The Annals of Statistics
1991, Vol. 19, No. 3, 1163–1190

THE VARIATIONAL FORM OF CERTAIN BAYES ESTIMATORS[1]

BY L. R. HAFF

University of California, San Diego

A general representation is obtained for the formal Bayes estimator of a parameter matrix. We assume that the prior distribution is symmetric in some sense, but it is not specified otherwise. The formal Bayes risk is minimized subject to order constraints by a variational technique; hence our representation is called "the variational form of the Bayes estimator" (VFBE). The VFBE is used to obtain estimators that have good frequency properties relative to the usual estimators. Such estimators are obtained for the mean vector and covariance matrix of a multivariate normal distribution. Also, for possibly nonnormal data, we give the VFBE of several Pearson means. A certain emphasis is placed on the problem of estimating the covariance matrix. For that problem, our constrained optimization provides an estimator with very good properties: Its eigenvalues are in the proper order, and they are not as distorted as those in the sample covariance matrix. The VFBE for the covariance matrix is related to an estimator of Stein. Of the two, the VFBE deals with order relations in a more natural way; that is, it is more criterion dependent. In addition, it is easier to compute than Stein's estimator, and a brief Monte Carlo simulation indicates that it has better risk properties as well.

Introduction and summary. A general representation for the Bayes estimator of a parameter matrix Ψ is obtained; it is dubbed the variational form of the Bayes estimator (VFBE). Some special cases are developed which outperform the usual estimators in the frequency sense. In this section we discuss the nature of the VFBE and outline the main results.

The main results pertain to estimating the mean vector and covariance matrix of the multivariate normal distribution, with an emphasis on the latter problem. Hence, we start with the following.

Notation. Decision theoretics. Assume that $\mathbf{X}$, $p \times 1$, has a multivariate normal distribution with mean vector θ and covariance matrix Σ. Assume also that $\mathbf{S}$, $p \times p$, has a Wishart distribution with matrix Σ, $k - p - 1 > 0$ and $\mathbf{X}$ and $\mathbf{S}$ are independent; that is,

$$(0.1)\quad \mathbf{X} \sim \mathcal{N}_p(\theta, \Sigma) \quad \text{and} \quad \mathbf{S} \sim \mathcal{W}_p(\Sigma, k) \quad \text{with } \mathbf{X} \text{ and } \mathbf{S} \text{ independent.}$$

Received January 1983; revised July 1990.
[1]Research supported by an NSF grant.
AMS 1980 *subject classifications*. Primary 62H12; secondary 62C99.
Key words and phrases. Mean vector, spherically symmetric estimators, covariance matrix, orthogonally invariant estimators, Pearson curves, unbiased estimation of risk, variational form of Bayes estimators, Euler equations, eigenvalue distortion in sample covariance matrix, estimation of eigenvalues.

As separate problems, we consider the estimation of θ and Σ. We assume that an estimator $\hat{\theta}$ suffers a loss

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^t \Sigma^{-1} (\hat{\theta} - \theta). \tag{0.2}$$

We assume also that $\hat{\Sigma}$ suffers

$$\begin{aligned} L_1(\hat{\Sigma}, \Sigma) &= \operatorname{tr}(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1}) - p \quad \text{or} \\ L_2(\hat{\Sigma}, \Sigma) &= \operatorname{tr}(\hat{\Sigma}\Sigma^{-1} - I)^2, \end{aligned} \tag{0.3}$$

where L_1 and L_2 are treated separately. In addition, we briefly consider the estimation of several Pearson means under quadratic loss: For $\mathbf{X}$ $(p \times 1)$ a vector of independent Pearson variables, $\theta \equiv E[\mathbf{X}|\theta]$ the vector of means and $\hat{\theta}$ an estimator of θ, we assume that

$$L_Q(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^t Q (\hat{\theta} - \theta), \tag{0.4}$$

with $Q = \operatorname{diag}(q_1, q_2, \ldots, q_p)$ a diagonal matrix of specified constants. We assume that only the modes of the Pearson distributions are unknown, but this framework does include several nonnormal cases that have appeared in the literature. The main purpose of this example is to display the VFBE in a nonnormal setting. See Haff and Johnson (1986) for further results and references.

Denote by $(\Psi, \hat{\Psi}, L)$ any of the preceding estimation problems. In addition, denote by $\Pi(\Psi)$ the prior distribution of Ψ. Then the risk and Bayes risk are given by

$$\rho(\hat{\Psi}, \Psi) = E\big[L(\hat{\Psi}, \Psi)\big|\Psi\big] \quad \text{and} \quad \rho(\hat{\Psi}, \Pi) = E\big[\rho(\hat{\Psi}, \Psi)\big], \tag{0.5}$$

respectively. [Often we write $E_\Psi(\cdot)$ instead of $E[\cdot|\Psi]$.] Given $\hat{\Psi}^*$ and $\hat{\Psi}$, two estimators of Ψ, then $\hat{\Psi}^*$ dominates $\hat{\Psi}$ if $\rho(\hat{\Psi}^*, \Psi) \le \rho(\hat{\Psi}, \Psi)$ for all Ψ, with strict inequality at some Ψ. Our strategy is to let the formal Bayes rule suggest estimators which perform well in the frequency sense. Particularly, we seek estimators which dominate the usual estimators.

The VFBE. Let $\mathbf{A}$ be a random matrix with p.d.f. $f(A|\Psi)$ with respect to Lebesgue measure on Euclidean space of appropriate dimension. Then the marginal density of $\mathbf{A}$ is given by

$$f_\Pi(A) = \int f(A|\Psi)\, d\Pi(\Psi). \tag{0.6}$$

The VFBE depends explicitly on $f_\Pi(A)$. Now we comment on the derivation of the VFBE and also on the applications of the result.

The VFBE is obtained as follows: For $(\Psi, \hat{\Psi}(\mathbf{A}), L)$ we assume that an unbiased estimator $\hat{\rho}(\hat{\Psi}, \Psi)$ of $\rho(\hat{\Psi}, \Psi)$ is available [i.e., $E_\Psi \hat{\rho}(\hat{\Psi}, \Psi) = \rho(\hat{\Psi}, \Psi)$]. (Such risk estimators exist for many estimation problems, and several refer-

ences are found in the following.) From Bayes' formula, we thus obtain

$$\rho(\hat{\Psi}, \Pi) = EE_{\Psi}\hat{\rho}(\hat{\Psi}, \Psi) = \int \hat{\rho}(\hat{\Psi}, \Psi) f_{\Pi}(A)\, dA, \tag{0.7}$$

provided $\int |\hat{\rho}(\hat{\Psi}, \Psi)| f_{\Pi}(A)\, dA < \infty$. Given that $\hat{\rho}(\hat{\Psi}, \Psi)$ is written as a function of $\hat{\Psi}$ and $\partial\hat{\Psi}/\partial A$, the last integral is minimized by solutions of the Euler equations. In this way, we obtain a general representation of the formal Bayes rule (i.e., the VFBE) that depends explicitly on f_{Π}. This representation is appropriate for an important class of problems that includes (0.1)–(0.4).

Typically, the VFBE is simplified if we assume that $\Pi(\Psi)$ is symmetric in some sense. Symmetry is often assumed in the following, but $\Pi(\Psi)$ is not specified otherwise. Following reduction by symmetry, we mostly replace $f_{\Pi}(A)$ (in the VFBE) by a function that is not a formal density in the sense of (0.6). Such substitutions are in line with our strategy of letting the VFBE suggest estimators with good frequency properties. These comments are illustrated by the following.

The prototypical case. For $\mathbf{X} \sim \mathcal{N}_p(\theta, I)$ and $\Pi(\theta)$ a (possibly improper) prior distribution, the formal Bayes estimator can be represented by

$$\hat{\boldsymbol{\theta}}_B = \mathbf{X} + \nabla \log f_{\Pi}(\mathbf{X}) \tag{0.8}$$

in which $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \ldots, \partial/\partial x_p)^t$; see Brown (1971) or Stein (1981). This representation is the prototype of the VFBE. It is our first example and it motivates the approach that is taken.

One way of using (0.8) is suggested by a result of Stein (1981); namely, if $p \geq 3$ and

$$\Sigma_i \partial^2 f_{\Pi}(X)^{1/2}/\partial x_i^2 \leq 0 \quad \text{a.e.} \tag{0.9}$$

[i.e., if $f_{\Pi}(X)$ is superharmonic], then $\mathbf{X}$ is dominated by (0.8). [Any superharmonic f_{Π} will do here; it need not be interpreted by (0.6).] Another way of using (0.8) proceeds as follows: If $\Pi(\theta)$ is spherically symmetric; that is, if $\Pi(\theta) = \Pi(O\theta)$ with O an arbitrary orthogonal matrix, then we obtain $f_{\Pi}(X) = w^{-(p-2)/2} g_{\Pi}(w)$ in which $g_{\Pi}(\cdot)$ is the density of $\mathbf{w} = \mathbf{X}^t\mathbf{X}$. Thus (0.8) becomes

$$\hat{\boldsymbol{\theta}}_B = [1 - \varphi_B(\mathbf{w})]\mathbf{X}, \tag{0.10}$$

in which $\varphi_B(\mathbf{w}) = (p-2)/\mathbf{w} - 2(d/dw)\log g_{\Pi}(\mathbf{w})$. Note that $\mathbf{w}$ is a maximal invariant under $\mathbf{X} \to O\mathbf{X}$ and $\theta \to O\theta$. If we set $g_{\Pi}(w) = \text{const.}$, then (0.10) becomes

$$\hat{\boldsymbol{\theta}}_S \equiv [1 - (p-2)/\mathbf{w}]\mathbf{X}, \qquad \mathbf{w} = \mathbf{X}^t\mathbf{X}, \tag{0.11}$$

the James–Stein (1961) estimator. It is known that (0.11) dominates $\mathbf{X}$ (and is thus minimax). This example is typical of our results for (0.2) or (0.3) when Π is suitably invariant. That is, for $\mathbf{w}$ judiciously defined, the VFBE features those terms which are prominent in a Stein-like estimator. Thus we might choose $g_{\Pi}(w)$ in a naive way, perhaps, and then check on the risk properties of the VFBE.

Observe that (0.11) admits a standard Bayesian interpretation; namely, $(d/dw)\log g_\Pi(\mathbf{w})$ may be negligible in (0.10) if Π *is specified* so that some 1-1 function of $\mathbf{w}$ has a "vague" distribution. In this connection, however, we have not found a prior distribution Π that yields $g_\Pi(w) = \text{const.}$ (a.e.).

Finally, some comments on terminology: It is clear from (0.10) that the VFBE is not unique (since any 1-1 function of $\mathbf{w}$ is also a maximal invariant). Consequently, for any problem $(\Psi, \hat{\Psi}, L)$, we refer to the VFBE with the understanding that both $\mathbf{w}$ and $g_\Pi(w)$ have been specified. Also, we will use the acronym VFBE though our estimator is often not Bayes or even formal Bayes. [It is Bayes if and only if f_Π (or g_Π) is interpreted by (0.6) with Π a proper prior distribution; it is formal Bayes if Π is improper.] This abuse of terminology allows for quick reference, while the precise nature of the estimator (formal Bayes or otherwise) will be obvious from the context.

Literature. Perhaps dozens of articles have referenced Stein's inequality (0.9) and its implication for minimax estimators. George (1986) used (0.9) to construct multiple shrinkage estimators (where $\Sigma = I$). For the Pearson curves problem (0.4), Haff and Johnson (1986) gave the appropriate generalization of (0.9) after a transformation to natural variables. Also, implicit in Corollary 4.5 of Haff and Johnson (1986) is a specialization of (0.9) for $g_\Pi(w)$.

1. The main results.

The constrained VFBE. First we develop a scheme for improving upon (0.11) and certain of its analogues. One major difficulty with these estimators is the following: When the VFBE is modified by replacing f_Π (or g_Π) by a function that is not a marginal p.d.f., the result can be dominated by certain of its truncated versions. [It is known, e.g., that (0.11) suffers this defect.] Therefore, we are guided by the solution of the following problem: For $\hat{\Psi}$ constrained to a complete class, and under suitable regularity,

$$\text{minimize} \int \hat{\rho}(\hat{\Psi}, \Psi) f_\Pi(A)\, dA, \tag{1.1}$$

given that f_Π (or g_Π) is *any* function for which this problem is well posed. [In particular, this problem is well posed if f_Π (or g_Π) is interpreted by (0.6) and if the integral is finite.] The solution of (1.1) describes the class of estimators with which we start. Our strategy is to exploit this class in order to find estimators with good frequency properties. In particular, it is enlarged by allowing naive choices for f_Π (or g_Π) which initially are suggested by results on Stein-like estimation.

The constrained problem (1.1) is solved by using a slack variable technique, and we return to (0.10) and the equations that follow it to introduce the main idea. First, since Π is orthogonally invariant, the feasible estimators are of the form $(1 - \varphi(\mathbf{w}))\mathbf{X}$ in which φ is real. From Anderson [(1984), page 91], any

such estimator is dominated by $(1 - \varphi^*(\mathbf{w}))\mathbf{X}$ in which $\varphi^*(\mathbf{w}) = \min\{1, \varphi(\mathbf{w})\}$. A complete class is thus formed by the spherically symmetric estimators with $\varphi(\mathbf{w}) \leq 1$. In the following, we show that the solution of (1.1) subject to $\varphi(\mathbf{w}) \leq 1$ is given by

$$\hat{\boldsymbol{\theta}} = [1 - \min\{1, \varphi_B(\mathbf{w})\}]\mathbf{X}, \tag{1.2}$$

which dominates (0.10). The slack variable technique is accompanied by a finite algorithm. In this example, the solution (1.2) requires a single step only. [Note that $g_\Pi(w) = \text{const.}$ yields the positive-part James–Stein estimator.]

Our main results on estimating means pertain to the unconstrained case. Though certain extensions of (1.2) might well be worthwhile, the main contribution of (1.2) in the present article is to motivate our work on estimating covariance matrices.

Estimation of the mean vector. Assume that $\mathbf{X} \sim \mathcal{N}_p(\theta, I)$ and $\mathbf{S} \sim \mathcal{W}_p(\Sigma, k)$ with both θ and Σ unknown, $\mathbf{X}$ and $\mathbf{S}$ independent. First, the natural extension of (0.8) is given by the general formula for the VFBE. A generalization of (0.9) is then obtained under which this estimator is minimax. A special case of this minimax result appears in Bilodeau and Kariya [(1989), Sect. 4]. Finally, the natural extension of (0.10) follows from the assumption that $\Pi(\theta|\Sigma)$ is an elliptical distribution. A special case of the latter was used by Lin and Tsai (1973) to develop minimax results. Our discussion on estimating means concludes with a brief illustration using Pearson curves.

Estimation of covariance matrices. Denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $\mathbf{l}_1 \geq \mathbf{l}_2 \geq \cdots \geq \mathbf{l}_p \geq 0$ the ordered eigenvalues of Σ and $\mathbf{S}$, respectively, and set $\lambda = (\lambda_1, \ldots, \lambda_p)$ and $\mathbf{l} = (\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p)$. Assume $\Pi(\Sigma)$ is orthogonally invariant, that is, $\Pi(O\Sigma O^t) = \Pi(\Sigma)$, so the VFBE is completely determined by estimates of λ. For a certain $g_\Pi(l)$, the analogue of φ_B [see (0.10)] is an estimate of λ introduced by Stein (1975) in a Rietz lecture. (This is our unconstrained solution, the "rough estimate.") This estimate is not always faithful to the natural order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Hence, the natural constraint in (1.1) is $\varphi_1 \geq \varphi_2 \geq \cdots \geq \varphi_p \geq 0$, where $\varphi_i = \varphi_i(\mathbf{l})$, $i = 1, \ldots, p$, are the feasible estimates. The minimization is done by the same technique that gave (1.2), and completely analogous results are obtained. Further, our constrained estimates of λ are seen to correct the well known distortion in the spectrum of $\mathbf{S}$.

It is natural to compare our estimator of Σ with that of Stein (1975), who obtained the rough estimates of λ by heuristic minimization of the unbiased estimator of the risk function (not the Bayes risk). He adjusted these by a pooling algorithm that gave a set of nonnegative estimates and then did isotonic regression on the result to impose the proper order. Several Monte Carlo studies have indicated that Stein's estimator does remarkably well in spite of its ad hoc nature. In fact, it might be the best estimator (in terms of

risk) that has appeared in the literature. There are indications, however, that our present estimator has some advantages and offers further improvement yet. A brief synopsis of these indications follows.

1. Our constrained minimization is criterion dependent; that is, it is dictated by the loss function L_1. Isotonic regression, on the other hand, is a separate criterion. Thus we expected some improvement over Stein's estimator in terms of risk. However, increased efficiency was not observed in our Monte Carlo output until we took values of p in the range $15 \leq p \leq 20$. The constrained VFBE then performed significantly better than Stein's estimator (in the statistical sense), but the reduction in risk relative to the usual estimator was only about 2%.
2. More favorable comparisons were made under L_2, even though our own estimator takes on an ad hoc nature in this case. Here, we rescaled the estimators (derived under L_1) as in Lin and Perlman (1985). This made comparisons under L_2 possible. For this case, the constrained VFBE outperformed Stein's estimator by some 4–10% (L_2). The results so far (for both L_1 and L_2) were obtained for $p \in \{5, 10, 15, 20\}$ and $k = 2p$. We checked other combinations with $p < k < 2p$. In particular, for $p = 20$, $k = 21$ and $\Sigma = I$, the VFBE was best (L_2) by 20%—a surprising result. For $\Sigma = I$, the favorable comparisons gradually dropped off to 9% as k increased from 21 to 40 (or $2p$).
3. Stein's estimator appears intractable for risk calculations, and we have not yet obtained dominance results for the VFBE either. Such results are not out of the question, however, especially for $p = 2$ or 3. To start with, the unbiased estimator of the risk function is valid and obtainable for the VFBE. It is valid for Stein's estimator, also, but it seems practically unobtainable in this case. [The validity of the unbiased estimator of the risk function for both of these follows essentially from the same argument that shows that the unbiased estimator of the risk function is valid for the positive-part versions (1.2).]

In addition to the work on estimation, some useful computational results are found in this article. Lemma 4.2 and Theorem 6.1 together provide an effective means of computing unbiased risk estimators in fairly complex situations. The point is illustrated by our results for loss function L_2. (Most of the L_2 details have been omitted, but they are available upon request.)

Several workers have either used our basic approach or referenced the computational results since the first version of this article was submitted. Among others, Muirhead and Verathaworn (1983) used our method to estimate eigenvalues in a two-population setting. Dey and Srinivasan (1985) and Dey (in several papers) have used Theorem 6.1 in various estimation problems. Also, Loh (1988a, b) made effective use of the slack variable technique. Loh's papers provide further evidence that our method is more efficient than Stein's (in terms of risk). Finally, the analogue of our variational method was worked out in the discrete setting by Alcaraz (1990), and minimax results were obtained for the estimation of several Poisson means.

2. The general Euler equations. Assume that $\mathbf{A}$ is a random matrix with p.d.f. $f(A|\Psi)$ with respect to the Lebesgue measure on R^n. Assume further that $\Psi = (\Psi_1, \Psi_2)$ in which $\Psi_i \in R^{p_i}$, $i = 1, 2$. The vector Ψ_1 is the parameter of interest and Ψ_2 may be regarded as a nuisance parameter. If nuisance parameters are not present then we set $\Psi \equiv \Psi_1$. [In $(\Sigma, \hat{\Sigma}, L_i)$, for example, $\Psi \equiv \Sigma$ is the parameter of interest and both S and Σ are embedded in $R^{p(p+1)/2}$.] Throughout, we assume that $L(\hat{\Psi}_1, \Psi_1)$ has been specified and, importantly, that an unbiased estimator $\hat{\rho}(\hat{\Psi}, \Psi_1)$ of $\rho(\hat{\Psi}_1, \Psi_1)$ is available. In this section, the VFBE of Ψ_1 is defined by the solution of a system of Euler equations.

Now assume that $(\Psi_1, \hat{\Psi}_1, L)$ is invariant under a transformation group and that the invariant estimators are of the form

$$\hat{\Psi}_1 = a[\mathbf{A}, \varphi(\mathbf{w})] \in R^{p_1} \tag{2.1}$$

for a specified function $a[\cdot, \cdot]$ in which (a) $\mathbf{w} \in R^m$ is a maximal invariant (b) φ is arbitrary, φ: $R^m \to R^q$; hence a: $R^{n+q} \to R^{p_1}$ for positive integers n, m, q and p_1.

The class of estimators under consideration is given by the following subset of the invariant class.

DEFINITION 2.1 (Feasible estimators). Assume that $(\Psi_1, \hat{\Psi}_1, L)$ is invariant under a transformation group and that the invariant estimators are given by (2.1). The feasible estimators are those in (2.1) for which an unbiased estimator $\hat{\rho}(\hat{\Psi}_1, \Psi_1)$ of $\rho(\hat{\Psi}_1, \Psi_1) \equiv E_\Psi L(\hat{\Psi}_1, \Psi_1)$ exists and can be written as

$$\hat{\rho}\big(\hat{\Psi}_1, \Psi_1\big) = \not{\rho}[\mathbf{w}, \varphi(\mathbf{w}), d\varphi(\mathbf{w})], \tag{2.2}$$

where $\not{\rho}$: $R^{m+(m+1)q} \to R$ in which $d\varphi(\mathbf{w}) \equiv [(\partial\varphi_i/\partial w_j)(\mathbf{w})]$, $q \times m$. If the problem is not invariant under a transformation group, then we set $\mathbf{w} = \mathbf{A}$ and the feasible estimators are those of the form $\hat{\Psi}_1 = a[\mathbf{A}, \varphi(\mathbf{A})] \equiv \varphi(\mathbf{A})$, $q = p_1$, for which an unbiased estimator of the risk function exists.

LITERATURE. Conditions under which unbiased estimators of the risk function exist for various situations are found in the papers of Hudson (1978), Haff (1979b, 1981), Berger (1980), Stein (1981) and Haff and Johnson (1986).

The VFBE is given by the following theorem.

THEOREM 2.1. *Let $g(w|\Psi)$ be the p.d.f. of $\mathbf{w}$ given Ψ with support on $w \subset R^m$. Let $\Pi(\Psi)$ be the (possibly improper) prior distribution of Ψ and set $g_\Pi(w) = \int g(w|\Psi)\, d\Pi(\Psi)$. For the problem $(\Psi_1, \hat{\Psi}_1, L)$, assume that a formal Bayes estimator exists and is a feasible estimator. In addition, assume the following.*

(i) *The function $\not{\rho}[w, \varphi(w), d\varphi(w)] \equiv \not{\rho}[w, \varphi(w), d\varphi(w)]g_\Pi(w)$ is twice continuously differentiable, $R^{m+(m+1)q} \to R$.*

(ii) *For any constant* ε *and any vector* $\beta(w) = (\beta_1(w), \beta_2(w), \ldots, \beta_q(w))^t$ *such that* $a[\mathbf{A}, \varphi(\mathbf{w}) + \varepsilon\beta(\mathbf{w})]$ *is a feasible estimator, we have*

$$\int_{\Gamma_j}\left[g_\Pi(w)|\beta_i(w)|\left(\sum_{c=1}^{m}(\partial\not{f}/\partial\varphi'_{ic})^2\right)^{1/2}\right]d\tau \to 0 \quad \text{as } j \to \infty,$$

for each $i = 1, 2, \ldots, q$. *The integral is over* Γ_j, *the surface of* w_j, $j = 1, 2, \ldots$, *where* w_j *is an increasing sequence of simple regions such that* $w_j \uparrow w$. *Also,* τ *is surface measure and* φ'_{ic} *is that argument of* $\not{f}$ *associated with the variable* $\partial\varphi_i/\partial w_c$.

Finally, if the problem $(\Psi_1, \hat{\Psi}_1, L)$ *is strictly convex in* $\hat{\Psi}_1$, *then the formal Bayes rule is unique and must be a solution of*

$$\nabla_\varphi \not{f} = \left[\nabla_w^t \nabla_{\varphi'}^t\right]^t \not{f} + \left[\nabla_{\varphi'}\not{f}\right]\left[\nabla_w \log g_\Pi(w)\right], \tag{2.3}$$

in which $\nabla_\varphi \equiv (\partial/\partial\varphi_1, \partial/\partial\varphi_2, \ldots, \partial/\partial\varphi_q)^t$, $\nabla_w \equiv (\partial/\partial w_1, \partial/\partial w_2, \ldots, \partial/\partial w_m)^t$ *and* $\nabla_{\varphi'} \equiv (\partial/\partial\varphi'_{ij})$, $q \times m$, *with* $\varphi'_{ij} \equiv \partial\varphi_i/\partial w_j$.

NOTATION. For future convenience, let us write [L] = [R1] + [R2] where [R1] depicts the first term on the right-hand side of (2.3) and so forth. This will simplify the presentation of special cases.

PROOF OF THEOREM 2.1. The uniqueness of the formal Bayes rule $a[\mathbf{A}, \varphi_\Pi(\mathbf{w})]$ follows from the convexity of the loss function. Now set $J_j[\varphi] = \int_{w_j} \not{f}[w, \varphi(w), d\varphi(w)]\,dw$ and let e_i be the ith unit vector in R^q. It is seen that the differential of $J_j[\varphi]$ in the ith direction is given by

$$(d/d\varepsilon)J_j[\varphi + \varepsilon\beta_i e_i]|_{\varepsilon=0} = \int_{w_j}\left[(\partial/\partial\varphi_i)\not{f} - \sum_{c=1}^{m}(\partial/\partial w_c)\left[(\partial/\partial\varphi'_{ic})\not{f}\right]\right]\beta_i\,dw$$

$$+ \int_{\Gamma_j}\langle G, \nu\rangle\beta_i\,d\tau$$

in which $\langle G, \nu\rangle$ is the usual inner product of

$$G = \left(\partial\not{f}/\partial\varphi'_{i1}, \partial\not{f}/\partial\varphi'_{i2}, \ldots, \partial\not{f}/\partial\varphi'_{im}\right)^t$$

and ν is the outward unit normal on Γ_j. The differential must be zero at the extremal $\varphi_\Pi(w)$. Note that the surface integral in the last equation goes to 0 as $j \to \infty$ by virtue of condition (ii) in the theorem. Thus, the first integral must be zero for each j, and this together with the arbitrary nature of β_i implies

$$(\partial/\partial\varphi_i)\not{f} - \sum_{c=1}^{m}(\partial/\partial w_c)\left[(\partial/\partial\varphi'_{ic})\not{f}\right] = 0,$$

$i = 1, 2, \ldots, q$, at $\varphi_\Pi(w)$. These equations can be written as

$$(\partial\not{f}/\partial\varphi_i)g_\Pi - \sum_{c=1}^{m}\left[(\partial^2\not{f}/\partial w_c\,\partial\varphi'_{ic})g_\Pi + (\partial\not{f}/\partial\varphi'_{ic})(\partial g_\Pi/\partial w_c)\right] = 0,$$

$i = 1, 2, \ldots, q$. Finally, (2.3) is obtained by dividing both sides of the last equation by g_Π. We omit further details. □

REMARK 2.1. Denote the unique solution of (2.3) by $\varphi_\Pi(w)$. This defines $a[\mathbf{A}, \varphi_\Pi(\mathbf{w})]$, a representation of the formal Bayes estimator that we call the variational form of the Bayes estimator (VFBE). Note the explicit dependence upon $\nabla_w \log g_\Pi(w)$.

REMARK 2.2. In (2.3), the notation $\nabla_{\varphi'} \not{\ }$ indicates postmultiplication of $\nabla_{\varphi'}$ by a scalar $\not{\ }$, followed by partial differentiation as indicated; thus the first term on the right-hand side of (2.3) is

$$\left[\nabla_w^t \nabla_{\varphi'}^t\right]^t \not{\ } \equiv \left[\nabla_w^t (\nabla_{\varphi'} \not{\ })^t\right]^t.$$

REMARK 2.3. Conditions (i) and (ii) of Theorem 2.1 are suggested by m-dimensional extensions of results found in Gelfand and Fomin [(1963), pages 152–154].

REMARK 2.4. We can extend Theorem 2.1 by allowing the arguments of $\not{\ }$ to include all the second-order terms $\partial^2 \varphi_k(w)/\partial w_i \, \partial w_j$. A case in which this structure is actually present is briefly discussed in Section 7.

3. Estimation of the mean vector. Assume the normal Wishart setup (0.1) with θ to be estimated subject to (0.2). In this section, the solutions of (2.3) are given for several special cases and a minimax theorem is proved where Σ is unknown. The solution of (2.3) is also displayed for the Pearson case.

Recall that $E(\theta|\mathbf{X})$, the usual expression for the Bayes rule, is computed by minimizing the posterior expected loss. The usual expression must be equivalent to the VFBE and connections between the two are illustrated among the special cases.

First, we need $\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\theta}}_\varphi, \theta)$ [recall (2.2) and (2.3)], and this estimator follows from two identities.

1. The normal identity of Stein (1981): For $\mathbf{X} \sim \mathcal{N}_p(\theta, \Sigma)$ and φ: $\mathbf{R}^p \to \mathbf{R}^p$, we have

$$E_\theta(\mathbf{X} - \theta)^t \Sigma^{-1} \varphi(\mathbf{X}) = E_\theta \, \nabla^t \varphi(\mathbf{X}), \tag{3.1}$$

   where φ is almost differentiable and $\nabla^t \varphi(\mathbf{X}) \equiv \Sigma_i \, \partial \varphi_i(\mathbf{X})/\partial \mathbf{x}_i$.
2. The Wishart identity of Stein (1977a, b) or Haff (1979a). Let $\mathbf{S} = (\mathbf{s}_{ij}) \sim \mathcal{W}_p(\Sigma, k)$ with Σ unknown. Also, let $D = (d_{ij})$ be a $p \times p$ operator matrix, $d_{ij} \equiv (1/2)(1 + \delta_{ij})(\partial/\partial s_{ij})$, where δ_{ij} is the Kronecker delta. For suitable $p \times p$ matrices $\mathbf{V} = (\mathbf{v}_{ij})$ we have

$$\operatorname{tr} E_\Sigma(\mathbf{V}\Sigma^{-1}) = 2 \operatorname{tr} E_\Sigma(D\mathbf{V}) + (k - p - 1) \operatorname{tr} E_\Sigma(\mathbf{S}^{-1}\mathbf{V}), \tag{3.2}$$

   in which $D\mathbf{V} \equiv (\Sigma_t \, d_{it} \mathbf{v}_{tj})$, $p \times p$. See Haff (1979b or 1981) for conditions

under which this identity is valid. See also Haff (1981, 1982) for certain generalizations.

Recall that the operators in (2.3) are given by

$$\nabla_\varphi \equiv \left(\partial/\partial\varphi_1, \partial/\partial\varphi_2, \ldots, \partial/\partial\varphi_q\right)^t, \qquad q \times 1,$$

$$\nabla_w \equiv \left(\partial/\partial w_1, \partial/\partial w_2, \ldots, \partial/\partial w_m\right)^t, \qquad m \times 1,$$

and

$$\nabla_{\varphi'} \equiv \left(\partial/\partial\varphi'_{ij}\right), \qquad q \times m,$$

where, in the latter, $\varphi'_{ij} \equiv \partial\varphi_i/\partial w_j$. These are distinguished from ∇ in the normal identity (3.1), except that ∇ and ∇_w will coincide whenever $\mathbf{A} = \mathbf{w}$. We assume throughout that the conditions of Theorem 2.1 are satisfied, and we shall use [L] = [R1] + [R2] to designate (2.3).

EXAMPLE 3.1 (Prototypical case). Again, let $\mathbf{X} \sim N_p(\theta, I)$. The solutions of (2.3) are given for both asymmetric and spherically symmetric priors. For the latter case, we derive (1.2) since this result motivates our main application, that is the estimation of Σ under L_1.

(i) $\Pi(\theta)$ asymmetrical. In terms of (2.1), we have $\Psi = \theta$, $\mathbf{A} = \mathbf{w} = \mathbf{X}$ and $n = m = q = p$. Now, for convenience, we write the feasible estimators as $\hat{\boldsymbol{\theta}}_\varphi = a[\mathbf{X}, \varphi(\mathbf{X})] = \mathbf{X} + \varphi(\mathbf{X})$ where φ: $R^p \to R^p$. It readily follows from (3.1) that an unbiased estimator of $\rho(\hat{\boldsymbol{\theta}}_\varphi, \theta)$ is given by

$$\hat{\boldsymbol{\rho}}\left(\hat{\boldsymbol{\theta}}_\varphi, \theta\right) = \not{\rho}[\mathbf{X}, \varphi(\mathbf{X}), d\varphi(\mathbf{X})] = p + 2\nabla^t\varphi(\mathbf{X}) + \varphi(\mathbf{X})^t\varphi(\mathbf{X}), \tag{3.3}$$

in which $d\varphi(\mathbf{X}) = \nabla\varphi(\mathbf{X})$, $p \times 1$. The terms in (2.3) are [L] = $2\varphi(\mathbf{X})$, [R1] = 0 and [R2] = $2\nabla \log f_\pi(\mathbf{X})$; thus (2.3) becomes $\varphi_\Pi(\mathbf{X}) = \nabla \log f_\Pi(\mathbf{X})$, which defines

$$\hat{\boldsymbol{\theta}}_B = \mathbf{X} + \nabla \log f_\Pi(\mathbf{X}). \tag{3.4}$$

How is (3.4) related to the more usual representation? Apart from our method, (3.4) is obtained from $E(\theta|\mathbf{X})$ if one observes that $\nabla f(x|\theta) = -(x - \theta)f(x|\theta)$ [see Stein (1981), page 1140]. Surely this can always be done; that is, the usual expression for the Bayes rule always gives the VFBE, after it is transformed by appropriate identities. [Note that the last equation is a key step in the derivation of (3.1) (with $\Sigma = I$).] Such conversions are academic, however, since they require that the VFBE be known in advance. On the other hand, (2.3) provides the appropriate generalization of (3.4) for a large class of useful models. The main point, of course, is that the VFBE can be used to find estimators with good frequency properties. While our method is endemic to problems where an unbiased estimator of the risk function is available, such problems do account for most of the published work on Stein-like estimation.

(ii) $\Pi(\theta)$ spherically symmetric. Again $\Psi = \theta$. Now $\mathbf{A} = \mathbf{X}$, $\mathbf{w} = \mathbf{X}^t\mathbf{X}$ and the feasible estimators are of the form $\hat{\boldsymbol{\theta}}_\varphi = (1 - \varphi(\mathbf{w}))\mathbf{X}$ for $\varphi(\mathbf{w})$ real; $n = p$

and $m = q = 1$. From (3.1), the unbiased estimator of the risk function is given by

$$\begin{aligned}\hat{\rho}\left(\hat{\theta}_{\varphi}, \theta\right) &= p - 2\nabla^{t}[\varphi(\mathbf{w})\mathbf{X}] + [\varphi(\mathbf{w})\mathbf{X}]^{t}[\varphi(\mathbf{w})\mathbf{X}] \\ &= p - 4\mathbf{w}\varphi'(\mathbf{w}) - 2p\varphi(\mathbf{w}) + \mathbf{w}\varphi^{2}(\mathbf{w}).\end{aligned} \tag{3.5}$$

(Recall that the definition of φ is specific to the problem.) Since $(1 - \varphi(\mathbf{w}))\mathbf{X}$ is dominated by $[1 - \min\{1, \varphi_B(\mathbf{w})\}]\,\mathbf{X}$ [Anderson (1984), page 91] we introduce the "slack variable"

$$\varepsilon(w)^{2} \equiv 1 - \varphi(w), \tag{3.6}$$

and then solve (2.3) for the extremal $\varepsilon_{\Pi}(w)^2 \equiv 1 - \varphi_{\Pi}(w)$. This determines an estimator for $(1 - \varphi_{\Pi}(\mathbf{w}))\mathbf{X}$ in which $\varphi_{\Pi}(\mathbf{w}) \leq 1$.

The solution is obtained as follows. First, from (3.5) and (3.6) we obtain

$$\begin{aligned}\hat{\rho}\left(\hat{\theta}_{\varphi}, \theta\right) &= p - 4\mathbf{w}\varphi'(\mathbf{w}) - 2p\varphi(\mathbf{w}) + \mathbf{w}\varphi^{2}(\mathbf{w}) \\ &= -p + \mathbf{w} + 8\mathbf{w}\varepsilon(\mathbf{w})\varepsilon'(\mathbf{w}) + 2(p - \mathbf{w})\varepsilon(\mathbf{w})^{2} + \mathbf{w}\varepsilon(\mathbf{w})^{4}.\end{aligned} \tag{3.7}$$

Denote the last expression by $\not{e}[\mathbf{w}, \varepsilon(\mathbf{w}), d\varepsilon(\mathbf{w})]$ in which $d\varepsilon(\mathbf{w}) = \varepsilon'(\mathbf{w})$. In (2.3) we have $\nabla_{\varepsilon} = d/d\varepsilon$ so that $[\mathrm{L}] = 8\mathbf{w}\varepsilon'(\mathbf{w}) + 4(p - \mathbf{w})\varepsilon(\mathbf{w}) + 4\mathbf{w}\varepsilon(\mathbf{w})^3$. The other operators are $\nabla_{\varepsilon'} = d/d\varepsilon'$ and $\nabla_{w} = d/dw$. Thus $[\mathrm{R1}] = 8\varepsilon(\mathbf{w}) + 8\mathbf{w}\varepsilon'(\mathbf{w})$ and $[\mathrm{R2}] = 8\mathbf{w}\varepsilon(\mathbf{w})(d/dw)\log g_{\Pi}(\mathbf{w})$. In summary, (2.3) becomes

$$\varepsilon(\mathbf{w})\left[(p - 2)/\mathbf{w} - 1 + \varepsilon(\mathbf{w})^{2} - 2(d/dw)\log g_{\Pi}(\mathbf{w})\right] = 0,$$

from which it follows that

$$\varepsilon_{\Pi}(\mathbf{w}) = 0 \quad \text{or} \quad \varepsilon_{\Pi}(\mathbf{w})^{2} = 1 - (p - 2)/\mathbf{w} + 2(d/dw)\log g_{\Pi}(\mathbf{w}).$$

Consequently, we obtain

$$\hat{\boldsymbol{\theta}} = \left[1 - (p - 2)/\mathbf{w} + 2(d/dw)\log g_{\Pi}(\mathbf{w})\right]^{+}\mathbf{X}, \tag{3.8}$$

in which $r^{+} = r$ if $r > 0$ and $r^{+} = 0$ if $r \leq 0$. (Recall that g_{Π} is any function for which the minimization is well posed.)

We shall let (3.8) and its analogues suggest naive choices for g_{Π}. In (3.8), for example, $g_{\Pi}(w) = \text{const.}$ is an important reference since it gives the positive-part James–Stein estimator. [It is easily seen that the formal Bayes risk does not exist if $g_{\Pi}(w) = \text{const.}$, so this choice extends the solution class (3.8).] This strategy is gainfully employed within the context of $(\Sigma, \hat{\Sigma}, L_1)$, a problem for which we have much less intuition.

LITERATURE. See Berger (1980) for ample references wherein generalizations of $\hat{\rho}(\hat{\theta}_{\varphi}, \theta)$ (3.5) are used to find minimax estimators of θ. Among these, one should see Stein (1974, 1981). Efron and Morris [(1976a), Berger (1976a, b, 1980) or Hudson (1978). Various theorems from these papers can be used to impose conditions on $g_{\Pi}(\mathbf{w})$ under which (3.8) is a minimax estimator. In this connection, see Efron and Morris (1976a), Theorem 2, page 16]; see also Berger [(1976b), Theorem 1, page 257].

We should mention that Brown (1971) used variational arguments of a different character than those presented in the present paper. Also, one should see Brown (1988) for a different derivation of positive-part estimators and a discussion of their properties. Other results on positive-part estimators are found in Bock (1987).

EXAMPLE 3.2 (Estimation of θ when Σ is unknown). Now consider $(\theta, \hat{\theta}, L)$ in which $\mathbf{X} \sim N_p(\theta, \Sigma)$ and $\mathbf{S} \sim W_p(\Sigma, k)$ with Σ unknown; $k - p - 1 > 0$ and $\mathbf{X}$ and $\mathbf{S}$ are independent.

(i) $\Pi(\theta, \Sigma)$ asymmetrical. In this case, we set $\mathbf{A} = \mathbf{w} = (\mathbf{X}, \mathbf{S})$ and $\Psi = (\theta, \Sigma)$ in which $\Psi_1 = \theta$ and $\Psi_2 = \Sigma$. The feasible estimators are now among functions of the form $a[\mathbf{X}, \varphi(\mathbf{W})] = \mathbf{X} + \varphi(\mathbf{W})$ with $\mathbf{W} = (\mathbf{X}, \mathbf{S})$ and $\varphi(\mathbf{W}) \in R^p$ [$n = m = (p^2 + 3p)/2$ and $q = p$]. From (3.1) and (3.2), the unbiased estimator of the risk function can be written as

$$\begin{aligned} &\not{\rho}[\mathbf{W}, \varphi(\mathbf{W}), d\varphi(\mathbf{W})] \\ &\quad = 4\varphi(\mathbf{W})^t D\varphi(\mathbf{W}) + (k - p - 1)\varphi(\mathbf{W})^t \mathbf{S}^{-1}\varphi(\mathbf{W}) \\ &\qquad + 2\nabla^t \varphi(\mathbf{W}) + p. \end{aligned} \tag{3.9}$$

Here [L] $= 4D\varphi + 2(k - p - 1)\mathbf{S}^{-1}\varphi$, [R1] $= 4D\varphi$ and

$$[\mathrm{R2}] = 2\nabla \log g_\Pi + 4[D \log g_\Pi]\varphi.$$

Thus it follows that (2.3) has the unique solution $\varphi_\Pi(\mathbf{W}) = [(k - p - 1)\mathbf{S}^{-1} - 2D \log g_\Pi]^{-1}\nabla \log g_\Pi$, and thus the appropriate generalization of (3.4) is

$$\hat{\theta}_B = \mathbf{X} + \left[(k - p - 1)\mathbf{S}^{-1} - 2D \log g_\Pi(\mathbf{W})\right]^{-1} \nabla \log g_\Pi(\mathbf{W}). \tag{3.10}$$

To verify [R1] and [R2], match $\mathbf{W} = (\mathbf{X}, \mathbf{S})$ with an element in R^n as follows:

$$\mathbf{w}_t = \begin{cases} \mathbf{X}_t & \text{for } t = 1, 2, \ldots, p, \\ \mathbf{s}_{ij} & \text{for } t = t_{ij} = (2p - i)(i - 1)/2 + j + p, \text{ provided } i \le j. \end{cases}$$

One thus obtains $\nabla_{\varphi'} \not{\rho} = (2I_{p\times p}, 4\Phi)$, $p \times n$, in which the elements of Φ are given by $\partial \not{\rho}/\partial \varphi'_{it}$ with $\varphi'_{it} \equiv \partial \varphi_i / \partial w_t$, $i = 1, \ldots, p$, $t = p + 1, \ldots, n$). Particularly,

$$\partial \not{\rho}/\partial \varphi'_{it} = \varphi_i \delta_{t, t_{ii}} + \tfrac{1}{2} \sum_{c=i+1}^{p} \varphi_c \delta_{t, t_{ic}} + \tfrac{1}{2} \sum_{c=1}^{i-1} \varphi_c \delta_{t, t_{ci}}$$

in which $\delta_{t,u}$ is Kronecker's delta on t and u. Further details are omitted.

(ii) $\Pi(\theta|\Sigma)$ elliptically symmetric. Lin and Tsai [(1973), page 143] gave a class of prior distributions $\Pi(\theta, \Sigma)$ which depend on θ only through $\theta^t \Sigma^{-1} \theta$ [see also Tiao and Zellner (1964) and Villegas (1969)]. Here the formal Bayes estimators are of the class $a[\mathbf{X}, \varphi(\mathbf{w})] = (1 - \varphi(\mathbf{w}))\mathbf{X}$, with $\mathbf{w} = \mathbf{X}^t \mathbf{S}^{-1} \mathbf{X}$ and $\varphi(\cdot)$ a real-valued function. [Now $\mathbf{A} = \mathbf{X}$ and Ψ is given as in (i).] From (3.9) it follows that the unbiased estimator of the risk function is

$$\begin{aligned} \not{\rho}[\mathbf{w}, \varphi(\mathbf{w}), d\varphi(\mathbf{w})] &= p - 2p\varphi(\mathbf{w}) - 4\mathbf{w}\varphi'(\mathbf{w}) - 4\mathbf{w}^2 \varphi(\mathbf{w})\varphi'(\mathbf{w}) \\ &\quad + (k - p - 1)\mathbf{w}\varphi^2(\mathbf{w}) \end{aligned}$$

in which $d\varphi(\mathbf{w}) \equiv \varphi'(\mathbf{w})$. It is left for the reader to show that (2.3) has the unique solution

$$\varphi_\Pi(\mathbf{w}) = \frac{(p-2) - 2\mathbf{w}(\log g_\Pi(\mathbf{w}))'}{\mathbf{w}\left[(k-p+3) + 2\mathbf{w}(\log g_\Pi(\mathbf{w}))'\right]}.$$

The formalism $g_\Pi(\mathbf{w}) = \text{const.}$ yields

$$\hat{\boldsymbol{\theta}}_S = [1 - (p-2)/\mathbf{w}(k-p+3)]\mathbf{X}, \tag{3.11}$$

the James–Stein (1961) estimator for this situation. Again, for $\mathbf{w}$ suitably defined, it appears that the most important aspect of the Bayes estimator from the frequentist point of view is featured by the VFBE.

Minimax estimators. Here we give a minimax theorem for the VFBE of Example 3.2(i). In particular, Stein's "superharmonic condition" (0.9) is extended to this setting. After some preliminaries, the result is stated as Theorem 3.1. A special case of the minimax theorem appears in Bilodeau and Kariya [(1989), Section 4].

NOTATION. For a $p \times p$ matrix V, let $V(\nabla\nabla^t)$ be that $p \times p$ operator defined by the premultiplication of $\nabla\nabla^t \equiv (\partial^2/\partial x_i\,\partial x_j)$ by V. For a scalar function h, we set $V(\nabla\nabla^t)h \equiv V(\partial^2 h/\partial x_i\,\partial x_j)$.

LEMMA 3.1. *For suitable functions* h: $R^p \to R$, *we have*

$$(\partial \log h/\partial x_i)(\partial \log h/\partial x_j) + 2\partial^2 \log h/\partial x_i\,\partial x_j \equiv \left[4\,\partial^2 h^{1/2}/\partial x_i\,\partial x_j\right]/h^{1/2}$$

or, equivalently, for any $p \times p$ *matrix* V,

$$(\nabla \log h)^t V(\nabla \log h) + 2\,\mathrm{tr}\left[V(\nabla\nabla^t)\log h\right] = 4\,\mathrm{tr}\left[V(\nabla\nabla^t)\right]h^{1/2}/h^{1/2}.$$

PROOF. The proof entails routine calculus only; we omit the details. □

In addition, we need the product rule for an operator $\tilde{D}$ whose elements are linear combinations of $\partial/\partial s_{ij}$, $i, j = 1, 2, \ldots, p$ [e.g., operators of the kind QD or $(QD)^t$ where Q is a matrix]. The result, Lemma 3.2, is used extensively in the following.

LEMMA 3.2. *Let* A *and* B *be matrix functions of* S. *Assuming that all partial derivatives and products exist as needed, we have*

$$\tilde{D}(AB) = (A^t\tilde{D}^t)^t B + (\tilde{D}A)B.$$

PROOF. Perform the formal multiplication and differentiate coordinatewise. We omit the details. □

THEOREM 3.1. *Assume that* $\mathbf{X} \sim N_p(\theta, \Sigma)$ *and* $\mathbf{S} \sim W_p(\Sigma, k)$ *with* $\mathbf{X}$ *and* $\mathbf{S}$ *independent,* θ *and* Σ *unknown,* $k - p - 1 > 0$ *and* $p \geq 3$. *For* $\mathbf{W} \equiv (\mathbf{X}, \mathbf{S})$,

let $\mathbf{H}(\mathbf{W})$ *be a* $p \times p$ *symmetric matrix, and let* $g_\Pi(\mathbf{W})$ *be a positive-valued function. Assume also that*

$$\hat{\boldsymbol{\theta}} = \mathbf{X} + \mathbf{H}(\mathbf{W})\nabla \log g_\Pi(\mathbf{W})$$

is a feasible estimator (recall Definition 2.1*). Then* $\hat{\boldsymbol{\theta}}$ *is a minimax estimator if*

$$\begin{aligned} &\varphi(\mathbf{W})^t\left[4D + (k-p-1)\mathbf{S}^{-1} - \mathbf{H}^{-1}\right]\varphi(\mathbf{W}) \\ &\quad + 4\,\mathrm{tr}\left[\mathbf{H}(\nabla\nabla^t)\right]g_\Pi(\mathbf{W})^{1/2}/g_\Pi(\mathbf{W})^{1/2} \\ &\quad + 2[\nabla^t\mathbf{H}][\nabla \log g_\Pi(\mathbf{W})] \le 0, \end{aligned} \tag{3.12}$$

where $\varphi(\mathbf{W}) \equiv \mathbf{H}(\mathbf{W})\nabla \log g_\Pi(\mathbf{W})$, $p \times 1$.

REMARK 3.1. Notice that the second term in (3.12) coincides with the Laplacian in (0.9) if we set $\mathbf{H} = I$. The first and third terms of (3.12) are due to the fact that $\mathbf{H}$ is essentially an estimator of the covariance matrix. (The latter point is clarified in the following.)

REMARK 3.2. One can show that the James–Stein estimator (3.11) satisfies inequality (3.12). [Set $\mathbf{H} = [1/(k-p+3)]\mathbf{S}$ and $g_\Pi(\mathbf{W}) \propto (\mathbf{X}^t\mathbf{S}^{-1}\mathbf{X})^{-(p-2)/2}$; hence $\varphi(\mathbf{W}) = \mathbf{H}\nabla \log g_\Pi(\mathbf{W}) = -c\mathbf{X}/\mathbf{X}^t\mathbf{S}^{-1}\mathbf{X}$ with $c = (p-2)/(k-p+3)$. Also use the fact that $D(\mathbf{X}^t\mathbf{S}^{-1}\mathbf{X}) = -\mathbf{S}^{-1}\mathbf{X}\mathbf{X}^t\mathbf{S}^{-1}$; see Haff (1982), Lemma 6(ii).]

PROOF OF THEOREM 3.1. It follows from (3.9) that

$$\begin{aligned} \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\theta}}, \theta) - \hat{\boldsymbol{\rho}}(\mathbf{X}, \theta) &= 4\varphi(\mathbf{W})^t D\varphi(\mathbf{W}) + (k-p-1)\varphi(\mathbf{W})^t\mathbf{S}^{-1}\varphi(\mathbf{W}) \\ &\quad + 2\nabla^t\varphi(\mathbf{W}). \end{aligned}$$

A routine calculation shows that

$$\begin{aligned} 2\nabla^t\varphi(\mathbf{W}) &= 2[\nabla^t\mathbf{H}][\nabla \log g_\Pi] + 2\,\mathrm{tr}\left[\mathbf{H}(\nabla\nabla^t)\right]\log g_\Pi \\ &= 2[\nabla^t\mathbf{H}][\nabla \log g_\Pi] - [\nabla \log g_\Pi]^t\mathbf{H}[\nabla \log g_\Pi] \\ &\quad + [\nabla \log g_\Pi]^t\mathbf{H}[\nabla \log g_\Pi] + 2\,\mathrm{tr}\left[\mathbf{H}(\nabla\nabla^t)\right]\log g_\Pi \\ &= 2[\nabla^t\mathbf{H}][\nabla \log g_\Pi] - \varphi(\mathbf{W})^t\mathbf{H}^{-1}\varphi(\mathbf{W}) \\ &\quad + 4\,\mathrm{tr}\left[\mathbf{H}(\nabla\nabla^t)\right]g_\Pi^{1/2}/g_\Pi^{1/2} \end{aligned}$$

(from Lemma 3.1). Thus the unbiased estimate of the difference in risk becomes

$$\begin{aligned} \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\theta}}, \theta) - \hat{\boldsymbol{\rho}}(\mathbf{X}, \theta) &= \varphi(\mathbf{W})^t\left[4D + (k-p-1)\mathbf{S}^{-1} - \mathbf{H}^{-1}\right]\varphi(\mathbf{W}) \\ &\quad + 4\,\mathrm{tr}\left[\mathbf{H}(\nabla\nabla^t)\right]g_\Pi^{1/2}/g_\Pi^{1/2} + 2[\nabla^t\mathbf{H}][\nabla \log g_\Pi] \end{aligned}$$

and the proof is complete. □

We will show that $\mathbf{H}$ is generally regarded as an estimator of Σ by relating the VFBE (3.10) to the more usual representation of the Bayes rule. The

relationship follows from these equations:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_B &= \mathbf{X} + \left[(k-p-1)\mathbf{S}^{-1} - 2D\log g_\Pi(\mathbf{W})\right]^{-1}\nabla \log g_\Pi(\mathbf{W}) \\ &= \mathbf{X} + \left[\int \Sigma^{-1} q(\Sigma|\mathbf{W})\, d\Sigma\right]^{-1} \nabla \log g_\Pi(\mathbf{W}) \\ &= \left[\int \Sigma^{-1} q(\Sigma|\mathbf{W})\, d\Sigma\right]^{-1}\left[\int\int \Sigma^{-1}\theta p(\theta,\Sigma|\mathbf{W})\, d\theta\, d\Sigma\right].\end{aligned} \tag{3.13}$$

Here $p(\theta,\Sigma|\mathbf{W})$ is the posterior p.d.f. of (θ,Σ) and $q(\Sigma|\mathbf{W}) \equiv \int p(\theta,\Sigma|\mathbf{W})\, d\theta$. From the second equation in (3.13) it is clear that $\mathbf{H}$ is an estimator of Σ. The third equation is the standard result one obtains by minimizing the posterior expected loss (actually a trivial calculation). These are related as follows. The second equation in (3.13) is obtained from the first one by using

$$-2D\log g_\Pi(\mathbf{W}) = -(k-p-1)\mathbf{S}^{-1} + \int\int \Sigma^{-1} p(\theta,\Sigma|\mathbf{W})\, d\theta\, d\Sigma, \tag{3.14}$$

the proof of which is left for the reader. The action of D in (3.14) is a key ingredient in the derivation of the Wishart identity (3.2). From the second line of (3.13), we obtain the third one by using

$$\begin{aligned}\nabla \log g_\Pi(\mathbf{W}) = &-\left[\int\int \Sigma^{-1} p(\theta,\Sigma|\mathbf{W})\, d\theta\, d\Sigma\right]\mathbf{X} \\ &+ \int\int \Sigma^{-1}\theta p(\theta,\Sigma|\mathbf{W})\, d\theta\, d\Sigma\end{aligned} \tag{3.15}$$

(the proof is again left for the reader). In its slightly disguised form, the gradient in (3.15) is used in the derivation of the normal identity (3.1).

We conclude this section with an illustration of the VFBE for some nonnormal cases. See Haff and Johnson [(1986), page 47] for the verification of certain details.

EXAMPLE 3.3 (Pearson curves). Let $\mathbf{X}$, $p \times 1$, be a vector of independent variates in which the p.d.f. of $\mathbf{X}_i$ is defined by

$$\frac{d}{dx} f(x_i|\Psi_i) = -\frac{x_i - \alpha_i}{\beta_{0i} + \beta_{1i}x_i + \beta_{2i}x_i^2} f(x_i|\Psi_i),$$

$i = 1, 2, \ldots, p$, with the mode α_i the only unknown parameter. Here Ψ_i is the mean, $\Psi_i = (\beta_{1i} + \alpha_i)/(1 - 2\beta_{2i})$ and an estimate is required of the vector of means Ψ, $p \times 1$, under the loss function (0.4). Let $\mathbf{c}_i(\mathbf{x}_i) \equiv (\beta_{0i} + \beta_{1i}\mathbf{x}_i + \beta_{2i}\mathbf{x}_i^2)/(1 - 2\beta_{2i})$, $i = 1, 2, \ldots, p$, be the ith component of $\mathbf{c}$, $p \times 1$, $\beta_{2i} \neq \frac{1}{2}$. Finally, consider the general class of estimators $\hat{\Psi} = \mathbf{X} + \varphi(\mathbf{X})$, $\varphi(\mathbf{X})$ a $p \times 1$ vector ($\mathbf{w} \equiv \mathbf{X}$). From Haff and Johnson (1986), the unbiased estimator of risk is

$$\not{r}[\mathbf{X}, \varphi(\mathbf{X}), d\varphi(\mathbf{X})] = \sum_{i=1}^{p} q_i\left[2\mathbf{c}_i(\mathbf{X}_i)\, \partial\varphi_i(\mathbf{X})/\partial x_i + \varphi_i(\mathbf{X})^2 + \mathbf{c}_i(\mathbf{X}_i)\right],$$

$$d\varphi(\mathbf{X}) \equiv \operatorname{diag}\left(\partial\varphi_1(\mathbf{X})/\partial x_1, \ldots, \partial\varphi_p(\mathbf{X})/\partial x_p\right).$$

The terms in (2.3) are [L] = $2Q\varphi(\mathbf{X})$, [R1] = $2Q(d\mathbf{c}_1/dx_1, \ldots, d\mathbf{c}_p/dx_p)^t$, $p \times 1$, and [R2] = $2Q\,\mathrm{diag}(\mathbf{c}_1(\mathbf{X}_1), \ldots, \mathbf{c}_p(\mathbf{X}_p))[\nabla \log f_n]$. It follows that the ith component of the solution of (2.3) is given by

$$\varphi_{i\Pi}(\mathbf{X}) = d\mathbf{c}_i(\mathbf{X}_i)/dx_i + \mathbf{c}_i(\mathbf{X}_i)(\partial/\partial x_i)\log f_\Pi(\mathbf{X})$$
$$= \mathbf{c}_i(\mathbf{X}_i)(\partial/\partial x_i)\log[\mathbf{c}_i(\mathbf{X}_i) f_\Pi(\mathbf{X})]$$

and hence the ith component of the VFBE by $\hat{\mathbf{\Psi}}_{iB} = \mathbf{X}_i + \varphi_{i\Pi}(\mathbf{X})$, $i = 1, 2, \ldots, p$.

4. Estimation of the covariance matrix: Unbiased estimators of risk functions. Assume that $\mathbf{S} \sim W_p(\Sigma, k)$, $k - p - 1 > 0$, with Σ to be estimated under L_1 or L_2 in (0.3). In this section, we provide the unbiased estimators of $\rho_i(\hat{\mathbf{\Sigma}}, \Sigma)$, $i = 1, 2$, up to certain matrix derivatives. These derivatives are specialized in Section 5 for orthogonally invariant estimators. With their determination, the estimation problems $(\Sigma, \hat{\mathbf{\Sigma}}, \rho_i)$, $i = 1, 2$, are treated in Sections 6 and 7, respectively.

We note first that $L_1(\hat{\mathbf{\Sigma}}, \Sigma) = \mathrm{tr}(\hat{\mathbf{\Sigma}}\Sigma^{-1}) - \log\det(\hat{\mathbf{\Sigma}}) + \log\det(\Sigma) - p$ and that the last two terms have no effect on comparisons between estimators. Thus we omit them and define $\rho_1(\hat{\mathbf{\Sigma}}, \Sigma)$ as

$$\rho_1(\hat{\mathbf{\Sigma}}, \Sigma) = E_\Sigma\big[\mathrm{tr}(\hat{\mathbf{\Sigma}}\Sigma^{-1}) - \log\det(\hat{\mathbf{\Sigma}})\big]. \tag{4.1}$$

Now expand the second loss function as $L_2(\hat{\mathbf{\Sigma}}, \Sigma) = \mathrm{tr}(\mathbf{T}\Sigma^{-1}) - 2\,\mathrm{tr}(\hat{\mathbf{\Sigma}}\Sigma^{-1}) + p$ in which $\mathbf{T} = \hat{\mathbf{\Sigma}}\Sigma^{-1}\hat{\mathbf{\Sigma}}$, so that

$$\rho_2(\hat{\mathbf{\Sigma}}, \Sigma) = E_\Sigma\big[\mathrm{tr}(\mathbf{T}\Sigma^{-1}) - 2\,\mathrm{tr}(\hat{\mathbf{\Sigma}}\Sigma^{-1}) + p\big]. \tag{4.2}$$

(It will be convenient to keep the constant p in this case.) In (4.1) and (4.2), the unbiased estimator of a general term $E\,\mathrm{tr}(\mathbf{V}\Sigma^{-1})$ is needed, where $\mathbf{V} = \mathbf{V}(\mathbf{S}, \Sigma)$, in order to provide $\hat{\boldsymbol{\rho}}_1(\hat{\mathbf{\Sigma}}, \Sigma)$ and $\hat{\boldsymbol{\rho}}_2(\hat{\mathbf{\Sigma}}, \Sigma)$. Note that the desired estimator is implicit in the Wishart identity (3.2).

The unbiased estimator $\hat{\boldsymbol{\rho}}_1(\hat{\mathbf{\Sigma}}, \Sigma)$ follows immediately from (3.2); that is, merely replace $\hat{\mathbf{\Sigma}}$ by $\mathbf{V}$ in (3.2). The result was due to Stein (1975).

THEOREM 4.1 [Stein (1975)]. *Let $\hat{\mathbf{\Sigma}}$ be an estimator of Σ for which* (3.2) *is a valid identity with $\mathbf{V} = \hat{\mathbf{\Sigma}}$. Then the unbiased estimator of $\rho_1(\hat{\mathbf{\Sigma}}, \Sigma)$ is*

$$\hat{\boldsymbol{\rho}}_1(\hat{\mathbf{\Sigma}}, \Sigma) = 2\,\mathrm{tr}(D\hat{\mathbf{\Sigma}}) + (k - p - 1)\mathrm{tr}(S^{-1}\hat{\mathbf{\Sigma}}) - \log\det(\hat{\mathbf{\Sigma}}). \tag{4.3}$$

PROOF. An immediate application of (3.2). □

Now we provide $\hat{\boldsymbol{\rho}}_2(\hat{\mathbf{\Sigma}}, \Sigma)$, the unbiased estimator of (4.2).

THEOREM 4.2. *Let $\hat{\mathbf{\Sigma}}$ be an estimator of Σ for which* (3.2) *is a valid identity when $\mathbf{V}$ is replaced by each of $\hat{\mathbf{\Sigma}}$, $\hat{\mathbf{\Sigma}}\Sigma^{-1}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Sigma}}\mathbf{S}^{-1}\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Sigma}}D\hat{\mathbf{\Sigma}}$. Then the*

unbiased estimator $\hat{\rho}_2(\hat{\Sigma}, \Sigma)$ *is given by*

$$\begin{aligned}\hat{\rho}_2(\hat{\Sigma}, \Sigma) &= 8\,\mathrm{tr}(\hat{\Sigma}D^2\hat{\Sigma}) + 8\,\mathrm{tr}(D\hat{\Sigma})^2 + 8(k-p-1)\,\mathrm{tr}(\mathbf{S}^{-1}\hat{\Sigma}D\hat{\Sigma}) \\ &\quad -(k-p-1)\left[\mathrm{tr}(\mathbf{S}^{-1}\hat{\Sigma})\right]^2 + (k-p-1)(k-p-2)\mathrm{tr}(\mathbf{S}^{-1}\hat{\Sigma})^2 \\ &\quad - 4\,\mathrm{tr}(D\hat{\Sigma}) - 2(k-p-1)\mathrm{tr}(\mathbf{S}^{-1}\hat{\Sigma}) + p.\end{aligned} \tag{4.4}$$

PROOF. In (4.2), the term $\mathrm{tr}(\mathbf{T}\Sigma^{-1})$ is quadratic in the coordinates of $\Sigma^{-1} = (\sigma^{ij})$. Hence two applications of (3.2) are required. Here we need to differentiate matrix products, and Lemma 3.2 is needed for this. Also, we need the fact that $\mathrm{tr}[(AD)^tC] = \mathrm{tr}(ADC^t)$, which is routine to verify. Further details are omitted since they are tedious and add nothing to the presentation. □

The reader who is unwilling to sort through the proof might wish to check (4.4) for a special case. For example, set $\hat{\Sigma} = (1/k)\mathbf{S}$. It is elementary that $\rho_2(\mathbf{S}/k, \Sigma) = p(p+1)/k$. This value is readily obtained from (4.4) by using the fact that $D\mathbf{S} = (p+1)I/2$, $p \times p$.

5. Orthogonally invariant estimators: Calculus on the eigenstructure. Recall that (4.3) requires the derivative $D\hat{\Sigma}$ and (4.4) requires both $D\hat{\Sigma}$ and $D^2\hat{\Sigma}$. We assume henceforth that $\Pi(\Sigma)$ is orthogonally invariant. The formal Bayes estimators under both L_1 and L_2 are thus orthogonally invariant and are given by

$$\hat{\Sigma} = \mathbf{R}\varphi(\mathbf{L})\mathbf{R}^t, \tag{5.1}$$

in which $\mathbf{L} = \mathrm{diag}(\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p)$, $\mathbf{l}_1 \geq \mathbf{l}_2 \geq \cdots \geq \mathbf{l}_p$. $\mathbf{R}\mathbf{R}^t = \mathbf{R}^t\mathbf{R} = I$ and $\varphi(\mathbf{L}) = \mathrm{diag}(\varphi_1(\mathbf{L}), \varphi_2(\mathbf{L}), \ldots, \varphi_p(\mathbf{L}))$, with $\varphi_i(\mathbf{L}) \geq 0$, $i = 1, 2, \ldots, p$. For this class of estimators, we give a formula for $D\hat{\Sigma}$ from which $D^{(n)}\hat{\Sigma}$ can be obtained by iteration.

Set $\mathbf{R} = (\mathbf{r}_{ij})$. The following derivatives appear in Stein's (1977a) notes:

$$\begin{aligned} d_{jm}\mathbf{l}_i &= \mathbf{r}_{ji}\mathbf{r}_{mi} \quad \text{and} \\ d_{mn}\mathbf{r}_{ij} &= (1/2)\sum_{a \neq j}\left[\mathbf{r}_{ia}/(\mathbf{l}_j - \mathbf{l}_a)\right]\left[\mathbf{r}_{ma}\mathbf{r}_{nj} + \mathbf{r}_{na}\mathbf{r}_{mj}\right]. \end{aligned} \tag{5.2}$$

In a more general setting, they appear in Wilkinson (1965). As a complement of Lemma 3.2, these equations are now given (without proof) in matrix form, that is, in terms of the action of $D = (d_{ij})$, $p \times p$. Theorem 5.1 then provides a useful formula for $D^{(n)}[\mathbf{R}_\varphi(\mathbf{L})\mathbf{R}^t]$.

LEMMA 5.1. *Let* $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p$ *be the eigenvalues and* $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_p$ *the corresponding eigenvectors of* $\mathbf{S}$, *a* $p \times p$ *positive definite matrix. Also, let* D *be the* $p \times p$ *operator defined in* (3.2). *Then we have*

(a) $$D\mathbf{l}_i = \mathbf{R}_i\mathbf{R}_i^t,\ p \times p, \qquad i = 1, 2, \ldots, p,$$

and

(b) $$D\mathbf{R}_i = \mathbf{l}_i^*\mathbf{R}_i \quad \text{where } \mathbf{l}_i^* = \tfrac{1}{2}\sum_{a\neq i} 1/(\mathbf{l}_i - \mathbf{l}_a).$$

The following is the main result of this section.

THEOREM 5.1. *Let* $\varphi(\mathbf{L}) = \operatorname{diag}(\varphi_1(\mathbf{L}), \varphi_2(\mathbf{L}), \ldots, \varphi_p(\mathbf{L}))$ *in which* $\varphi(\mathbf{L})$ *is differentiable on* $\{\mathbf{l}_1 \geq \mathbf{l}_2 \geq \cdots \geq \mathbf{l}_p\}$. *Then*

$$D\big[\mathbf{R}\varphi(\mathbf{L})\mathbf{R}^t\big] = \mathbf{R}\varphi^{(1)}(\mathbf{L})\mathbf{R}^t,$$

in which $\varphi^{(1)}(\mathbf{L}) = \operatorname{diag}(\varphi_1^{(1)}(\mathbf{L}), \varphi_2^{(1)}(\mathbf{L}), \ldots, \varphi_p^{(1)}(\mathbf{L}))$,

$$\varphi_i^{(1)}(\mathbf{L}) = \tfrac{1}{2}\sum_{a\neq i}\big[\varphi_i(\mathbf{L}) - \varphi_a(\mathbf{L})\big]/(\mathbf{l}_i - \mathbf{l}_a) + \partial\varphi_i(\mathbf{L})/\partial l_i, \tag{5.3}$$
$$i = 1, 2, \ldots, p.$$

REMARK 5.1. From Theorem 5.1, it is clear that $D^{(n)}\hat{\mathbf{\Sigma}}$ can be obtained by recursion provided that $\varphi(\mathbf{L})$ is smooth enough.

PROOF OF THEOREM 5.1. From Lemma 3.2 we obtain

$$D\mathbf{R}\varphi\mathbf{R}^t = \big(\varphi\mathbf{R}^t D\big)^t\mathbf{R}^t + \Big[\big(\mathbf{R}^t D\big)^t\varphi\Big]\mathbf{R}^t + (D\mathbf{R})\varphi\mathbf{R}^t.$$

Denote these successive terms by A, B and C, respectively. First, we show that $A = \mathbf{R}N\mathbf{R}^t$ in which $N = \operatorname{diag}(n_1, n_2, \ldots, n_p)$, $n_i = (1/2)\sum_{b\neq i}\varphi_b/(\mathbf{l}_b - \mathbf{l}_i)$. The matrix

$$A = \Big(\sum_b \varphi_b \sum_a \mathbf{r}_{ab}\, d_{ai}\mathbf{r}_{jb}\Big) = \tfrac{1}{2}\Big(\sum_b \sum_{a\neq b} \mathbf{r}_{ja}\mathbf{r}_{ia}\varphi_b/(\mathbf{l}_b - \mathbf{l}_a)\Big)$$
$$= \tfrac{1}{2}\sum_b \sum_{a\neq b} \mathbf{R}_a\mathbf{R}_a^t\varphi_b/(\mathbf{l}_b - \mathbf{l}_a) = \sum_b \mathbf{R}N^{(b)}\mathbf{R}^t - I,$$

in which $N^{(b)}$ is a diagonal matrix $N_i^{(b)} = 1$ if $i = b$ and $N_i^{(b)} = (1/2)\varphi_b/(\mathbf{l}_b - \mathbf{l}_i)$ if $i \neq b$. The claim for A is established by noting that $\sum_b N_i^{(b)} - 1 = (1/2)\sum_{b\neq i}\varphi_b/(\mathbf{l}_b - \mathbf{l}_i)$. Next we show that

$$B = \mathbf{R}\operatorname{diag}(\partial\varphi_1/\partial l_i, \partial\varphi_2/\partial l_2, \ldots, \partial\varphi_p/\partial l_p)\mathbf{R}^t.$$

The derivative is $(\mathbf{R}^t D)^t\varphi = (\sum_a \mathbf{r}_{aj}\, d_{ai}\varphi_{jj})$. Further,

$$\sum_a \mathbf{r}_{aj}\, d_{ai}\varphi_j = \sum_a \mathbf{r}_{aj}\sum_c (\partial\varphi_j/\partial l_c)\, d_{ai} l_c = \sum_a \mathbf{r}_{aj}\sum_c (\partial\varphi_j/\partial l_c)\mathbf{r}_{ac}\mathbf{r}_{ic}$$
$$= \sum_c (\partial\varphi_j/\partial l_c)\Big(\sum_a \mathbf{r}_{aj}\mathbf{r}_{ac}\Big)\mathbf{r}_{ic} = (\partial\varphi_j/\partial l_j)\mathbf{r}_{ij}.$$

Hence, $(\mathbf{R}^t D)^t\varphi = [(\partial\varphi_j/\partial l_j)\mathbf{r}_{ij}]$. The claim for B now follows from $[(\mathbf{R}^t D)^t\varphi]\mathbf{R}^t = (\sum_a(\partial\phi_a/\partial l_a)\mathbf{r}_{ia}\mathbf{r}_{ja}) = \sum_a(\partial\varphi_a/\partial l_a)\mathbf{R}_a\mathbf{R}_a^t$. Finally, we show that $C = \mathbf{R}M\mathbf{R}^t$, $M = \operatorname{diag}(m_1, m_2, \ldots, m_p)$, where $m_j = \varphi_j l_j^*$ (recall Lemma 5.1). This is immediate since $(D\mathbf{R})\varphi\mathbf{R}^t = (\mathbf{l}_1^*\mathbf{R}_1, \mathbf{l}_2^*\mathbf{R}_2, \ldots, \mathbf{l}_p^*\mathbf{R}_p)\varphi\mathbf{R}^t = \mathbf{R}[\operatorname{diag}(\mathbf{l}_1^*, \mathbf{l}_2^*, \ldots, \mathbf{l}_p^*)]\varphi\mathbf{R}^t = \mathbf{R}M\mathbf{R}^t$. In summary, it is readily seen that $A + B + C = \mathbf{R}\varphi^{(1)}(\mathbf{L})\mathbf{R}^t$ and the proof is complete. □

6. Results for loss function L_1.

Preliminary remarks. Denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ the ordered eigenvalues of Σ and recall that an orthogonally invariant estimator of Σ is obtained by specifying $\varphi_i(\mathbf{L})$, an estimator of λ_i, $i = 1, 2, \ldots, p$. In this section, an estimator of Stein (1975) is compared with an estimator suggested by the VFBE under L_1 (both are orthogonally invariant). The two estimators are closely related, but they impose the natural constraint $[\varphi_1(\mathbf{L}) \geq \varphi_2(\mathbf{L}) \geq \cdots \geq \varphi_p(\mathbf{L}) \geq 0]$ in different ways.

Stein's (1975) method of computing $\varphi_i(\mathbf{L})$, $i = 1, 2, \ldots, p$, subject to $\varphi_1(\mathbf{L}) \geq \varphi_2(\mathbf{L}) \geq \cdots \geq \varphi_p(\mathbf{L}) \geq 0$ has a certain drawback; namely, his corrections of initial (disordered) estimates are not driven by the risk function. (A brief summary is given below.) By comparison, our own method proceeds as follows: Since $\int \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\Sigma}}, \Sigma) f_\Pi(S)\, dS = \int \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\Sigma}}, \Sigma) g_\Pi(l)\, dl$ in which $\mathbf{l} = (\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p)^t$ and $g_\Pi(l)$ is the marginal p.d.f. of $\mathbf{l}$, our estimates of $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are constructed from the solution of

$$\begin{aligned} &\text{minimize } \int \hat{\rho}(\hat{\Sigma}, \Sigma) g_\Pi(l)\, dl \\ &\text{subject to } \varphi_1(\mathbf{L}) \geq \varphi_2(\mathbf{L}) \geq \cdots \geq \varphi_p(\mathbf{L}) \geq 0 \end{aligned} \tag{6.1}$$

The prior $\Pi(\Sigma)$ is orthogonally invariant, but not specified otherwise. Consequently, g_Π is not determined by a particular prior distribution. We will start with the solution class indexed by the set of all g_Π for which (6.1) is well posed. Then we will examine a naive choice for g_Π which is suggested by Stein's (1975) work.

In the unconstrained solution, the substitution $g_\Pi(l) = \det(S^{-1})$ yields eigenvalue estimates identical to Stein's unconstrained estimates. This choice for g_Π is used throughout since it enables us to compare the two estimators. It is not clear, however, that this is optimal in any way. In particular, it is not clear that $\det(S^{-1})$ is the marginal distribution for any improper prior distribution. The problem of choosing g_Π is virtually an open one.

On the method of solution: Similar to (3.6), we introduce slack variables $\varepsilon_1(\mathbf{L}), \varepsilon_2(\mathbf{L}), \ldots, \varepsilon_p(\mathbf{L})$ defined by

$$\varphi_1 - \varphi_2 \equiv \varepsilon_1^2, \qquad \varphi_2 - \varphi_3 \equiv \varepsilon_2^2, \qquad \ldots, \qquad \varphi_p \equiv \varepsilon_p^2 \tag{6.2}$$

and we give a finite algorithm for solving (6.1) in terms of the ε's. The solutions are close analogues of the positive-part solutions (3.8).

Finally, why compare the VFBE with Stein's (1975) estimator? One standard for judging the frequency performance of new estimators is the minimax estimator of James and Stein (1961). It has constant risk and it dominates $(1/k)\mathbf{S}$, the unbiased estimator. In recent years, however, Stein's (1975) estimator has often been the standard, and there are two reasons why this is so. First, the constant risk minimax estimator never beats $(1/k)\mathbf{S}$ by very much [see Stein (1977a, b) for comparisons]. Second, the Monte Carlo results of Lin and Perlman (1985) indicate that Stein's estimator beats $(1/k)\mathbf{S}$ and

various other competitors by a substantial amount—over a significant portion of the parameter space, anyway. (Stein's estimator does best when $\lambda_i = c$, $i = 1, 2, \ldots, p$, and its performance decreases as the λ's become more dispersive.) We should add that Lin and Perlman (1985) devised other estimators which perform well in particular situations, as did Dey and Srinivasan (1985). See Haff (1979a, b, 1980) for related work. Overall, it appears that Stein's estimator and the VFBE outperform any of the others when the eigenvalues are close, with the VFBE doing best.

Stein's (1975) *estimator: A brief sketch.* The specialization of (4.3) for orthogonally invariant estimators follows from Theorem 5.1 and a simple identity,

$$\sum_b \sum_{b \neq i} (\varphi_i - \varphi_b)/(\mathbf{l}_i - \mathbf{l}_b) = 2\sum_i \sum_{b>i} (\varphi_i - \varphi_b)/(\mathbf{l}_i - \mathbf{l}_b)$$
$$= 2\sum_i \sum_{b \neq i} \varphi_i/(\mathbf{l}_i - \mathbf{l}_b).$$

For the latter simply combine the (i, b) and (b, i) terms. Thus the unbiased estimator of the risk function is given by

$$\hat{\boldsymbol{\rho}}_1(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = 2\sum_i \sum_{b \neq i} \varphi_i(\mathbf{L})/(\mathbf{l}_i - \mathbf{l}_b) + 2\sum_i \partial\varphi_i(\mathbf{L})/\partial l_i$$
$$+ (k - p - 1)\sum_i \varphi_i(\mathbf{L})/\mathbf{l}_i - \sum_i \log \varphi_i(\mathbf{L}), \tag{6.3}$$

which appeared first in Stein's (1975) Rietz lecture.

Stein (1975) ignored the partial derivatives in (6.3) and minimized the resulting approximation of $\hat{\boldsymbol{\rho}}_1$. This approach yields rough eigenvalue estimates, namely,

$$\varphi_a^s = \mathbf{l}_a \Big/ \left[(k - p + 1) + 2\mathbf{l}_a \sum_{i \neq a} 1/(\mathbf{l}_a - \mathbf{l}_i)\right], \qquad a = 1, 2, \ldots, p. \tag{6.4}$$

Note that the constraint in (6.1) might be violated by these; that is, they can be out of order and any but the first ($a = 1$) can be negative. For the denominators in (6.4), let us set

$$\alpha_a(\mathbf{L}) = (k - p + 1) + 2\mathbf{l}_a \sum_{i \neq a} 1/(\mathbf{l}_a - \mathbf{l}_1), \qquad a = 1, 2, \ldots, p. \tag{6.5}$$

Stein corrected the estimates (6.4) by using (a) an algorithm which pooled adjacent pairs (l_a, α_a) to ensure positive eigenvalue estimates, followed by (b) an isotonic regression on these positive estimates. See Lin and Perlman (1985) for a detailed description of steps (a) and (b). [Also see Barlow et al. (1972) for the standard work on isotonic regression.] In the following, a return from steps (a) and (b) is simply called Stein's correction, and we denote these corrected terms by $\bar{\varphi}_j^s$, $i = 1, 2, \ldots, p$. Accordingly, Stein's (1975) estimator of the

covariance matrix is given by

$$\hat{\mathbf{\Sigma}}_S = \mathbf{R}\bar{\varphi}^S(\mathbf{L})\mathbf{R}^t, \quad \text{where } \bar{\varphi}^S = \operatorname{diag}\left(\bar{\varphi}_j^S, \ldots, \bar{\varphi}_p^S\right). \tag{6.6}$$

The solution of (6.1). In (6.2), note that the φ_i are recovered by

$$\varphi_i(\mathbf{L}) = \sum_{t \geq i} \varepsilon_t(\mathbf{L})^2, \qquad i = 1, 2, \ldots, p. \tag{6.7}$$

Now rewrite (2.3) in terms of the slack variables; viz., in the notation of (2.3) (where $\hat{\rho}_1 = \not{\rho}$) we have

$$\begin{aligned} \not{\rho}[\mathbf{l}, \varepsilon(\mathbf{l}), d\varepsilon(\mathbf{l})] &= 2\sum_i \sum_{b \neq i} \sum_{t \geq i} \varepsilon_t^2/(\mathbf{l}_i - \mathbf{l}_b) + 4\sum_i \sum_{t \geq i} \varepsilon_t \, \partial\varepsilon_t/\partial l_i \\ &\quad + (k - p - 1)\sum_i (1/\mathbf{l}_i)\sum_{t \geq i} \varepsilon_t^2 - \sum_i \log\left[\sum_{t \geq i} \varepsilon_t^2\right], \end{aligned} \tag{6.8}$$

in which $\mathbf{w} = \mathbf{l} = (\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p)^t$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)^t$ and $d\varepsilon = (\partial\varepsilon_i/\partial l_j)$, $p \times p$. The terms in (2.3) are thus given by

$$\begin{aligned} [\mathrm{L}] = \left(\sum_{a=1}^{i} \left\{ 4\varepsilon_i \sum_{b \neq a} 1/(l_a - l_b) + 4\,\partial\varepsilon_i/\partial l_a \right.\right. \\ \left.\left. + 2(k - p - 1)\varepsilon_i(1/\mathbf{l}_a) - 2\varepsilon_i\left[\sum_{t \geq a} \varepsilon_t^2\right]^{-1} \right\}\right), \end{aligned}$$

a $p \times 1$ vector,

$$[\mathrm{R1}] = \left(\sum_{a=1}^{i} 4\,\partial\varepsilon_i/\partial l_a \right), \qquad p \times 1,$$

(notice that [R1] cancels with the second term in [L]) and

$$[\mathrm{R2}] = \left(4\varepsilon_i \sum_{a=1}^{i} \partial \log g_{\Pi}(L)/\partial l_a \right), \qquad p \times 1.$$

Finally, the ith equation of the system [L] = [R1] + [R2] is given by

$$\begin{aligned} \varepsilon_i \sum_{a=1}^{i} \left[2\sum_{b \neq a} 1/(\mathbf{l}_a - \mathbf{l}_b) + (k - p - 1)(1/\mathbf{l}_a) \right. \\ \left. - 1/\varphi_a - 2\partial \log g_{\Pi}/\partial l_a \right] = 0, \end{aligned} \tag{6.9}$$

$i = 1, 2, \ldots, p$. This is a triangular system of equations whose solution is readily described.

For the remainder of this section we interpret the solution of (6.9) in terms of φ_a^s, $a = 1, 2, \ldots, p$ [recall (6.4)]. Accordingly, we henceforth set

$$g_{\Pi}(l) \propto \det(S^{-1}), \tag{6.10}$$

so that $\partial \log g_{\Pi}(l)/\partial l_a = -1/l_a$. From the latter, the system (6.9) becomes

$$\varepsilon_i \sum_{a=1}^{i} \left[(\varphi_a^s)^{-1} - (\varphi_a)^{-1} \right] = 0, \qquad i = 1, 2, \ldots, p$$

or, schematically,

$$\begin{aligned} 1/\varphi_1^s &= 1/\varphi_1 && \text{or} \quad \varepsilon_1 = 0, \\ 1/\varphi_1^s + 1/\varphi_2^s &= 1/\varphi_1 + 1/\varphi_2 && \text{or} \quad \varepsilon_2 = 0, \\ &\vdots \\ \sum_{a=1}^{p} 1/\varphi_a^s &= \sum_{a=1}^{p} 1/\varphi_a && \text{or} \quad \varepsilon_p = 0. \end{aligned} \tag{6.11}$$

[Recall that $\varphi_i(\mathbf{L}) = \sum_{t \geq i} \varepsilon_t(\mathbf{L})^2$, $i = 1, 2, \ldots, p$.] This system is solved by a finite algorithm, and a FORTRAN routine is available from the author upon request. Let us denote the solution of (6.11) by $\varphi_1^H \geq \varphi_2^H \geq \cdots \geq \varphi_p^H$ and hence the VFBE by

$$\hat{\Sigma}_H = \mathbf{R}\varphi^H(\mathbf{L})\mathbf{R}^t \quad \text{where } \varphi^H = \operatorname{diag}\left(\varphi_1^H, \ldots, \varphi_p^H\right). \tag{6.12}$$

A qualitative comparison. Let us informally compare the eigenvalue estimates in (6.6) with those in (6.12). Among $\overline{\varphi}_1^s, \overline{\varphi}_2^s, \ldots, \overline{\varphi}_p^s$ (Stein's correction) we find pooled estimates of the kind

$$\overline{\varphi}_j^s = \overline{\varphi}_{j+1}^s = \overline{\varphi}_{j+2}^s = (\mathbf{l}_j + \mathbf{l}_{j+1} + \mathbf{l}_{j+2})/(\alpha_j + \alpha_{j+1} + \alpha_{j+2}), \tag{6.13}$$

for example [recall (6.4)], where in general the actual number of consecutive rough estimates that are pooled depends on the pattern of violations. [Again, see Lin and Perlman (1985) for details.] By comparison, $\varphi_1^H, \varphi_2^H, \ldots, \varphi_p^H$ typically include harmonic averages of consecutive rough estimates. Thus, the VFBE analogue of (6.13) is given by

$$\varphi_j^H = \varphi_{j+1}^H = \varphi_{j+2}^H = 3\Big/\left[(1/\varphi_j^s) + (1/\varphi_{j+1}^s) + (1/\varphi_{j+2}^s)\right]. \tag{6.14}$$

If the required order is not violated by the rough estimates, then it is seen that $\varphi_j^H = \overline{\varphi}_j^s = \varphi_j^s$, $i = 1, 2, \ldots, p$.

On correcting the eigenvalue distortion in the sample covariance matrix. From Jensen's inequality, it follows that $E[(1/k)\mathbf{l}_1] \geq \lambda_1$ and $E[(1/k)\mathbf{l}_p] \leq \lambda_p$. Furthermore, this distortion is greatest whenever $\Sigma = \sigma^2 I$. It can be seen that the VFBE corrects this distortion in natural way. In particular, let $\varphi_1^H, \ldots, \varphi_p^H$ be the eigenvalue estimates which define the VFBE with $g_{\Pi}(l) \propto \det(S^{-1})$. Then it can be seen that

(i) $$(1/k)\mathbf{l}_p \leq \varphi_p^H \quad \text{a.e.}$$

and

(ii) $$[(1/k)\mathbf{l}_1] \geq \varphi_1^H,$$

where $\varphi_1^H = h/\sum_{i=1}^{h}(1/\varphi_1^S)$ and h satisfies $p \geq 2h - 1$. The proof of this statement is routine, so the details have been omitted.

The Monte Carlo results. We close this section with some Monte Carlo comparisons between $\hat{\Sigma}_S$ and $\hat{\Sigma}_H$. First, note that the best scalar multiples of $\mathbf{S}$ for L_1 and L_2 are given by $\mathbf{S}/k$ and $\mathbf{S}/(k + p + 1)$, respectively [see Haff (1980)]. In each case, we denote the "best scalar multiple" by $\hat{\Sigma}_0$, since its value will be obvious from the context. The simulated risk functions of $\hat{\Sigma}_S$ and $\hat{\Sigma}_H$ are compared relative to that of $\hat{\Sigma}_0$.

While $\hat{\Sigma}_S$ and $\hat{\Sigma}_H$ were obtained under L_1, we can still compare them under L_2 in a meaningful way. That is, suppose that $\hat{\Sigma}$ is derived under L_1. Then $\hat{\Sigma}$ can be examined under L_2 if it is rescaled as $k\hat{\Sigma}/(k + p + 1)$. [Lin and Perlman (1985) thus compared $k\hat{\Sigma}_S/(k + p + 1)$ with various other estimators under L_2]. For our purposes, we are compelled to examine this rescaling since the VFBE's under L_2 are fairly complicated.

The simulation was done as follows. For each $p \in \{5, 10, 15, 20\}$ and $k = 2p$, we generated 100 i.i.d. matrices from a $W_p(I, k)$ distribution. These observations were then transformed into $W_p(\Sigma, k)$ matrices for each of

$$\begin{aligned} \Sigma_a &= \operatorname{diag}(1, 1, \ldots, 1), \\ \Sigma_b &= \operatorname{diag}(2p, 1, 1, \ldots, 1), \\ \Sigma_c &= \operatorname{diag}(p, p-1, p-2, \ldots, 1). \end{aligned} \tag{6.15}$$

(We can assume Σ is diagonal since the problem is orthogonally invariant.) At each combination of p, L_i, $i = 1, 2$, and Σ [in (6.15)], we computed $\hat{\Sigma}_0^{(j)}$, $\hat{\Sigma}_S^{(j)}$ and $\hat{\Sigma}_H^{(j)}$, $j = 1, 2, \ldots, 100$. Then we proceeded as follows.

(a) The 100 paired differences (p.diff.'s) $L(\hat{\Sigma}_S^{(j)}, \Sigma) - L(\hat{\Sigma}_H^{(j)}, \Sigma)$, $j = 1, 2, \ldots, 100$, were used to compute

$$\text{mean p.diff.} \pm 2 \text{ standard deviations (of mean p.diff.)},$$

an approximate 95% confidence interval for $R(\hat{\Sigma}_S, \Sigma) - R(\hat{\Sigma}_H, \Sigma)$. At each combination of p, L_i, $i = 1, 2$, and Σ, this interval appears as the top entry in Table 1.

(b) We recorded $\bar{L}$, the sample mean loss for $\hat{\Sigma}_S$, $\hat{\Sigma}_H$ and $\hat{\Sigma}_0$ (at each combination). Also, following Lin and Perlman (1985), we recorded the percentage reduction in the average loss (PRIAL) relative to $\hat{\Sigma}_0$; that is, we recorded

$$\text{PRIAL} = 100\frac{\bar{L}(\hat{\Sigma}_0, \Sigma) - \bar{L}(\hat{\Sigma}, \Sigma)}{\bar{L}(\hat{\Sigma}_0, \Sigma)} \tag{6.16}$$

for both $\hat{\Sigma}_S$ and $\hat{\Sigma}_H$. In Table 1, the bottom entry (at each combination) records the ordered pair

$$\text{(PRIAL:Stein's estimator, PRIAL:VFBE)}. \tag{6.17}$$

As indicated above, $\hat{\Sigma}_0$ depends on the loss function being used and, accordingly, $\hat{\Sigma}_S$ and $\hat{\Sigma}_H$ are rescaled by $k/(k + p + 1)$ for comparisons under L_2.

TABLE 1
Mean p.diff. ± 2 *standard deviations (of mean* p.diff.) *and* (PRIAL:*Stein's estimator*, PRIAL:VFBE)

	$p = 5$	$p = 10$	$p = 15$	$p = 20$
		L_1		
Σ_a	0.034 ± 0.030 (63%, 65%)	0.060 ± 0.029 (79%, 81%)	0.056 ± 0.025 (85%, 87%)	0.087 ± 0.026 (89%, 90%)
Σ_b	0.004 ± 0.017* (49%, 49%)	0.040 ± 0.020 (70%, 72%)	0.058 ± 0.023 (80%, 81%)	0.097 ± 0.020 (84%, 85%)
Σ_c	−0.012 ± 0.020* (45%, 45%)	−0.012 ± 0.020* (48%, 48%)	−0.002 ± 0.014* (47%, 47%)	−0.005 ± 0.011* (49%, 49%)
		L_2		
Σ_a	0.180 ± 0.028 (30%, 39%)	0.354 ± 0.032 (41%, 51%)	0.479 ± 0.040 (47%, 56%)	0.615 ± 0.040 (49%, 58%)
Σ_b	0.099 ± 0.020 (23%, 29%)	0.269 ± 0.020 (38%, 45%)	0.437 ± 0.034 (43%, 52%)	0.538 ± 0.020 (46%, 53%)
Σ_c	0.036 ± 0.020 (24%, 26%)	0.044 ± 0.020 (22%, 23%)	0.054 ± 0.020 (21%, 22%)	0.076 ± 0.014 (22%, 23%)

*The observed difference in risk between Stein's estimator and the VFBE was not significant at the 95% level. The other differences were increasingly significant with increasing values of p.

The simulation was done for increasing values of p, since we conjectured that the VFBE should do increasingly better than Stein's estimator as p increases. (This was based upon the fact that our constrained optimization is more criterion dependent than Stein's correction, so we should be doing better.) For the given levels of p and $k = 2p$, our conjecture was confirmed, but the (L_1) differences in PRIAL were small. In particular, for L_1, the VFBE outperformed Stein's estimator at Σ_a and Σ_b by only 1–2% in PRIAL terms. The differences for Σ_c were not significant at all.

For loss function L_2, however, the differences in performance were more pronounced. At Σ_a and Σ_b, the VFBE was best (in PRIAL terms) by some 4–10%; at Σ_c, it was best by 1–3%.

We also performed simulations for various k at each fixed p. For $p < k < 2p$, our results were more striking than those reported above (for $p = 2k$). To cite one example, for $p = 20$, $k = 21$ and $\Sigma = \Sigma_a$, we recorded [$\bar{L} \pm 2$ standard deviations ($\bar{L}$), PRIAL] at both loss functions for both Stein's estimator and the VFBE. The results were as follows:

$$
\begin{aligned}
&L_1 \quad \text{Stein's estimator:} && 4.08 \pm 0.48, \quad 78\%,\\
&\qquad\qquad \text{VFBE:} && 3.60 \pm 0.48, \quad 80\%,\\
&L_2 \quad \text{Stein's estimator:} && 7.47 \pm 0.18, \quad 25\%,\\
&\qquad\qquad \text{VFBE:} && 5.54 \pm 0.38, \quad 45\%.
\end{aligned}
\tag{6.18}
$$

While the 2% improvement in PRIAL under L_1 was typical of all of our results for that loss function, the 20% improvement under L_2 was a bit surprising.

One parting remark about these simulations: There are important applications in which $p > 100$, say, for which efficiency of estimation is a vital concern. The performance of the VFBE relative to Stein's estimator in high dimensions (> 20) has not been determined. Future simulations on this scale would be of practical interest.

7. Results for loss function L_2. A brief summary. The formal Bayes estimators of Σ under L_2 are again of the form $\hat{\Sigma} = \mathbf{R}\varphi(\mathbf{L})\mathbf{R}^t$. In this case, however, the risk calculations are fairly complicated, so the unconstrained version of the VFBE is stated without proof. Nonetheless, we record a corollary of Sharma (1983) which provides an independent confirmation of the result.

From Theorems 4.2 and 5.1, it can be seen that $\hat{\rho}_2(\hat{\Sigma}, \Sigma)$ can be expressed as a function of the φ_i, $i = 1, 2, \ldots, p$, and certain first and second partial derivatives of the φ_i. Consequently, we can minimize

$$\rho_2(\hat{\Sigma}, \Pi) = \int_{R^m} \hat{\rho}_2(\hat{\Sigma}, \Sigma) g_\Pi(L)\, dL \tag{7.1}$$

(and thus obtain the VFBE) by extending Theorem 2.1 in a natural way. (Details are available upon request from the author.)

Denote by $\mathscr{V}$ the set of $p \times p$ positive definite matrices and by $\mathscr{O}(p)$ the orthogonal group of $p \times p$ matrices. Then we have

$$g_\Pi(l) \propto \left[\prod_{i=1}^{p} l_i\right]^{(k-p-1)/2} \left[\prod_{i<j} (l_i - l_j)\right] g^*(L), \tag{7.2}$$

in which

$$g^*(L) = \int_{\mathscr{O}(p)} \int_{\mathscr{V}} |\Sigma|^{-k/2} e\,\mathrm{tr}(-\Sigma^{-1} H L H^t/2)\, d\Pi(\Sigma)\, dH.$$

[See Muirhead (1982) or Anderson (1984).]

The VFBE is given by the following.

THEOREM 7.1. *Assume that* $\Pi(\Sigma)$ *is orthogonally invariant and that* (7.1) *is a valid expression for the formal Bayes risk. If the VFBE exists, then it is given by* $\mathbf{R}\varphi^B(\mathbf{L})\mathbf{R}^t$ *in which the diagonal elements of* φ^B *are given by the solution of a linear system*

$$A\Phi = C$$

in which $\Phi \equiv (\varphi_1, \varphi_1, \ldots, \varphi_1)^t$, $p \times 1$, $A = (a_{ij})$, $p \times p$, *with*

$$a_{ij} = \begin{cases} 2\dfrac{\partial^2}{\partial l_i^2} \log g^*(L) + 2\left[\dfrac{\partial}{\partial l_i} \log g^*(L)\right]^2, & i = j, \\ \left(\dfrac{1}{l_i - l_j}\right)\left[\dfrac{\partial}{\partial l_i} \log g^*(L) - \dfrac{\partial}{\partial l_j} \log g^*(L)\right], & i \neq j, \end{cases}$$

and $C = (c_i)$, $p \times 1$, $c_i = -\partial/\partial l_i \log g^*(L)$, $i = 1, 2, \ldots, p$.

PROOF. Omitted. □

The proof of Theorem 7.1 is very tedious. However, the statement is supported by an important special case. Under the conjugate prior distribution

$$\Sigma^{-1} \sim W_p[(1/\gamma)I, k'] \quad \text{with } k' \text{ and } \gamma \text{ specified.} \tag{7.3}$$

It is known that the Bayes estimator is given by

$$\hat{\Sigma}_B = (\mathbf{S} + \gamma I)/(k + k' + p + 1). \tag{7.4}$$

The estimator specified by Theorem 7.1 must be identical to that in (7.3), of course, a fact actually confirmed by Sharma (1983). Sharma's observation is recorded as the following corollary.

COROLLARY 7.1. *Let* $\varphi^B(L)$, $p \times 1$, *be the solution of* $A\Phi = C$, *with the latter specified by Theorem* 7.1. *Further, let* $d\Pi(\Sigma^{-1})$ *be given by* (7.3) *with* k' *and* γ *both specified. Then*

$$\hat{\Sigma}_B = \mathbf{R}\varphi^B(\mathbf{L})\mathbf{R}^t = (\mathbf{S} + \gamma I)/(k + k' + p + 1).$$

PROOF. The result readily follows from $g^*(L) = [\prod_{i=1}^p (l_i + \gamma)]^{-(k+k')/2}$. Further details are omitted. □

Acknowledgments. The author is grateful to Carl FitzGerald, who suggested some preliminary calculations that led to this study. He is grateful, also, to several editors and two patient referees for many helpful suggestions. Finally, he is indebted to J. O. Berger for conversations which helped further improve the manuscript.

REFERENCES

ALCARAZ, J. E. (1990). The minimax estimation of Poisson means with the variational form of a restricted Bayes estimator. Unpublished manuscript.

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

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M., and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

BERGER, J. (1976a). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.* **4** 223–226.

BERGER, J. (1976b). Minimax estimation of a multivariate normal mean under arbitrary quadratic loss. *J. Multivariate Anal.* **6** 256–264.

BERGER, J. (1980). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameter. *Ann. Statist.* **8** 545–571.

BERGER, J. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) **1** 109–141. Academic, New York.

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERGER, J. and HAFF, L. R. (1981). A class of minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Statist. Decisions* **1** 105–129.

BILODEAU, M. and KARIYA, T. (1989). Minimax estimators in the normal MANOVA model. *J. Multivariate Anal.* **28** 260–270.

BOCK, M. E. (1987). Shrinkage estimators: Pseudo-Bayes rules for normal mean vectors. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 281–297. Springer, New York.

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

BROWN, L. D. (1988). The differential inequality of a statistical estimation problem. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 299–324. Springer, New York.

CHEN, SHUN-YU (1988). Restricted risk Bayes estimation for the mean of multivariate normal distribution. *J. Multivariate Anal.* **24** 207–217.

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

EATON, M. L. (1970). Some problems in covariance estimation (preliminary report). Technical Report 49, Dept. Statistics, Stanford Univ.

EFRON, B. and MORRIS, C. (1976a). Families of minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* **4** 11–21.

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

GELFAND, I. M. and FOMIN, S. V. (1963). *Calculus of Variations* (revised English edition) (Richard A. Silverman, transl. and ed.). Prentice-Hall, Englewood Cliffs, N.J.

GEORGE, E. I. (1986). Minimax multiple shrinkage estimation. *Ann. Statist.* **14** 188–205.

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 convariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *Ann. Statist.* **7** 1264–1276.

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

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

HAFF, L. R. (1981). Further identities for the Wishart distribution with applications in regression. *Canad. J. Statist.* **9** 215–224.

HAFF, L. R. (1982). Identities for the inverse Wishart distribution with computational results in linear and quadratic discrimination. *Sankhyā Ser. B* **44** 245–258.

HAFF, L. R. and JOHNSON, R. W. (1986). The superharmonic condition for simultaneous estimation of means in exponential families. *Canad. J. Statist.* **14** 43–54.

HUDSON, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6** 478–484.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 361–380. Univ. California Press, Berkeley.

JOHNSON, R. W. (1984). Simultaneous estimation of generalized Pearson means. Ph.D dissertation, Dept. Mathematics, Univ. of California, San Diego.

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

LIN, P.-E. and TSAI H.-L. (1973). Generalised Bayes minimax estimators of the multivariate normal mean with unknown covariance matrix. *Ann. Statist.* **1** 142–145.

LOH, W.-L. (1988a). Estimating covariance matrices I. Technical Report 88-37, Dept. Statistics, Purdue Univ.

LOH, W.-L. (1988b). Estimating the common mean of two multivariate normal distributions. Technical Report 88-48. Dept. Statistics, Purdue Univ.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

MUIRHEAD, R. J. and VERATHAWORN, F. (1983). On estimating the latent roots of $\Sigma_1\Sigma_2^{-1}$. In *Multivariate Analysis IV* (P. R. Krishnaiah, ed.) 431–447. North-Holland, Amsterdam.

OLKIN, I. and SELLIAH, J. B. (1977). Estimating covariance in a multivariate normal distribution. In *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. S. Moore, eds.) 313–326. Academic, New York..

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

SELLIAH, J. B. (1964). Estimation and testing problems in a Wishart distribution. Technical Report 10, Dept. Statistics, Stanford Univ.

SHARMA, D. (1983). Personal communication.

SHARMA, D. and KRISHNAMOORTHY (1984). Empirical Bayes estimators of normal covariance matrix. Unpublished manuscript.

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

STEIN, C. (1974). Estimation of the mean of a multivariate normal distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics* (J. Hájek, ed.). Univ. Karlova, Prague.

STEIN, C. (1975). Estimation of a covariance matrix. Rietz Lecture, 1975 Annual Meeting of American Statistical Association, Atlanta, Ga.

STEIN, C. (1977a). Personal communication. 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* (I. A. Ibragimov and M. S. Nikulin, eds.). *Proceedings of Scientific Seminars of the Steklov Institute* **74** 4–65. Leningrad Division, Leningrad.

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

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

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

TAKEMURA, A. (1983). An orthogonally univariant minimax estimator of the covariance matrix of a multivariate normal distribution. Technical Report 9, Dept. Statistics, Stanford Univ.

TIAO, G. C. and ZELLNER, A. (1964). On the Bayesian estimation of multivariate regression. *J. Roy. Statist. Soc. Ser. B* **26** 277–285.

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

WILKINSON, J. H. (1965). *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

DEPARTMENT OF MATHEMATICS C-012
UNIVERSITY OF CALIFORNIA
LA JOLLA, CALIFORNIA 92093