first the special case where $\boldsymbol{\theta} = \mathbf{0}$ and $\mathbf{C} = \text{diag}(\phi_1, \ldots, \phi_p)$. Then, following Bartlett and Kendall (1946), the distribution

of l_j, the log of the jth diagonal element of $\mathbf{S}$ is for $n \geq 10$, closely approximated by a normal distribution with mean $\alpha_j = \log \phi_j$ and variance $2n^{-1}$. Consider the empirical Bayes shrinkage estimators

$$\alpha_j^* = \tilde{\omega} l_j + (1 - \tilde{\omega})\bar{l} \qquad j = 1, \ldots, p, \tag{1}$$

where

$$1 - \tilde{\omega} = \min\left[2n^{-1}p \Big/ \sum \left(l_j - \bar{l}\right)^2, 1\right]. \tag{2}$$

Efron and Morris (1973) develop excellent mean squared error (mse) properties for the α_j^* (assuming that the above normal approximation for the distribution of l_j is exact), which are valid for any $p \geq 3$. For large p, the mse of the α_j^* multiplies the approximate mse $= 2pn^{-1}$ of the maximum likelihood estimators $\hat{\alpha}_j = l_j$ by a factor equal to $\Sigma(\alpha_j - \bar{\alpha})^2/[\Sigma(\alpha_j - \bar{\alpha})^2 + 2pn^{-1}]$ [see also Leonard (1976)]. Efron and Morris (1976) and Haff (1980) obtain similarly convincing properties for shrinkage estimators of the form (1.2), under a variety of choices of loss function for a general covariance matrix $\mathbf{C}$.

The estimator $\mathbf{A}^*$ for $\mathbf{A}$ corresponding to (3.7) and $\boldsymbol{\theta} = \mathbf{0}$, takes the form of a matrix weighted average of $\boldsymbol{\Lambda} = \log(\tilde{\mathbf{U}}/n)$ with $\tilde{\mathbf{U}} = \Sigma_i \mathbf{y}_i \mathbf{y}_i^T$, and a matrix of intraclass form, whose elements are themselves estimated from the data. In Tables 1, 2 and 3 we report the total mse's for the diagonal terms of $\mathbf{A}^*$ and the nondiagonal terms of $\mathbf{A}^*$, for various choices of the parameters $\kappa_1 = \log \sigma_1^2$ and $\kappa_2 = \log \sigma_2^2$, three different choices $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$ of the true covariance matrix $\mathbf{C}$, $p = 6$ and $n = 50$. The numbers in parentheses give the corresponding total mse's for the diagonal terms of $\mathbf{C}^* = \exp(\mathbf{A}^*)$ and the nondiagonal terms. These results permit consideration of the risk properties of either $\mathbf{A}^*$ or $\mathbf{C}^*$ under any loss function, which is a linear combination of the total of the diagonal squared errors and the total of the nondiagonal squared errors for estimators of $\mathbf{A}$ or $\mathbf{C}$. All results are based upon 10,000 simulations with the error of simulation measured by a coefficient of variation of less than 1%, and refer to the exact distribution of $\tilde{\mathbf{U}}$.

The results in Table 1 correspond to the choice for the true covariance matrix $\mathbf{C}$ of an intraclass matrix $\mathbf{C}_1$, with all variances equal to two and unit covariances. The maximum likelihood estimator $\boldsymbol{\Lambda} = \log(\tilde{\mathbf{U}}/50)$ of $\mathbf{A}$ possessed diagonal component of total risk equal to 0.273, and nondiagonal component of total risk equal to 0.307. However, the estimator $\tilde{\mathbf{A}} = \log(\tilde{\mathbf{U}}/47)$ adjusted these components to 0.242 and 0.307 respectively, and $\tilde{\mathbf{U}}/47$ was the best integer divisor of $\tilde{\mathbf{U}}$ in terms of both components of the mse for estimators of $\mathbf{A}$. However, the approximate Bayes estimators (3.7), substantially improve both components of risk for a large range of values of κ_1 and κ_2 when compared with any integer divisor of $\tilde{\mathbf{U}}$. We anticipate that the exact posterior mean matrix of $\mathbf{A}$ would possess even better risk properties, but these are overtedious to compute. The maximum likelihood estimator $\mathbf{S} = \tilde{\mathbf{U}}/50$ possessed diagonal and nondiagonal components of total risk equal to 0.960 and 1.501 respectively, and the best integer divisor of $\tilde{\mathbf{U}}$, when considering components of mse for estimators of $\mathbf{C}$, was $\tilde{\mathbf{U}}/53$; the latter reduced the compo-

TABLE 1
Components of risk (covariance matrix $\mathbf{C}_1$*)*[1]

κ_2	κ_1 = −2		κ_1 = −1		κ_1 = 0		κ_1 = 1		κ_1 = 2	
−2	0.197	0.240	0.238	0.241	0.258	0.241	0.266	0.242	0.269	0.242
	(0.764	1.339)	(0.856	1.353)	(0.903	1.360)	(0.923	1.363)	(0.930	1.364)
−1	0.198	0.278	0.238	0.280	0.259	0.280	0.267	0.280	0.270	0.280
	(0.778	1.420)	(0.871	1.433)	(0.918	1.440)	(0.937	1.443)	(0.945	1.444)
0	0.198	0.295	0.239	0.296	0.259	0.296	0.267	0.296	0.270	0.296
	(0.785	1.455)	(0.878	1.468)	(0.924	1.475)	(0.944	1.478)	(0.951	1.479)
1	0.198	0.301	0.239	0.302	0.259	0.303	0.267	0.303	0.271	0.303
	(0.788	1.469)	(0.880	1.482)	(0.927	1.488)	(0.946	1.491)	(0.954	1.492)
2	0.198	0.304	0.239	0.305	0.259	0.305	0.267	0.305	0.271	0.305
	(0.789	1.474)	(0.881	1.487)	(0.928	1.494)	(0.947	1.496)	(0.955	1.497)

[1]For each value of κ_1 and κ_2, the figure in the first column denotes the diagonal component of risk for estimators of $\mathbf{A}$, and the figure in the second column denotes the nondiagonal component of risk. The figures in parentheses describe similar quantities for estimators of $\mathbf{C}$.

nents of $\mathbf{C}$ to 0.930 and 1.384. The entries in parentheses in Table 1 illustrate that our Bayesian estimator $\mathbf{C}^* = \exp(\mathbf{A}^*)$ performs much better for many values of κ_1 and κ_2.

The results in Table 2 involve a choice $\mathbf{C}_2$ of the true covariance matrix $\mathbf{C}$ which is of autoregressive form, with unit variances, and (j, k)th correlation equal to $0.7^{|j-k|}$. However, the approximate Bayes estimator (3.7) still shrinks $\mathbf{\Lambda}$ towards intraclass form, that is, toward an incorrect hypothesis. The diagonal and nondiagonal components of total risk for $\mathbf{\Lambda}$ now equal 0.259 and 0.290, while $\tilde{\mathbf{\Lambda}} = \log(\tilde{\mathbf{U}}/47)$ adjusts these to 0.228 and 0.290. Note that

TABLE 2
Components of risk (covariance matrix $\mathbf{C}_2$*)*[1]

κ_2	κ_1 = −2		κ_1 = −1		κ_1 = 0		κ_1 = 1		κ_1 = 2	
−2	0.196	0.242	0.227	0.242	0.244	0.242	0.250	0.242	0.253	0.242
	(0.193	0.333)	(0.209	0.336)	(0.217	0.338)	(0.221	0.339)	(0.222	0.339)
−1	0.197	0.264	0.229	0.265	0.246	0.265	0.253	0.265	0.256	0.265
	(0.199	0.349)	(0.216	0.353)	(0.225	0.355)	(0.228	0.355)	(0.230	0.356)
0	0.198	0.278	0.230	0.279	0.247	0.280	0.254	0.279	0.257	0.281
	(0.204	0.361)	(0.221	0.361)	(0.230	0.366)	(0.234	0.367)	(0.235	0.367)
1	0.198	0.284	0.230	0.285	0.247	0.286	0.254	0.286	0.257	0.286
	(0.206	0.361)	(0.223	0.369)	(0.232	0.371)	(0.236	0.372)	(0.238	0.373)
2	0.198	0.286	0.230	0.287	0.247	0.288	0.255	0.288	0.257	0.288
	(0.207	0.368)	(0.224	0.372)	(0.233	0.374)	(0.237	0.374)	(0.239	0.375)

[1]For each value of κ_1 and κ_2, the figure in the first column denotes the diagonal component of risk for estimators of $\mathbf{A}$, and the figure in the second column denotes the nondiagonal component of risk. The figures in parentheses describe similar quantities for estimators of $\mathbf{C}$.

TABLE 3
Components of risk (covariance matrix $\mathbf{C}_3$*)*[1]

κ_2	κ_1 = −2		−1		0		1		2	
−2	0.662	0.181	0.304	0.186	0.261	0.190	0.265	0.192	0.269	0.192
	(211.0	144.8)	(108.5	93.7)	(87.2	83.0)	(84.6	81.6)	(84.5	81.5)
−1	0.669	0.205	0.307	0.210	0.263	0.214	0.266	0.216	0.270	0.217
	(192.9	129.3)	(101.3	88.6)	(86.5	83.9)	(86.7	85.0)	(87.8	86.0)
0	0.671	0.215	0.308	0.221	0.264	0.225	0.267	0.227	0.271	0.228
	(186.1	123.9)	(99.0	87.6)	(86.9	85.3)	(88.3	87.6)	(89.9	88.9)
1	0.672	0.219	0.309	0.225	0.264	0.229	0.267	0.232	0.271	0.232
	(183.5	122.0)	(198.2	87.3)	(87.1	86.1)	(89.0	88.7)	(90.8	90.3)
2	0.672	0.221	0.309	0.227	0.264	0.231	0.267	0.233	0.271	0.234
	(182.6	121.3)	(98.0	87.3)	(87.2	86.4)	(89.3	88.2)	(91.1	90.8)

[1]For each value of κ_1 and κ_2, the figure in the first column denotes the diagonal component of risk for estimators of $\mathbf{A}$, and the figure in the second column denotes the nondiagonal component of risk. The figures in parentheses describe similar quantities for estimators of $\mathbf{C}$.

$\kappa_1 = -1$ and $\kappa_2 = -2$ or -1 still lead to choices of $\mathbf{A}^*$ with superior risk properties to both Λ and $\tilde{\Lambda}$. The corresponding components of risk for $\mathbf{S} = \tilde{\mathbf{U}}/50$ were 0.240 and 0.376, and the estimator $\tilde{\mathbf{U}}/53$ reduced these components to 0.233 and 0.347. However, $\mathbf{C}^*$ is superior when $\kappa_1 = -2$ or -1 and $\kappa_2 = -1$.

The results in Table 3 are based upon a choice $\mathbf{C}_3$ of true covariance matrix $\mathbf{C}$ which possesses diagonal terms equal to 1, 4, 9, 16, 25 and 36, and all correlations equal to 0.7. While we would not anticipate our shrinkages toward intraclass form to perform well, they can in fact still fare quite reasonably, for some choices of κ_1 and κ_2, indicating that our procedures are potentially reasonably robust to the choice of hypothesized model. The diagonal and nondiagonal components of risk for Λ were 0.274 and 0.236, and for $\tilde{\Lambda} = \log(\tilde{\mathbf{U}}/46)$ these reduced to 0.242 and 0.236. These values may be compared with the components of risk for our approximate Bayes estimator when $\kappa_1 = 2$ and $\kappa_2 = -2$, that is, 0.269 and 0.192. In addition to the values in Table 3, the choices $\kappa_1 = 2$ and $\kappa_2 = -3$ yield components of risk equal to 0.266 and 0.151, respectively.

For covariance matrix $\mathbf{C}_3$, the diagonal and nondiagonal components of risk for $\mathbf{S} = \tilde{\mathbf{U}}/50$ were 92.8 and 92.2, and the multiple $\tilde{\mathbf{U}}/52$ reduced these to 89.0 and 87.2, respectively. The choices $\kappa_1 = 2$ and $\kappa_2 = -2$ yield components of risk 84.5 and 81.5. Our results suggest that, as far as risk properties are concerned, we can still do well even if we smooth Λ towards an incorrect null hypothesis, but that it is always essentially important to carefully judge, via κ_1 and κ_2, how much to smooth toward any particular hypothesis. Hence the hierarchical and empirical Bayes procedures of Section 4 deserve close attention.

APPENDIX 3

Representations of exchangeable distributions. Suppose that the upper triangular elements $\{a_{ij}\}$, of the symmetric matrix $\mathbf{A} = \log \mathbf{C}$, satisfy

$$(1) \qquad a_{ij} = f\left(\mu, \lambda_i, \lambda_j, \lambda_{ij}^{AB}, \xi_i \delta_{ij}\right) \qquad i = 1, \ldots, j;\ j = 1, \ldots, p,$$

where δ_{ij} denotes the Kronecker delta function, f denotes some mapping from R^5 to R^1 and the $p(p+5)/2 + 1$ variables $\mu, \lambda_1, \ldots, \lambda_p, \xi_1, \ldots, \xi_p$ and λ_{ij}^{AB} $(i = 1, \ldots, j;\ j = 1, \ldots, p)$ are independent, and each uniformly distributed on (0, 1). Then (1) provides a very general exchangeable distribution for $\mathbf{A}$, and hence for the positive definite matrix $\mathbf{C} = \exp(\mathbf{A})$, which modifies a suggestion made by Aldous (1981), by constraining $\mathbf{A}$ to be symmetric. The coauthors and Grant Izmirlian are currently investigating the necessity, as $p \to \infty$, for a structure of the form (1).

It can alternatively be assumed that

$$(2) \qquad a_{ij} = \mu + \lambda_i + \lambda_j + \lambda_{ij}^{AB} + \xi_i \delta_{ij} \qquad i = 1, \ldots, j;\ j = 1, \ldots, p,$$

where μ possesses conditional c.d.f. G_μ, and $\{\lambda_i\}$, $\{\xi_i\}$ and $\{\lambda_{ij}^{AB}\}$ comprise three independent random samples, conditional on their respective common c.d.f.'s G_λ, G_{AB} and G_ξ. A further joint distribution may be assigned to the c.d.f.'s G_μ, G_λ, G_{AB} and G_ξ. Again, the distribution of $\mathbf{C}$ is exchangeable.

APPENDIX 4

Numerical results. In Section 8, the analysis of the Project Talent data refers to the following nine matrices.

Matrix A (sample correlation matrix):

0.100	0.760	0.681	0.734	0.608	0.559	0.434	0.647
0.760	0.100	0.598	0.679	0.463	0.489	0.518	0.482
0.681	0.598	0.100	0.663	0.407	0.367	0.401	0.681
0.734	0.679	0.663	0.100	0.591	0.494	0.506	0.631
0.608	0.463	0.407	0.591	0.100	0.517	0.402	0.469
0.559	0.489	0.367	0.494	0.517	0.100	0.548	0.531
0.434	0.518	0.401	0.506	0.402	0.548	0.100	0.493
0.647	0.482	0.681	0.631	0.469	0.531	0.493	0.100

Matrix A2 (maximum likelihood matrix Λ):

−5.356	0.694	0.378	0.475	0.395	0.299	0.018	0.405
0.694	−4.692	0.304	0.504	0.139	0.219	0.355	0.056
0.378	0.304	−5.720	0.391	0.047	−0.034	0.092	0.527
0.475	0.504	0.391	−4.115	0.453	0.189	0.306	0.441
0.395	0.139	0.047	0.453	−3.772	0.363	0.190	0.210
0.299	0.219	−0.034	0.189	0.363	−4.004	0.438	0.351
0.018	0.355	0.092	0.306	0.190	0.438	−3.656	0.316
0.405	0.056	0.527	0.441	0.210	0.351	0.316	−4.158

Matrix A3 (estimated standard errors for $\mathbf{\Lambda}$):

0.150	0.107	0.108	0.102	0.099	0.101	0.097	0.102
0.107	0.151	0.105	0.107	0.105	0.106	0.105	0.106
0.108	0.105	0.155	0.100	0.096	0.099	0.094	0.101
0.102	0.107	0.100	0.149	0.108	0.108	0.107	0.108
0.099	0.105	0.096	0.108	0.154	0.110	0.110	0.108
0.101	0.106	0.099	0.108	0.110	0.154	0.110	0.109
0.097	0.105	0.094	0.107	0.110	0.110	0.155	0.108
0.102	0.106	0.101	0.108	0.108	0.109	0.108	0.152

Matrix A4 (posterior expectation of $\mathbf{A}$ under uniform prior):

−5.292	0.695	0.379	0.473	0.395	0.294	0.014	0.408
0.695	−4.630	0.302	0.501	0.135	0.219	0.354	0.052
0.379	0.302	−5.657	0.392	0.051	−0.033	0.092	0.527
0.473	0.501	0.392	−4.057	0.453	0.189	0.307	0.440
0.395	0.136	0.051	0.453	−3.709	0.363	0.186	0.211
0.294	0.219	−0.033	0.189	0.363	−3.940	0.433	0.350
0.014	0.353	0.092	0.307	0.186	0.433	−3.588	0.320
0.408	0.052	0.527	0.440	0.211	0.350	0.320	−4.097

Matrix A5 (posterior standard errors for $\mathbf{A}$ under uniform prior):

0.152	0.108	0.112	0.104	0.101	0.105	0.100	0.104
0.108	0.154	0.110	0.110	0.108	0.108	0.109	0.108
0.112	0.110	0.160	0.104	0.101	0.101	0.097	0.104
0.104	0.110	0.104	0.147	0.111	0.110	0.108	0.110
0.101	0.108	0.101	0.111	0.162	0.114	0.114	0.109
0.105	0.108	0.101	0.110	0.114	0.170	0.114	0.113
0.100	0.109	0.097	0.108	0.114	0.114	0.167	0.112
0.104	0.108	0.104	0.110	0.109	0.113	0.112	0.156

Matrix A6 (posterior expectation of $\mathbf{A}$ under empirical exchangeable prior):

−5.210	0.528	0.342	0.421	0.363	0.300	0.133	0.365
0.528	−4.619	0.289	0.420	0.204	0.249	0.330	0.160
0.342	0.289	−5.599	0.340	0.125	0.077	0.153	0.422
0.421	0.420	0.340	−4.056	0.393	0.240	0.307	0.386
0.363	0.204	0.125	0.394	−3.767	0.332	0.232	0.248
0.300	0.249	0.077	0.240	0.332	−3.981	0.380	0.332
0.133	0.330	0.153	0.307	0.232	0.380	−3.649	0.312
0.365	0.160	0.422	0.386	0.248	0.332	0.312	−4.121

Matrix A7 (posterior standard errors for $\mathbf{A}$ under empirical exchangeable prior):

0.153	0.088	0.089	0.085	0.082	0.084	0.081	0.084
0.088	0.150	0.085	0.086	0.084	0.085	0.085	0.085
0.089	0.085	0.158	0.082	0.080	0.081	0.080	0.084
0.085	0.086	0.082	0.146	0.086	0.086	0.086	0.087
0.082	0.084	0.080	0.086	0.153	0.086	0.086	0.086
0.084	0.085	0.081	0.086	0.086	0.152	0.087	0.086
0.081	0.085	0.080	0.086	0.086	0.087	0.151	0.085
0.084	0.085	0.084	0.087	0.086	0.086	0.085	0.150

Matrix A8 (posterior expectation of population correlation matrix under empirical exchangeable prior):

0.100	0.701	0.653	0.680	0.597	0.573	0.487	0.627
0.701	0.100	0.589	0.630	0.489	0.512	0.528	0.506
0.653	0.589	0.100	0.624	0.462	0.445	0.463	0.636
0.680	0.630	0.624	0.100	0.565	0.510	0.514	0.595
0.597	0.489	0.462	0.565	0.100	0.509	0.438	0.487
0.573	0.512	0.445	0.510	0.509	0.100	0.526	0.532
0.487	0.528	0.463	0.514	0.438	0.526	0.100	0.504
0.627	0.506	0.636	0.595	0.487	0.532	0.504	0.100

Matrix A9 (posterior expectation of **A** under hierarchical exchangeable prior):

−5.243	0.550	0.345	0.423	0.370	0.303	0.126	0.376
0.550	−4.640	0.284	0.429	0.196	0.246	0.338	0.147
0.345	0.284	−5.615	0.342	0.126	0.067	0.145	0.432
0.423	0.429	0.342	−4.042	0.398	0.240	0.309	0.393
0.370	0.196	0.126	0.398	−3.743	0.329	0.226	0.246
0.303	0.246	0.067	0.241	0.329	−3.939	0.373	0.332
0.126	0.338	0.145	0.309	0.226	0.373	−3.638	0.314
0.376	0.147	0.432	0.393	0.246	0.332	0.314	−4.118

Acknowledgments. The authors are indebted to Dennis Cox for suggesting the exponential matrix transformation, and describing some of the basic properties of this transformation to Ben Noble for advice concerning Volterra integral equations and to Dennis Lindley for suggesting the problem to the first co-author in 1970. Gwyn Evans, Tom Stroud, Tom Y. M. Chiu, Kam-Wah Tsui, Richard Johnson, Christian Ritter, Irwin Guttman, Grant Izmirlian, Jim Press, Jim Dickey, an Associate Editor and two referees all gave valuable advice.

REFERENCES

ALBERT, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *J. Amer. Statist. Assoc.* **83** 1037–1044.

ALDOUS, D. J. (1981). Representations for partially exchangeable array of random variables. *J. Multivariate Anal.* **11** 581–598.

ANDERSON, T. W. (1984). *Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley, New York.

BARTLETT, M. S. and KENDALL, D. G. (1946). The statistical analysis of variance—heterogeneity and the logarithmic transformation. *J. Roy. Statist. Soc. Ser. B* **8** 128–138.

BELLMAN, R. (1970). *Introduction to Matrix Analysis*. McGraw-Hill, New York.

CHAMBERLAIN, G. and LEAMER, E. E. (1976). Matrix weighted averages and posterior bounds. *J. Roy. Statist. Soc. Ser. B* **38** 73–84.

CHEN, C. F. (1979). Bayesian inference for a normal dispersion matrix and its applications to stochastic multiple regression analysis. *J. Roy. Statist. Soc. Ser. B* **41** 235–248.

COHEN, A. (1974). To pool or not to pool in hypothesis testing. *J. Amer. Statist. Assoc.* **69** 721–725.

COOLEY, W. W. and LOYNES, B. P. R. (1971). *Multivariate Data Analysis*. Wiley, New York.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DICKEY, J. M., LINDLEY, D. V. and PRESS, S. J. (1985). Bayesian estimation of the dispersion matrix of a multivariate normal distribution. *Comm. Statist. Theory Methods* **14** 1019–1034.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

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

EVANS, I. G. (1965). Bayesian estimation of parameters of a multivariate normal distribution. *J. Roy. Statist. Soc. Ser. B* **27** 279–283.

FLANAGAN, J. C., DAVIS, F. B., DAILEY, J. T., SHAYCOFT, M. F., ORR, D. B., GOLDBERG, I. and NEYMAN, C. A. (1964). *Project TALENT*. Univ. Pittsburgh.

GEWEKE, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *J. Econometrics* **57** 1317–1339.

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

HSU, J. S. J., LEONARD, T. and TSUI, K. W. (1991). Statistical inference for multiple choice tests. *Psychometrika* **55** 327–348.

JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41** 851–864.

KASS, R. E. and STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* **84** 717–726.

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* **60** 297–308.

LEONARD, T. (1975). A Bayesian approach to the linear model with unequal variances. *Technometrics* **17** 95–102.

LEONARD, T. (1976). Some alternative approaches to multiparameter estimation. *Biometrika* **63** 69–76.

LEONARD, T. (1977). A Bayesian approach to some multinomial estimation and pre-testing problems. *J. Amer. Statist. Assoc.* **72** 869–874.

LEONARD, T., HSU, J. S. J. and RITTER, C. (1993). The Laplacian T-approximation in Bayesian inference. Unpublished manuscript.

LEONARD, T., HSU, J. S. J. and TSUI, K. (1989). Bayesian marginal inference. *J. Amer. Statist. Assoc.* **84** 1051–1058.

LEONARD, T. and NOVICK, M. R. (1986). Bayesian full rank marginalization for two-way contingency tables. *Journal of Educational Statistics* **11** 33–56.

LEONARD, T. and ORD, K. (1976). An investigation of the F test procedure as an estimation short-cut. *J. Roy. Statist. Soc. Ser. B* **38** 95–98.

LINDLEY, D. V. (1971). The estimation of many parameters. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 435–455. Holt, Rinehart, and Winston, Toronto.

MORRIS, C. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65.

O'HAGAN, A. (1976). On posterior joint and marginal modes. *Biometrika* **63** 329–333.

PRESS, S. J. (1992). Bayes testing for equality of multivariate normal covariance matrices. Paper presented at the NSF–NBER Seminar on Bayesian Inference in Econometrics and Statistics, Washington Univ., St. Louis, April 1992.

STEIN, C. (1975). Estimation of a covariance matrix. IMS–ASA Annual Meeting (and unpublished lecture notes).

STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1210 WEST DAYTON STREET
MADISON, WISCONSIN 53706-1693

DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
UNIVERSITY OF CALIFORNIA
SANTA BARBARA, CALIFORNIA 93106