The Annals of Statistics
1980, Vol. 8, No. 4, 716–761

A ROBUST GENERALIZED BAYES ESTIMATOR AND CONFIDENCE REGION FOR A MULTIVARIATE NORMAL MEAN[1]

BY JAMES BERGER

Purdue University

It is observed that in selecting an alternative to the usual maximum likelihood estimator, δ^0, of a multivariate normal mean, it is important to take into account prior information. Prior information in the form of a prior mean and a prior covariance matrix is considered. A generalized Bayes estimator is developed which is significantly better than δ^0 if this prior information is correct and yet is very robust with respect to misspecification of the prior information. An associated confidence region is also constructed, and is shown to have very attractive size and probability of coverage.

TABLE OF CONTENTS

1. Introduction. Let $X = (X_1, \cdots, X_p)^t$ have a p-variate distribution with mean vector $\theta = (\theta_1, \cdots, \theta_p)^t$ and nonsingular covariance matrix Σ. (Σ will be assumed known until Section 5.) It is desired to estimate θ using an estimator $\delta(X) = (\delta_1(X), \cdots, \delta_p(X))^t$ and under a quadratic loss $L(\theta, \delta) = (\delta - \theta)^t Q(\delta - \theta)$, Q being a positive definite $(p \times p)$ matrix. Two common problems giving rise to this setup are (i) estimating a multivariate mean where X is the vector of sample

Received September 1978; revised February 1979.

[1]Research supported by the National Science Foundation, Grants MCS76-06627 and MCS76-06627A2, and by the John Simon Guggenheim Memorial Foundation.

AMS 1970 *subject classifications.* Primary 62F15; secondary 62F10, 62F25.

Key words and phrases. Robust generalized Bayes estimators, multivariate normal mean, quadratic loss, risk, confidence ellipsoids, size, probability of coverage.

means, and (ii) estimating a vector, θ, of regression coefficients where $X = (D'D)^{-1}D'Y$ is the least squares estimator and $\Sigma = \sigma^2(D'D)^{-1}$, D being the design matrix and σ^2 the variance of the errors in the observation Y.

The usual estimator $\delta^0(X) = X$ has been observed to have several deficiencies. These include:

1. It is inadmissible if $p \geqslant 3$. Indeed an estimator δ' can be found with $R(\theta, \delta') < R(\theta, \delta^0) = \operatorname{tr} Q\Sigma$ for all θ, where $R(\theta, \delta) = E_\theta L(\theta, \delta(X))$ is the expected loss. This was first noticed by Stein (1955).

2. It does not use often existing prior information or relationships among the coordinates, such as when the θ_i are a sample from a superpopulation.

3. When X is the least squares estimator from a regression problem, δ^0 is unstable in that $(D'D)$ is often nearly singular, so that small changes in the observation Y result in very large changes in the estimates of the regression coefficients. (This problem has given rise to the theory of ridge regression, introduced by Hoerl and Kennard (1970).)

In attempting to improve upon δ^0, a number of different approaches have been taken. For the most part, these can be categorized into three areas, according to the nature of the resulting estimator.

The first category consists of approaches resulting in estimators which are linear (i.e., of the form $\delta(X) = CX + \mu$, C a matrix and μ a vector). For example, the Bayesian approach with normal priors and the original form of ridge regression (with a fixed ridge constant) give rise to linear estimators.

The second category of approaches consists of those for which coordinates of θ are set equal to zero, the remainder being estimated in a standard way. For example, preliminary test estimators and typical regression procedures which select the "significant" regression coefficients (effectively setting the others equal to zero) are of this type.

The third category consists of approaches leading to estimators which satisfy

$$\delta(x) = (I - T/(x'Cx))x + o(|x|^{-1}) \quad \text{as} \quad |x| \to \infty, \tag{1.1}$$

where T and C are $(p \times p)$ matrices, $|x|$ is the Euclidean norm of x, and "o" is the usual little oh notation. For example, minimax, empirical Bayes, Bayes with t-like priors, and the usual stochastic ridge regression approaches all result in estimators of the form (1.1). (In stochastic ridge regression, the ridge constant is usually estimated from the data using the inverse of some quadratic form in X.)

A number of articles dealing with the above approaches are listed in the references. Unfortunately, the number of articles is by now too large to allow discussion of each contribution specifically, and even too large to reasonably include all in the references. Therefore, only the latest articles and articles specifically referred to in this paper are listed. References to earlier works can be found in these articles.

In looking for an alternative to δ^0, only the third category of estimators will be considered. Linear estimators have the well-known disadvantage of a lack of

robustness with respect to the assumptions under which they are derived. For example, if a Bayesian approach with a normal prior were taken, the resulting linear estimator would have infinite Bayes risk if the true prior were Cauchy. (By "true prior" is meant that prior distribution which the person would choose if an infinite amount of time were available for introspection and comparisons among alternate possibilities. In any real situation the actual prior distribution chosen will necessarily be only an approximation to this true prior distribution.) In contrast, estimators of the form (1.1) tend to be considerably more robust. Some evidence of this will be presented later (see also Rubin (1977)).

Estimators from the second category will not be considered for two reasons. First, if, indeed, estimation is the sole goal, then it has generally been found that discontinuous procedures (such as preliminary test estimators) can be improved upon by smooth shrinkage procedures satisfying (1.1). Of course, there are often compelling reasons (in regression, for example) to try for model simplification by setting "nonsignificant" coordinates equal to zero. The goal then is not simply estimation, however, and it seems simplest to approach the problem in two stages —decide first which coordinates are to be set equal to zero, and then use a good estimation procedure on the remaining coordinates. The first stage is outside of the scope of this paper, while for the second stage using a smooth estimator is desirable.

In choosing among estimators of the form (1.1), one is presented with a wide array of principles to go by. The key in choosing among these principles lies in observing the behavior of the estimators—namely, that the estimators perform well (have risk significantly better than δ^0) only in specific regions of the parameter space R^p. Outside these regions they have risks which are either essentially equivalent to, or possibly worse than, δ^0. (This is basically due to the fact that δ^0 is minimax, so that no uniformly large improvement in risk is possible.) Since the region of significant improvement differs from estimator to estimator, it seems inescapable that choosing an estimator can best be done by choosing the region of θ over which improvement in risk is desired. In other words, prior knowledge must come into play in effectively choosing an estimator of the form (1.1). (As we shall see, this prior knowledge can be quite vague, such as merely believing that the prior distribution of the θ_i is exchangeable. See Section 4 for a discussion of this.)

Note that the above reasoning is not the usual rationality argument for being Bayesian, but, instead, a seemingly inevitable conclusion of the particular problem being considered. Indeed, if it is felt that there is no prior information whatsoever available, then δ^0 might as well be used, since the "chance" that θ would happen to be in the region of significant improvement of a competing estimator would be negligible. In the remainder of the paper comments will often be phrased in Bayesian terms, not necessarily because a prior distribution on θ is thought to exist, but because it seems necessary to act as if one does exist if a good alternative to δ^0 is to be chosen.

The above considerations also point out the difficulty in meaningfully comparing estimators of the form (1.1) by numerical studies. In numerical studies, the θ at

which the estimators are evaluated must be chosen in some fashion, and different estimators will perform best depending on how the θ are chosen. This point was raised by Efron and Morris, Bingham and Larntz, and Thisted in the discussion of Dempster, Schatzoff, and Wermuth (1977).

It would be enormously difficult to specify prior information for a particular problem, and then choose among all available estimators of the form (1.1) according to which does best for that particular set of prior beliefs. Instead, an estimator should be developed which allows the direct incorporation of prior information in order to adjust its region of significant improvement. This and other desirable properties of an estimator are listed below.

1. δ should readily allow incorporation of prior information.

2. δ should be robust with respect to misspecification of prior information. Equivalently, δ should not have risk $R(\theta, \delta)$ seriously worse than $R(\theta, \delta^0) = \text{tr}(Q\Sigma)$ over a significant region of the parameter space.

3. δ should be expressible in a closed form, relatively simple formula, not only for ease of calculation but also to enable examination for unintuitive or unappealing features.

4. δ should be stable in a ridge sense (providing this is consistent with 1).

5. δ should be admissible (or nearly so).

6. δ should have the following "empirical Bayes" property. Assume $\Sigma = \sigma^2 I$ and that the θ_i are a random sample from a prior distribution with mean 0 and variance τ^2. Then $\lim_{p\to\infty} |X|^2/p = \sigma^2 + \tau^2$ with probability one. The estimator

$$\delta(X) = (1 - p\sigma^2/|X|^2)X \tag{1.2}$$

is thus very close to the optimal linear Bayes estimator $\delta^L(X) = (1 - \sigma^2/(\sigma^2 + \tau^2))X$, while having a risk uniformly better than δ^0—a very desirable situation. (See Efron and Morris (1973a) for further discussion.)

7. δ should have good associated confidence regions for θ.

The rationale for property 1 has been discussed. Property 2 is also crucial, in that while it is necessary to make use of prior information to significantly improve upon δ^0, we do not want to run the risk of being significantly worse than δ^0 if the (often vague) prior information is wrong. Bayesians might disagree with defining robustness with respect to misspecification of the prior information in terms of the relationship of $R(\theta, \delta)$ to $R(\theta, \delta^0) = \text{tr}(Q\Sigma)$, as done here. To a Bayesian, a more relevant concept would be robustness of the posterior distribution or of the posterior expected loss. A thorough discussion of this issue would take us too far afield. (See Berger (1980) for such a discussion.) We content ourselves here with the observation that when X is "extreme" (by which we mean not plausible according to prior beliefs) it is reasonable to suggest that a "robust posterior Bayesian" would doubt his prior beliefs, and want $\delta(X)$ to be close to $\delta^0(X) = X$ (the rule that corresponds to no prior information). (See Hill (1974) for discussion of this.) Since extreme X correspond to large θ, a comparison of $R(\theta, \delta)$ and $R(\theta, \delta^0)$ for large θ is often an appropriate way of investigating this relationship.

Property 3 seems important, partly to make the estimator more attractive to practitioners, but also to make a thorough analysis of the estimator possible. Properties 4, 5, and 6 are all appealing, but perhaps will not be compelling to some statisticians, depending on their philosophical viewpoint. Property 7 is of considerable importance in typical applications of estimation. Section 3 will be devoted to the development and analysis of an interesting set of confidence regions.

In attempting to verify the above properties for a proposed estimator, numbers 2, 5, and 7 cause the most difficulty. In checking 2, Berger (1976b) can be useful, though numerical studies are probably necessary. The only certain method of ensuring that 5 is satisfied is to develop δ as an admissible generalized Bayes estimator. (Brown (1971) shows that an estimator must be generalized Bayes to be admissible.) Trying to verify that an estimator is "nearly" admissible is difficult. A useful negative result is given in Berger and Srinivasan (1978), namely that estimators satisfying (1.1) are approximations to generalized Bayes estimators (up to a $o(|x|^{-1})$ term) if and only if $T = k\Sigma C$ for some constant k.

Estimators so far proposed do not fully satisfy the above list of properties. The only estimators of the form (1.1) which allow the incorporation of prior information are empirical Bayes estimators (see, for example, Efron and Morris (1973a) and Rolph (1976)) and Bayes estimators arising from flat-tailed prior densities (see, for example, Box and Tiao (1968), certain examples in Lindley and Smith (1972), Hill (1974), Dickey (1974) and Leonard (1976)). Unfortunately the estimators which have been developed using these approaches cannot be written in closed form (except for a few special cases of Q, Σ, and prior information), making a meaningful analysis of them very difficult. In Section 2 a reasonable generalized Bayes estimator is developed which does satisfy the above seven properties.

Before proceeding, a word is in order as to what type of prior input is envisaged. Recall that the real goal is to decide what region of the parameter space is of greatest importance. A relatively simple approach would be to specify an ellipsoid of interest. This ellipsoid could be written as $\{\theta : (\theta - \mu)^t A^{-1}(\theta - \mu) \leq p\}$. Alternatively it seems plausible to assume the availability of a prior mean vector, μ, for θ, and also of a variance (or covariance) matrix A which reflects the believed accuracy of the guess, μ. In either case the prior input is conveniently summarized by μ and A. Only rarely will additional prior knowledge (such as knowledge of the functional form of the prior) be available. Hence it is desired to construct an estimator which can make use of μ and A, but which requires no further knowledge of the prior in order to be better than δ^0. The estimator should also be robust in the sense that if μ and A do not reflect the true value of θ (or the true prior of θ for Bayesians), the estimator should not be significantly worse than δ^0. Further discussion of the prior input is given in Section 4, where it is shown how to incorporate into the above framework such things as a belief in exchangeability of the prior, or a belief that certain linear restrictions on θ hold.

It should be emphasized at the outset that the only "real" prior beliefs are those in μ and A. The procedures we recommend will be developed with respect to a

particular (generalized) prior distribution which allows incorporation of μ and A. This prior should in a sense be considered a technical artifact, however, and not as the believed prior distribution. The prior was chosen because it does seem to accurately reflect μ and A, because it is flat-tailed (which leads to robust procedures), and because it leads to explicit procedures which can be easily evaluated. It is reasonable to think of the prior as a very conservative (and hence not necessarily optimal) representation of the prior beliefs in μ and A.

The development and analysis in the following sections is rather lengthy. For convenience, therefore, we present here the actual estimator and confidence region that are suggested for use (when $p \geqslant 3$). Define, for convenience,

$$\|X - \mu\|^2 = [1 - 2/p](X - \mu)'(\Sigma + A)^{-1}(X - \mu).$$

The suggested estimator is

$$\delta^*(X) = X - \frac{(p-2)[1 - h_{n^*}(\|X - \mu\|^2)]}{(X - \mu)'(\Sigma + A)^{-1}(X - \mu)}\Sigma(\Sigma + A)^{-1}(X - \mu),$$

where $n^* = (p - 2)/2$ and h_{n^*} is given by (2.7). The suggested confidence region is

$$C^*(X) = \{\theta \in R^p : [\theta - \delta^*(X)]'\Sigma^*(X)^{-1}[\theta - \delta^*(X)] \leqslant k(\alpha)\},$$

where $k(\alpha)$ is the $100(1 - \alpha)$th percentile of the chi-square distribution with p degrees of freedom, and (letting $v = \|X - \mu\|^2$)

$$\Sigma^*(X) = \Sigma - \frac{(p-2)[1 - h_{n^*}(v)]}{(X - \mu)'(\Sigma + A)^{-1}(X - \mu)}\Sigma(\Sigma + A)^{-1}\Sigma$$
$$+ \frac{p(p-2)[1 - h_{n^*}(v)][1 - h_{(n^*+1)}(v)]}{[(X - \mu)'(\Sigma + A)^{-1}(X - \mu)]^2}$$
$$\times \Sigma(\Sigma + A)^{-1}(X - \mu)(X - \mu)'(\Sigma + A)^{-1}\Sigma.$$

For the situation $\Sigma = \sigma^2\Sigma_0$, where Σ_0 is known but σ^2 is unknown, assume a random variable S^2 (independent of X) is observable, and that S^2/σ^2 has a chi-square distribution with m degrees of freedom. Then δ^* and C^* should be used with Σ replaced by

$$\hat{\Sigma} = [S^2/(m + 2)]\Sigma_0$$

and

$$k(\alpha) = (1 + 2/m)pF_{p,m}(1 - \alpha),$$

where $F_{p,m}(1 - \alpha)$ is the $100(1 - \alpha)$th percentile of the F distribution with p and m degrees of freedom. (This is discussed in Section 5.)

2. The generalized Bayes estimator. In this section (and Sections 3 and 5) it will be assumed for convenience that $\mu = 0$. This can be effected by a simple translation of the problem, and so can be assumed without loss of generality.

The notation $\det(B)$, $\mathrm{tr}(B)$, and $\mathrm{ch}_{\max}(B)$ will be used to denote the determinant, trace, and maximum characteristic root of a matrix B. Also, E will be used to denote expectation, with subscripts denoting parameter values and superscripts denoting random variables with respect to which the expectation is to be taken. When obvious, subscripts and superscripts will be deleted.

2.1. *Development of the estimator.* Let C be a $(p \times p)$ symmetric matrix such that $(C - \Sigma)$ is positive semidefinite. Define $B(\lambda) = \lambda^{-1}C - \Sigma$, for $0 < \lambda < 1$. For $n > 0$, consider the generalized prior density

$$g_n(\theta) = \int_0^1 [\det\{B(\lambda)\}]^{-\frac{1}{2}} \exp\{-\theta' B(\lambda)^{-1}\theta/2\} \lambda^{(n-1-p/2)}\, d\lambda.$$

Note that the conditional density of θ given λ is normal with mean 0 and covariance matrix $B(\lambda)$, while λ has the (generalized) density $(2\pi)^{p/2}\lambda^{(n-1-p/2)}$ on $(0, 1)$. This prior is a generalization of one considered in Berger (1976a), and for $\Sigma = C = I$ was first introduced by Strawderman (1971). (Judge and Bock (1977) give a good discussion of these special cases.)

Several aspects of g_n are interesting to observe. First, it can be shown that asymptotically (for large $|\theta|$) g_n behaves like $k(\theta' C^{-1}\theta)^{-n}$ for some constant k. Thus larger n correspond to "sharper tails" for the prior. It can also be checked that g_n has finite mass for $n > p/2$.

For certain C, n, and p, $g_n(\theta)$ can be calculated explicitly. For example, if $C = c\Sigma (c \geqslant 1)$, $p = 4$, and $n = (p - 2)/2 = 1$, then

$$g_n(\theta) = k\big(1 - \exp\{-\theta'\Sigma^{-1}\theta/[2(c-1)]\}\big)/\theta'\Sigma^{-1}\theta.$$

The actual form of g_n is not of great importance, however, since we do not really think that g_n is the true prior. Indeed, in one respect g_n is rather unnatural, in that it depends on Σ, the covariance matrix of X. (Thus, in a sampling situation, the prior changes as the sample size increases.) Intuitively this is unappealing. Prior information should, by definition, be independent of the experiment. Recall, however, that the goal is to obtain robust Bayesian procedures. Robustness can only be evaluated with regard to the particular experimental setup, so it is not necessarily unreasonable for the prior to depend on the experiment. For example, as the sample size increases and the sample information becomes more accurate, robustness becomes easier to obtain and one might be willing to be more daring in the choice of a prior (or, more specifically, in the quantification of true prior beliefs).

The above discussion is all rather speculative, however, and our basic attitude is that the proof is in the pudding. There are many situations in which reasonable priors give unreasonable results and questionable priors give very good results. Thus a necessary part of any truly convincing Bayesian analysis is a thorough examination of the resulting procedures, an analysis with respect to classical as well as Bayesian criteria. Hopefully the extensive analysis in the remainder of the paper will convince the reader that, whatever its faults, the prior g_n works.

We begin with the calculation of δ^n, the generalized Bayes estimator of θ with respect to g_n. For those not interested in the calculation, the result is given in expressions (2.4), (2.6) and (2.7).

Since the loss is quadratic, δ^n is simply the mean of the posterior distribution of θ given x. Hence

$$\delta^n(X) = \frac{\int \theta \exp\{-(X-\theta)'\Sigma^{-1}(X-\theta)/2\}\, g_n(\theta)\, d\theta}{\int \exp\{-(X-\theta)'\Sigma^{-1}(X-\theta)/2\}\, g_n(\theta)\, d\theta}.$$

It is straightforward to check that $g_n(\theta)$ has finite mass over any compact neighborhood of zero. This, along with the fact that $g_n(\theta)$ is bounded outside a neighborhood of zero, allows interchanging the order of integration in the numerator above to get

$$\begin{aligned}\int \theta \exp\{-(X-\theta)'\Sigma^{-1}(X-\theta)/2\}\, g_n(\theta)\, d\theta \\ = \int_0^1 \int \theta \exp\{-[(X-\theta)'\Sigma^{-1}(X-\theta) + \theta' B(\lambda)^{-1}\theta]/2\} d\theta \\ \times [\det\{B(\lambda)\}]^{-\frac{1}{2}} \lambda^{(n-1-p/2)} d\lambda.\end{aligned}$$

Completing squares and integrating out over θ in the last expression results in the equivalent formula

$$\begin{aligned}\int_0^1 \left[(\Sigma^{-1} + B(\lambda)^{-1})^{-1}\Sigma^{-1}X\right] \exp\left\{-X'\left[\Sigma^{-1} - \Sigma^{-1}(\Sigma^{-1} + B(\lambda)^{-1})^{-1}\Sigma^{-1}\right]X/2\right\} \\ \times \left[\det(\Sigma^{-1} + B(\lambda)^{-1})\right]^{-\frac{1}{2}} \left[\det B(\lambda)\right]^{-\frac{1}{2}} \lambda^{(n-1-p/2)}\, d\lambda.\end{aligned}$$

Using the matrix identities

$$\begin{aligned}[\Sigma^{-1} + B(\lambda)^{-1}]B(\lambda) &= \Sigma^{-1}B(\lambda) + I = \Sigma^{-1}C/\lambda, \\ [\Sigma^{-1} + B(\lambda)^{-1}]^{-1} &= \Sigma - \Sigma[\Sigma + B(\lambda)]^{-1}\Sigma \\ &= \Sigma - \Sigma(C/\lambda)^{-1}\Sigma = \Sigma - \lambda\Sigma C^{-1}\Sigma,\end{aligned} \tag{2.1}$$

it can be concluded that

$$\begin{aligned}\int \theta \exp\{-(X-\theta)'\Sigma^{-1}(X-\theta)/2\}\, g_n(\theta)\, d\theta \\ = \int_0^1 (I - \lambda\Sigma C^{-1})X \exp\{-\lambda X'C^{-1}X/2\}[\det(\Sigma^{-1}C)]^{-\frac{1}{2}} \lambda^{n-1}\, d\lambda.\end{aligned}$$

A similar calculation verifies that

$$\begin{aligned}\int_0^1 \exp\{-(X-\theta)'\Sigma^{-1}(X-\theta)/2\}\, g_n(\theta)\, d\theta \\ = \int_0^1 \exp\{-\lambda X'C^{-1}X/2\}[\det(\Sigma^{-1}C)]^{-\frac{1}{2}} \lambda^{n-1}\, d\lambda.\end{aligned} \tag{2.2}$$

Hence, defining

$$\|X\|^2 = X'C^{-1}X$$

and

$$r_n(v) = \frac{v\int_0^1 \lambda^n \exp\{-\lambda v/2\}\, d\lambda}{\int_0^1 \lambda^{(n-1)} \exp\{-\lambda v/2\}\, d\lambda}, \tag{2.3}$$

it follows that

$$\delta^n(X) = \left(I - \frac{r_n(\|X\|^2)\Sigma C^{-1}}{\|X\|^2}\right)X. \tag{2.4}$$

This calculation could have heuristically been done more quickly by first doing the analysis for θ conditional on λ and X, and then integrating out over the formal posterior density (on (0, 1)) of λ given X, namely

$$\exp\{-\lambda v/2\}\lambda^{(n-1)}/\int_0^1 \exp\{-\lambda v/2\}\lambda^{(n-1)}\,d\lambda, \tag{2.5}$$

where $v = \|X\|^2$. From (2.3) it is clear that the mean of this posterior density is $r_n(v)/v$. Such an analysis would not be rigorous, however, due to the fact that $g_n(\theta)$ is not a proper density.

An integration by parts in the numerator of (2.3) èstablishes that

$$r_n(v) = 2n\left(1 - \left[n\int_0^1 \lambda^{n-1}\exp\{-(\lambda - 1)v/2\}\,d\lambda\right]^{-1}\right). \tag{2.6}$$

Integration by parts also shows that

(2.7)

$$h_n(v) = \left[n\int_0^1 \lambda^{n-1}\exp\{-(\lambda - 1)v/2\}\,d\lambda\right]^{-1} = \left[\sum_{i=0}^{\infty}\frac{\Gamma(n+1)(v/2)^i}{\Gamma(n+1+i)}\right]^{-1}$$

$$= \frac{(v/2)^n}{n!\left[\exp\{v/2\} - \sum_{i=0}^{n-1}(v/2)^i/i!\right]} \quad \text{if } n \text{ is an integer}$$

$$= \frac{(v/2)^n}{\Gamma(n+1)\left[\exp\{v/2\}\operatorname{erf}\{(v/2)^{\frac{1}{2}}\} - \sum_{i=0}^{(n-3/2)}(v/2)^{(i+1/2)}/\Gamma\left(i+\tfrac{3}{2}\right)\right]}$$

$$\text{if } n - \tfrac{1}{2} \text{ is an integer,}$$

where $\operatorname{erf}((v/2)^{\frac{1}{2}}) = (2/\pi)^{\frac{1}{2}}\int_0^{v^{\frac{1}{2}}}\exp\{-t^2/2\}\,dt$. The last expressions in (2.7) are particularly convenient for calculation. The following lemma gives several properties of r_n which will be needed in the evaluation of δ^n.

LEMMA 2.1.1. *If $n > 0$ and $v > 0$, then*

(i) $0 < r_n(v) < 2n$;
(ii) $r_n(v)$ *is increasing in* v;
(iii) $\lim_{v\to\infty} r_n(v) = 2n$;
(iv) $\lim_{v\to 0}[r_n(v)/\{nv/(n+1)\}] = 1$;
(v) $\lim_{n\to\infty} r_n(v) = v$;
(vi) $\lim_{n\to\infty}[r_n(2nc)/(2n\{\min(1, c)\})] = 1$;
(vii) $r_n(v)/v$ *is decreasing in* v;
(viii) $\lim_{v\to\infty}[r_n'(v)/\{\exp(-v/2)(v/2)^n/\Gamma(n)\}] = 1$, *where* $r_n'(v) = (d/dv)\,r_n(v)$;
(ix) $\lim_{v\to\infty} v^m r_n'(v) = 0$ *for all* $m > 0$;
(x) $r_n(v)/v \leqslant n/(n+1)$;
(xi) $\lim_{v\to\infty} v^m(r_n(v) - 2n) = 0$ *for all* $m > 0$.

PROOF. Parts (i), (ii) and (iii) follow immediately from (2.6) and the first expression in (2.7). Parts (iv) and (v) follow from the first expression in (2.7), after noticing that the first two terms of the summation are dominant as $v \to 0$ or $n \to \infty$.

To prove part (vi), the first expression in (2.7) will again be used. For fixed i, the ith term of the summation satisfies

$$\lim_{n\to\infty}\left[(nc)^i\Gamma(n+1)/\Gamma(n+1+i)\right] = c^i.$$

Hence

$$\lim_{n\to\infty}\left[\sum_{i=0}^{\infty}\frac{(nc)^i\Gamma(n+1)}{\Gamma(n+1+i)}\right]^{-1} = \left[\sum_{i=0}^{\infty}c^i\right]^{-1} = \begin{cases} 0 & \text{if } c \geqslant 1 \\ (1-c) & \text{if } 0 < c < 1 \end{cases}.$$

The result follows.

To prove part (vii), consider λ as a random variable with the density (2.5). It is easy to check that this density has decreasing monotone likelihood ratio in v, and hence that the expected value of λ must be decreasing in v. But since the expected value of λ is simply $r_n(v)/v$, the conclusion follows.

To prove part (viii), observe that a calculation using (2.6) gives

$$r_n'(v) = \frac{\exp\{-v/2\}\int_0^1\lambda^{(n-1)}(1-\lambda)\exp\{-\lambda v/2\}\,d\lambda}{\left[\int_0^1\lambda^{(n-1)}\exp\{-\lambda v/2\}\,d\lambda\right]^2} \tag{2.8}$$

But

$$\begin{aligned}\int_0^1\lambda^m\exp\{-\lambda v/2\}\,d\lambda &= (v/2)^{-(m+1)}\int_0^{v/2}\lambda^m\exp\{-\lambda\}\,d\lambda \\ &= (v/2)^{-(m+1)}(\Gamma(m+1) - o(v^{-1})).\end{aligned}$$

Using this in (2.8) gives the desired result.

Part (ix) follows immediately from part (viii). Part (x) follows from parts (iv) and (vii). Part (xi) follows from (2.6) and the first expression in (2.7). □

The first question which arises is how should n and C be chosen? The choice of n that is recommended is $n = (p-2)/2$. The estimator δ^n can then be easily calculated using (2.4) and (2.7), and the resulting estimator will be seen to have many nice properties. Note that by Lemma 2.1.1 (iii), $\lim_{v\to\infty} r_n(v) = 2n = (p-2)$ for this choice of n. When $C = \Sigma = I$, the estimator $\delta^{(p-2)/2}(X)$ is thus similar to the original Stein estimator $\delta(X) = (1 - [p-2]/|X|^2)X$. Further justification for this choice of n will be seen later.

As a guide to choosing C, consider the situation where the prior is known to have mean zero and covariance matrix A. If absolutely certain about this, one would probably be fairly happy to use the best linear estimator (in terms of Bayes risk.) This best linear estimator is easily calculated to be

$$\delta(X) = \left(I - \Sigma(\Sigma + A)^{-1}\right)X.$$

This suggests choosing $C = \rho(\Sigma + A)$, ρ a constant, for then

$$\delta^n(X) = \left(I - \frac{r_n(\|X\|^2)\Sigma(\Sigma + A)^{-1}}{X'(\Sigma + A)^{-1}X}\right)X,$$

where $\|X\|^2 = X'(\Sigma + A)^{-1}X/\rho$. This estimator "corrects" X in the direction $\Sigma(\Sigma + A)^{-1}X$ exactly as does the best linear estimator, but controls the amount of correction in a way which is quite reasonable. To see this, note that if A is the "correct" prior covariance matrix, then (unconditionally) X has mean 0 and covariance matrix $(\Sigma + A)$. Therefore $\lim_{p\to\infty} X'(\Sigma + A)^{-1}X/p = 1$ with probability one. Hence $\|X\|^2 \sim p/\rho$ for large p. ("$\sim$" denotes approximate equality.) By Lemma 2.2.1 (vi), it follows that, for large p,

$$r_{(p-2)/2}(\|X\|^2) \sim p(\min\{1, 1/\rho\}).$$

Thus, if A is correct, p is large, and $\rho \leqslant 1$, then

$$\delta^{(p-2)/2}(\|X\|^2) \sim (I - \Sigma(\Sigma + A)^{-1})X$$

as would be desired. If, on the other hand, A is wrong, or θ is not in the region expected (i.e., near zero), then $[X'(\Sigma + A)^{-1}X]$ will tend to be much larger than $r_{(p-2)/2}(\|X\|^2)$ (which, recall, is bounded by $(p - 2)$), and $\delta^{(p-2)/2}$ will correct $\delta^0(X) = X$ very little.

The above considerations are not meant to prove anything, but merely to indicate why the suggested estimator is reasonable. Note, in particular, that choosing $2n \sim p$ (as in $n = (p - 2)/2$) was necessary to obtain the desired convergence to the best linear estimator for large p.

A decision must also be made as to what value of ρ to use. Note that ρ affects δ^n only through $r_n(X'(\Sigma + A)^{-1}X/\rho)$. It is clear from Lemma 2.1.1 (ii) that r_n is decreasing in ρ, so that larger ρ result in more conservative estimators (in that they are closer to $\delta^0(X) = X$). One reasonable choice of ρ follows from the observation (using Lemma 2.1.1 (iv)) that for small X

$$\delta^n(X) \sim \left(I - \frac{n}{\rho(n + 1)}\Sigma(\Sigma + A)^{-1}\right)X.$$

Since a small X strongly supports the prior beliefs and the best linear estimator is reasonable in such a situation, the indicated choice of ρ is

$$\rho = n/(n + 1). \tag{2.9}$$

A slight annoyance is that in the development of δ^n we were constrained to have $C \geqslant \Sigma$. For $C = \rho(\Sigma + A)$, this is satisfied only if $\rho \geqslant \text{ch}_{\max}\{\Sigma(\Sigma + A)^{-1}\}$. This inequality will typically hold for $\rho = n/(n + 1)$, but even if it does not there seems to be no clear reason to worry about it. The estimator δ^n will be very reasonable and appropriate even if $\rho < \text{ch}_{\max}\{\Sigma(\Sigma + A)^{-1}\}$ (although it will no longer clearly be generalized Bayes). Hence, for simplicity, the choice in (2.9) is recommended.

As a summary of the preceding arguments, the estimator suggested for use is

$$\delta^*(X) = \left(I - \frac{r^*\left(X'(\Sigma + A)^{-1}X/\rho^*\right)\Sigma(\Sigma + A)^{-1}}{X'(\Sigma + A)^{-1}X}\right)X, \tag{2.10}$$

where $r^* = r_{(p-2)/2}$ and $\rho^* = (p - 2)/p$.

2.2. *Evaluation of* δ^*. The estimator δ^* will now be examined carefully to see if it satisfies properties 1 through 7 given in Section 1. Some of what follows pertains to the whole class of estimators δ^n, while some refers specifically to δ^*. Which is being discussed will clearly be indicated.

PROPERTY 1. δ^* readily allows the incorporation of prior knowledge, which was the main goal. The question arises as to whether the incorporation of the prior knowledge, A, leads to a significant improvement for estimators of this form (assuming the prior knowledge is approximately correct). To investigate this question p-variate normal priors $\xi(\theta)$, with mean 0 and covariance matrix τB were considered. Note that these priors are not really close to the priors $g_n(\theta)$ in terms of tail behavior. Therefore, we are not loading the dice in favor of the estimator δ^*. (Of course the priors ξ have the same mean (here assumed to be zero) as the priors g_n, and this is a significant similarity. If the priors differed in this respect also, however, there would be little to compare, since drastically different priors will, of course, give quite different conclusions.) The Bayes risks $r(\xi, \delta) = \int R(\theta, \delta)\xi(\theta)\, d\theta$ of three estimators, $\delta^{\tau B}$, δ^B, and δ^I were compared. $\delta^{\tau B}$ is δ^* with the "correct" choice $A = \tau B$. δ^B is δ^* with $A = B$, meaning the wrong scale factor is being used. δ^I is δ^* with $A = I$, so that an entirely wrong covariance matrix is being used. Typical of the numerical results obtained are those given in the first three rows of Table 1. The calculations there are for $p = 6$, $Q = \Sigma = I$, and B diagonal with diagonal elements $\{.1, .5, 1, 3, 6, 16\}$. (Note that this is a wide spread of variances (for $\tau = 1$ anyway), in that some coordinates have comparatively small sample variance, some have comparatively small prior variance, and some are in between.) For varying τ, the Bayes risks of $\delta^{\tau B}$, δ^B, and δ^I are given in Table 1. $\delta^{\tau B}$ is clearly best, while δ^B is significantly better than δ^I. Thus it appears that the incorporation of prior knowledge in δ^* is quite worthwhile, though it need not be absolutely correct in order to achieve significant gains. Note that the Bayes risk of the usual

TABLE 1
Bayes risks

τ	.25	.50	.75	1.0	2.0	5.0	10.0	25.0	50.0
$\delta^{\tau B}$	3.19	3.64	3.90	4.12	4.59	5.10	5.39	5.69	5.82
δ^B	3.43	3.73	3.92	4.12	4.67	5.28	5.60	5.82	5.91
δ^I	3.77	4.38	4.70	4.95	5.41	5.74	5.86	5.94	5.97
$\delta_L^{\tau B}$	2.16	2.82	3.21	3.47	4.08	4.77	5.12	5.58	5.76
δ_L^B	2.78	3.01	3.24	3.47	4.39	7.15	11.75	25.56	48.56
δ_L^I	3.16	4.83	6.49	8.15	14.80	34.80	68.00	167.8	334.0

estimator is $r(\xi, \delta^0) = 6$. (The entries in the table were calculated by simulation and have a standard error of .02.)

PROPERTY 2. δ^* is quite robust with respect to misspecification of prior information, and has a risk $R(\theta, \delta^*)$ which compares favorably with $R(\theta, \delta^0)$.

The robustness is indicated by Table 1. δ^B and δ^I use (at say $\tau = 50$) drastically wrong prior information, and yet still have better Bayes risks than δ^0. Indeed it can be shown that $\lim_{\tau\to\infty} r(\xi, \delta^*) = 6$ no matter what fixed A is used in δ^* (since $\lim_{\|x\|\to\infty} |\delta^*(x) - x| = 0$). This compares quite favorably with the corresponding situation when linear Bayes estimators are used. The estimators $\delta_L^{\tau B}$, δ_L^B, and δ_L^I in Table 1 are the linear estimators defined by

$$\delta_L^A(X) = \left(I - \Sigma(\Sigma + A)^{-1}\right)X.$$

Thus $\delta_L^{\tau B}$ is the optimum linear and, indeed, optimum Bayes estimator for the situation of Table 1. δ_L^B and δ_L^I correspond to misspecified prior information. The risks given in Table 1 show the nonrobustness of the linear estimators compared to δ^*. The case for estimators such as δ^* would be even more telling if prior densities with flat tails were used. (We are looking at linear estimators on their home ground so to speak.)

Studies of Bayes risks alone tend to put estimators such as δ^* in a very flattering light. To discover the seamier side of such estimators, it is important to look at the regular risk $R(\theta, \delta^*)$ in comparison with $R(\theta, \delta^0)$. Since δ^* generally pulls $\delta^0(X) = X$ closer to zero, it can be expected that $R(\theta, \delta^*) < R(\theta, \delta^0)$ for θ in a neighborhood of zero. It also usually turns out to be true that $R(\theta, \delta^*) < R(\theta, \delta^0)$ for θ in certain directions of the parameter space. The reverse inequality can hold in other directions. Of usefulness in analyzing this behavior are the results of Berger (1976b). (See also Brown (1979)). Using Theorem 1, Lemma 1, and Lemma 2 of Berger (1976b), together with Lemma 2.1.1 (i), (iii), and (ix) of this paper, it can be shown that

$$\Delta(\theta) = R(\theta, \delta^n) - R(\theta, \delta^0) = \frac{-4n}{\theta' C^{-1}\theta}\left\{\operatorname{tr}(\Sigma Q \Sigma C^{-1}) - \frac{(2+n)\theta' C^{-1}\Sigma Q \Sigma C^{-1}\theta}{\theta' C^{-1}\theta}\right\} + o(|\theta|^{-2}). \tag{2.11}$$

Note immediately that $\Delta(\theta) \to 0$ as $|\theta| \to \infty$. Furthermore, if $|\theta|$ is large enough, it follows that $\Delta(\theta) < 0$ if

$$\frac{(2+n)\theta' C^{-1}\Sigma Q \Sigma C^{-1}\theta}{\theta' C^{-1}\theta} < \operatorname{tr}(\Sigma Q \Sigma C^{-1}). \tag{2.12}$$

From (2.12) can be determined the directions in which $\Delta(\theta) < 0$ for large $|\theta|$. Usually, $R(\theta, \delta^n)$ will be less than $R(\theta, \delta^0)$ for all θ in these directions. Note in particular that if (2.12) holds for all θ, or, equivalently, if

$$(2+n)\operatorname{ch}_{\max}(\Sigma Q \Sigma C^{-1}) < \operatorname{tr}(\Sigma Q \Sigma C^{-1}), \tag{2.13}$$

then $\Delta(\theta) < 0$ for large $|\theta|$. Indeed, using Theorem 1 of Berger (1976c), the following stronger result can be obtained:

THEOREM 2.2.1. *If* $(2 + n)\text{ch}_{\max}(\Sigma Q \Sigma C^{-1}) \leqslant \text{tr}(\Sigma Q \Sigma C^{-1})$, *then* $R(\theta, \delta^n) \leqslant R(\theta, \delta^0)$ *for all* θ. (δ^n *is hence minimax.*)

PROOF. Since $n > 0$, the condition of the theorem clearly ensures that $p \geqslant 3$. Assumptions (i) and (ii) of Theorem 1 of Berger (1976c) follow immediately. Assumptions (iii) and (iv) of Berger (1976c) can be verified by a simple calculation of ∇r_n (the gradient of r_n), together with Lemma 2.1.1 (ii), (iii) and (ix). Assumption (v) of Berger (1976c) is satisfied by the condition given in the theorem. The conclusion follows. □

COROLLARY 2.2.2. δ^* *is minimax if* $p \geqslant 3$ *and*

$$(2.14) \qquad (p + 2)\text{ch}_{\max}\{\Sigma Q \Sigma(\Sigma + A)^{-1}\} \leqslant 2\,\text{tr}\{\Sigma Q \Sigma(\Sigma + A)^{-1}\}.$$

This is true in particular if

(a) $Q = c_1 I$, $\Sigma = c_2 I$, $A = c_3 I$, $c_1 > 0$, $c_2 > 0$, $c_3 \geqslant 0$;
(b) $Q = c(I + \Sigma^{-1}A)\Sigma^{-1}$, $c > 0$; *or*
(c) $A = c\Sigma Q \Sigma - \Sigma$, $c \geqslant \text{ch}_{\max}(\Sigma^{-1}Q^{-1})$.

PROOF. Immediate. □

Thus, whenever (2.14) holds, δ^* will have risk smaller than $R(\theta, \delta^0)$ for all θ, a very nice situation. Note, in particular, that this will be true for the symmetric situation described in (a) of the corollary. (The result of Corollary 2.2.1 was obtained in the particular case $A = [\text{ch}_{\max}(\Sigma^{-1}Q^{-1})\Sigma Q \Sigma - \Sigma]$ in Berger (1976a), and in the case $Q = \Sigma = I$ and $A = 0$ in Strawderman (1971).)

Unfortunately, for nonsymmetric problems (2.14) will not typically be satisfied. It is instructive to examine the risk function $R(\theta, \delta^*)$ in such a situation by numerical methods. The situation considered was $p = 6$, $Q = I$, A diagonal with diagonal elements $\{1.55, 2, 2, 2, 2, 6.5\}$, and Σ diagonal with diagonal elements $\{.1, 1, 1, 1, 1, 10\}$. This is a case of considerable nonsymmetry where δ^* can be expected to be worse than δ^0 in certain directions of the parameter space. The normalized risk function $R(\theta, \delta^*)/\text{tr}(Q\Sigma)$ was numerically computed along the six coordinates axes and along the line $\theta = |\theta|(1, 1, 1, 1, 1, 1)^t/6^{\frac{1}{2}}$. From (2.12) we would expect that $\Delta(\theta) < 0$ along all these lines except the θ_6 axis. The numerical results in Figure 1 bear this out. ($R(\theta, \delta^0)/\text{tr}(Q\Sigma) = 1$ is the constant line on the graph.)

The risk of δ^* along the θ_6 axis appears to be seriously worse than that of δ^0, but recall that the prior belief is roughly that θ_6 has mean 0 and variance 6.5. If indeed θ_6 turns out to be 5 standard deviations away from zero, some penalty must be expected. Note, at least, that the penalty is bounded. For comparison purposes, the normalized risk $R(\theta, \delta_L)/\text{tr}(Q\Sigma)$ along the θ_6 axis of the optimum linear Bayes estimator is also given in Figure 1. The comparative robustness of δ^* is clear.

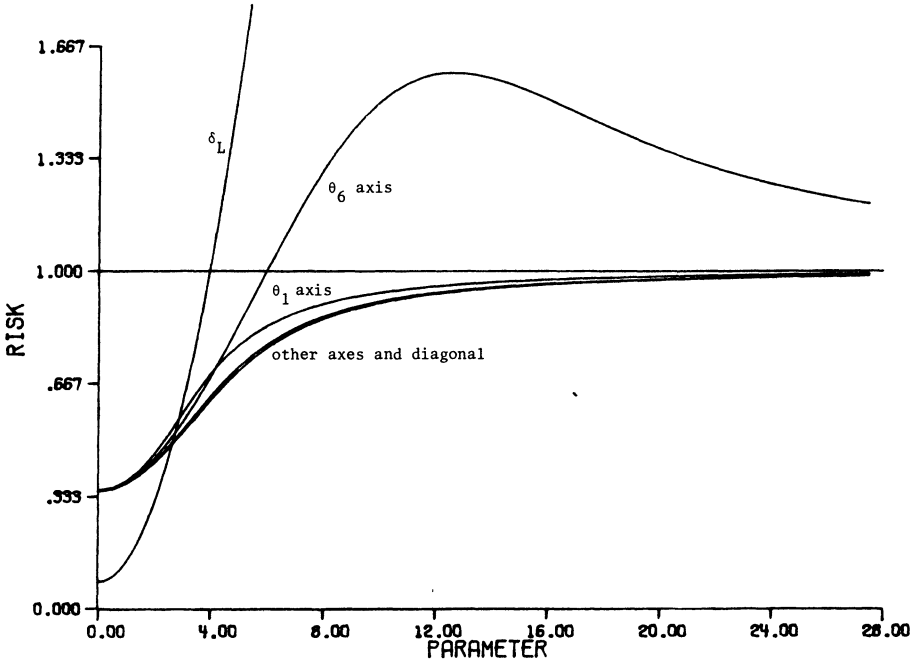

Figure 1.

Similar behavior has been observed in all realistic examples investigated. We have not found a plausible situation in which the risk of δ^* is bad for θ near the prior beliefs. The worst hypothetical case is when, say, $q_i = \varepsilon$ for $1 \leqslant i \leqslant p - 1$, where ε is much smaller than q_p and p is large. The major contribution to the risk is then from the pth coordinate, and it can be checked that δ_p^* will behave roughly like the linear Bayes estimator for reasonable values of θ. The risk of δ^* can thus get fairly bad if only one coordinate of θ is important.

Figure 1 makes it graphically clear that in using δ^*, minimaxity will often be sacrificed. It seems reasonable, however, to give up minimaxity in unimportant areas of the parameter space in order to achieve sizeable improvement elsewhere. Minimax estimators do not appear to be able to achieve the sizeable gains in (Bayes) risk offered by δ^* in nonsymmetric problems. See Berger (1979) for a discussion of this (along with the development of a minimax estimator that at least allows the incorporation of prior information).

δ^* has another advantage over minimax estimators in applications. This is that the loss matrix Q need not be known in order to calculate δ^*. (On the other hand, Q plays a crucial role in all minimax estimators.) In applications it is usually much easier to obtain prior information (like μ and A) from a client, than it is to obtain Q. (People will readily guess where θ is, but are reluctant to say which coordinates are more important than others.) This point was also made by Morris (1977).

PROPERTY 3. δ^* is clearly relatively easy to calculate, use and analyze.

PROPERTY 4. Stochastic ridge estimators make no formal allowance for prior information, but they are similar to δ^n with the choices $C = [\text{ch}_{\max}(\Sigma)]I$ and $n = p/2$. Hence estimators δ^n can be found with about the same "stability" as stochastic ridge estimators (which, however, may not be as stable as fixed constant ridge estimators). (See Casella (1977) for some definitions of stability). The prior input into an estimator seems far more important than its stability, however, so no attempt was made in choosing δ^* to force it to be stable.

PROPERTY 5. As in Berger (1976a), the results of Brown (1971) (in particular Theorem 6.4.2) can be used to show that δ^n is admissible if $n \geqslant (p-2)/2$ and $\rho \geqslant \text{ch}_{\max}[\Sigma(\Sigma + A)^{-1}]$, but inadmissible if $n < (p-2)/2$. Thus δ^* is admissible if $\text{ch}_{\max}[\Sigma(\Sigma + A)^{-1}] \leqslant (p-2)/p$.

As indicated previously, the flatter the tails of a prior density are, the more robust the generalized Bayes estimator derived from the prior tends to be. Since, for $g_n(\theta)$, smaller n correspond to flatter tails, it appears that δ^* is about as robust as possible (in terms of choice of n), while preserving admissibility. This was another reason for choosing $n = (p-2)/2$.

PROPERTY 6. The discussion leading to the choice of δ^* in Section 2.1 showed that δ^* has a crude empirical Bayes property—namely that if A is chosen correctly and p is large, then δ^* is approximately the optimum linear Bayes estimator. For the symmetric empirical Bayes situation discussed in Section 1, the following stronger empirical Bayes property can be obtained.

Assume $\Sigma = \sigma^2 I$, $A = cI$, and the θ_i are a sample from a prior distribution with mean zero and variance $\tau^2 > c$. Note that for δ^*,

$$\lim_{p\to\infty} \frac{\|X\|^2}{p} = \frac{|X|^2}{p\rho^*(\sigma^2 + c)} = \frac{(\sigma^2 + \tau^2)}{(\sigma^2 + c)} \qquad \text{(with probability one).}$$

Lemma 2.1.1 (vi) can be used to conclude that

$$\lim_{p\to\infty} \frac{r^*(\|X\|^2)}{p} = \min\left\{1, \frac{\sigma^2 + \tau^2}{(\sigma^2 + c)}\right\} = 1 \qquad \text{(with probability one).}$$

Hence $\delta^*(X)$ behaves like (1.2) as desired.

3. Confidence regions for θ. While there has been a great deal of research on multivariate estimation of θ, there has been comparatively little on the development of improved confidence regions for θ. The theoretical works of Brown (1966) and Joshi (1967) established that the usual confidence region could be improved upon, but did not provide explicit improved confidence regions. By the usual confidence region is meant

$$C^0(X) = \{\theta : (X-\theta)'\Sigma^{-1}(X-\theta) \leqslant k(\alpha)\},$$

where $k(\alpha)$ is the $100(1-\alpha)$th percentile of the chi-square distribution with p degrees of freedom. Stein (1962) and (1974) suggests certain confidence regions for

large p (based on heuristic considerations), but leaves open the question of what to do for small or moderate p. Faith (1976) in the symmetric situation ($\Sigma = A = I$) develops Bayesian confidence regions using priors similar to $g_n(\theta)$, and gives convincing numerical and theoretical arguments to support their superiority over C^0. Unfortunately, his confidence regions are difficult to work with, having a complicated shape arising from their Bayesian derivation.

Morris (1977) suggests an appealing way to proceed in a Bayesian fashion with a resulting confidence region which is fairly simple. In the symmetric situation he considers the prior $g_n(\theta)$ with $n = (p-2)/2$ and $C = I$, and calculates the posterior mean, $\delta^n(X)$, and posterior covariance matrix, $\Sigma_n(X)$. He uses the diagonal elements of $\Sigma_n(X)$ to construct confidence intervals for the θ_i, centered at $\delta_i^n(X)$. The resulting confidence region is simple and yet hopefully retains the advantage of a robust Bayesian approach. We will differ somewhat from Morris by considering confidence ellipsoids based on the entire $\Sigma_n(X)$, and, of course, dealing with the nonsymmetric situation.

3.1. *Development of the confidence region.* The first step is the calculation of $\Sigma_n(X)$, the covariance matrix of the posterior distribution of θ given X (for the prior $g_n(\theta)$). Clearly, $\Sigma_n(X)$ is given by

(3.1)
$$\Sigma_n(X) = \frac{\int[\theta - \delta^n(X)][\theta - \delta^n(X)]^t \exp\{-(X-\theta)^t\Sigma^{-1}(X-\theta)/2\}\, g_n(\theta)\, d\theta}{\int \exp\{-(X-\theta)^t\Sigma^{-1}(X-\theta)/2\}\, g_n(\theta)\, d\theta}.$$

Completing squares and interchanging orders of integration, as in Section 2, gives that the numerator of (3.1) is

(3.2)
$$\int_0^1 \exp\left\{-X^t\left[\Sigma^{-1} - \Sigma^{-1}\left(\Sigma^{-1} + B(\lambda)^{-1}\right)^{-1}\Sigma^{-1}\right]X/2\right\}\left[\det B(\lambda)\right]^{-\frac{1}{2}}\lambda^{(n-1-p/2)}$$
$$\times \int_{R^p}\left[\theta\theta^t - \delta^n(X)\delta^n(X)^t\right]\exp\left\{-(\theta - z)^t\left(\Sigma^{-1} + B(\lambda)^{-1}\right)(\theta - z)/2\right\}\, d\theta\, d\lambda,$$

where $z = (\Sigma^{-1} + B(\lambda)^{-1})^{-1}\Sigma^{-1}X$. Replacing $[\theta\theta^t]$ by $[(\theta - z)(\theta - z)^t + \theta z^t + z\theta^t - zz^t]$ and integrating over θ, the inside integral in (3.2) is equal to

$$\left[\det\left(\Sigma^{-1} + B(\lambda)^{-1}\right)\right]^{-\frac{1}{2}}\left[\left(\Sigma^{-1} + B(\lambda)^{-1}\right)^{-1} + zz^t - \delta^n(X)\delta^n(X)^t\right].$$

Using this along with the identities in (2.1) and the definitions of $\delta^n(X)$ and z, the expression (3.2) can be calculated to be

(3.3)
$$\int_0^1 \exp\{-\lambda\|X\|^2/2\}\lambda^{n-1}\left[\det(\Sigma^{-1}C)\right]^{-\frac{1}{2}}$$
$$\left[\Sigma - \lambda\Sigma C^{-1}\Sigma + \left(\frac{r_n(\|X\|^2)}{\|X\|^2} - \lambda\right)(\Sigma C^{-1}XX^t + XX^tC^{-1}\Sigma)\right.$$
$$\left. + \left(\lambda^2 - \frac{r_n^2(\|X\|^2)}{\|X\|^4}\right)\Sigma C^{-1}XX^tC^{-1}\Sigma\right] d\lambda.$$

Using (2.2), (2.3), (3.1), (3.3), and defining

$$t_n(v) = \frac{v^2 \int_0^1 \exp\{-\lambda v/2\}\lambda^{n+1}\, d\lambda}{\int_0^1 \exp\{-\lambda v/2\}\lambda^{n-1}\, d\lambda}, \tag{3.4}$$

it follows that

$$\Sigma_n(X) = \Sigma - \frac{r_n(\|X\|^2)}{\|X\|^2}\Sigma C^{-1}\Sigma + \frac{\left[t_n(\|X\|^2) - r_n^2(\|X\|^2)\right]}{\|X\|^4}\Sigma C^{-1}XX^tC^{-1}\Sigma. \tag{3.5}$$

From (2.3) and (2.6) it is clear that

$$t_n(v) = r_n(v)r_{n+1}(v) = r_n(v)\left[2(n+1) + v\right] - 2nv. \tag{3.6}$$

The following properties of t_n will be needed.

LEMMA 3.1.1. *If $n > 0$, then*

(i) $0 < t_n(v) < 4n(n+1)$;
(ii) $\lim_{v\to\infty} t_n(v) = 4n(n+1)$;
(iii) $\lim_{v\to 0}[t_n(v)/\{nv^2/(n+2)\}] = 1$;
(iv) $t_n(v) - r_n^2(v) = 2r_n(v) - 2vr_n'(v)$;
(v) $0 < t_n(v) - r_n^2(v) < 2r_n(v)$.

PROOF. Parts (i), (ii), and (iii) follow from (3.6) and Lemma 2.1.1. To prove part (iv), note that differentiating in (2.3) gives

$$r_n'(v) = \frac{r_n(v)}{v} - \frac{t_n(v)}{2v} + \frac{r_n^2(v)}{2v}.$$

Rearranging terms gives the desired result. The upper bound in part (v) follows immediately from (iv) and Lemma 2.1.1 (ii). To establish the lower bound, assume that λ is distributed as in (2.5), and note from (2.3) and (3.4) that

$$\left[t_n(v) - r_n^2(v)\right]/v^2 = E[\lambda^2] - (E[\lambda])^2 = E[\lambda - E(\lambda)]^2 > 0.$$

□

The following lemma will be needed later on, and provides an interesting bound on $\Sigma_n(X)$. For two $(p \times p)$ matrices A and B, let $A \leqslant B$ mean that $(B - A)$ is positive semidefinite.

LEMMA 3.1.2.

$$\Sigma - \frac{n}{(n+1)}\Sigma C^{-1}\Sigma \leqslant \Sigma_n(X) \leqslant \Sigma + \frac{n}{(n+1)}\Sigma C^{-1}\Sigma. \tag{3.7}$$

PROOF. The lower bound follows from (3.5), using Lemma 3.1.1 (v) and Lemma 2.1.1 (x). The upper bound follows from Lemma 3.1.1 (v), Lemma 2.1.1 (x), and the fact that $\|X\|^{-2}C^{-\frac{1}{2}}XX^tC^{-\frac{1}{2}} \leqslant I$. □

The lower bound in Lemma 3.1.2 is sharp, in that $\Sigma_n(0) = \Sigma - (n/n+1)\Sigma C^{-1}\Sigma$. (This follows from Lemma 3.1.1 (iii) and Lemma 2.1.1 (iv).) The upper

bound is not sharp in that the rank one matrix $C^{-\frac{1}{2}}XX'C^{-\frac{1}{2}}$ was bounded by the rank p matrix $\|X\|^2 I$.

The confidence regions that will be considered are the ellipsoids

$$(3.8)\quad C^n(X) = \left\{\theta \in R^p : [\theta - \delta^n(X)]'\Sigma_n(X)^{-1}[\theta - \delta^n(X)] \leqslant k(\alpha)\right\},$$

where $k(\alpha)$ is the $100(1-\alpha)$th percentile of the chi-square distribution with p degrees of freedom. Note that these are not the true Bayesian confidence sets for the priors g_n, but are only approximations based on the posterior means and covariances. They do have a familiar shape, however, and are quite easy to work with. In the calculation of $\Sigma_n(X)^{-1}$, the following lemma is useful.

LEMMA 3.1.3. *If Y is a $(p \times 1)$ vector and B a $(p \times p)$ matrix, then*

$$(I + YY'B)^{-1} = \left(I - [1 + Y'BY]^{-1}YY'B\right).$$

PROOF. Calculation. □

For convenience, define

$$(3.9)\quad u = u(\|X\|^2) = \frac{r_n(\|X\|^2)}{\|X\|^2}, \qquad w = w(\|X\|^2) = \frac{t_n(\|X\|^2) - r_n^2(\|X\|^2)}{\|X\|^4}.$$

(As observed in the proofs of Lemmas 2.1.1 and 3.1.1, these quantities can be interpreted as the mean and variance of the formal posterior distribution of λ given X. The dependence of u and w on n will be suppressed.)

Letting $B = (\Sigma - u\Sigma C^{-1}\Sigma)^{-1}$ and $Y = \Sigma C^{-1}X$, Lemma 3.1.3 can be applied to (3.5) to give

(3.10)

$$\begin{aligned}\Sigma_n(X)^{-1} &= (\Sigma - u\Sigma C^{-1}\Sigma)^{-1}\left(I + w\Sigma C^{-1}XX'C^{-1}\Sigma[\Sigma - u\Sigma C^{-1}\Sigma]^{-1}\right)^{-1}\\ &= B\left(I - [1 + wY'BY]^{-1}wYY'B\right).\end{aligned}$$

Thus the calculational problem is reduced to finding $B = (\Sigma - u\Sigma C^{-1}\Sigma)^{-1}$. If, in particular, $\Sigma = I$ and $C = \rho I$, then

$$(3.11)\qquad \Sigma_n(X)^{-1} = (1 - u/\rho)^{-1}\left(I - wXX'/[\rho^2 - \rho u + w|X|^2]\right).$$

The particular choice of n and C that is recommended is $n = (p-2)/2$ and $C = \rho^*(\Sigma + A)$ (ρ^* defined in (2.10)), so that the resulting confidence region is centered at δ^*. Let $C^*(X)$, $\Sigma^*(X)$, and t^* denote $C^n(X)$, $\Sigma_n(X)$, and t_n for these choices of n and C. Thus

$$\begin{aligned}(3.12)\quad \Sigma^*(X) = \Sigma &- \frac{r^*(\|X\|^2)}{\rho^*\|X\|^2}\Sigma(\Sigma + A)^{-1}\Sigma\\ &+ \frac{\left[t^*(\|X\|^2) - r^*(\|X\|^2)^2\right]}{\rho^{*2}\|X\|^4}\Sigma(\Sigma + A)^{-1}XX'(\Sigma + A)^{-1}\Sigma,\end{aligned}$$

where $\|X\|^2 = X'(\Sigma + A)^{-1}X/\rho^*$, and

$$C^*(X) = \{\theta : [\theta - \delta^*(X)]'\Sigma^*(X)^{-1}[\theta - \delta^*(X)] \leqslant k(\alpha)\}. \tag{3.13}$$

It is interesting to consider certain intuitive explanations for the terms of $\Sigma^*(X)$. Note first that in the standard Bayesian model where θ has a multivariate normal distribution with mean vector zero and covariance matrix A, the posterior covariance matrix is

$$(\Sigma^{-1} + A^{-1})^{-1} = \Sigma - \Sigma(\Sigma + A)^{-1}\Sigma. \tag{3.14}$$

In Section 2.1 it was shown that if A is the correct prior covariance matrix and p is large, then $r^*(\|X\|^2)/(\rho^*\|X\|^2) \sim 1$. Hence the first two terms of $\Sigma^*(X)$ behave like (3.14) when A is correct and p is large. On the other hand, if the A used is incorrect, then $r^*(\|X\|^2)/(\rho^*\|X\|^2)$ will usually be small and $\Sigma^*(X)$ will behave more like Σ. Note that the last term of $\Sigma^*(X)$ is relatively insignificant in large p situations, since it is a rank one matrix.

Another appealing facet of the large p behavior of $C^*(X)$ is that, for the symmetric situation ($\Sigma = I, A = \tau I$), $C^*(X)$ is similar to the confidence region suggested by Stein (1962). Indeed when $\|X\|^2 \geqslant p$ (the likely situation for large p), then $r^*(\|X\|^2) \sim p$, so (ignoring the rank one third term)

$$C^*(X) \sim \{\theta : |\theta - (1 - p/|X|^2)X|^2 \leqslant (1 - p/|X|^2)\kappa(\alpha)\},$$

which is the confidence region suggested by Stein up to first order terms.

The third term of $\Sigma^*(X)$ seems rather strange at first sight. It has a very reasonable intuitive explanation, however. Note that the characteristic vector corresponding to the nonzero characteristic root of the third term of $\Sigma^*(X)$ is $z = \Sigma(\Sigma + A)^{-1}X$. Hence in the direction of z, the contribution of the third term is positive. (The confidence ellipsoid is widened.) In directions perpendicular to z the third term is zero and the confidence ellipsoid is narrowed.

To intuitively explain this phenomenon, note that δ^* (at which Σ^* is centered) performs relatively badly when it "corrects" X along the same line that contains θ. (Correcting only along a line results in essentially a one dimensional problem.) δ^* achieves its gains when correcting those X for which the direction of correction is close to perpendicular to $(X - \theta)$. This phenomenon is exhibited in Figure 2, where θ is shown with four symmetrically placed possible X values (intended to crudely represent a spherically symmetric distribution). Assume the simple estimator $\delta^*(X) = (1 - r^*(|X|^2)/|X|^2)X$ is being used, so the X values will be shrunk towards zero. Clearly the effect of δ^* upon x_1 and x_3 (the x's corrected along the line containing θ) is harmful, in that the average distance of $\delta^*(x_1)$ and $\delta^*(x_3)$ from θ is larger than the average distance of x_1 and x_3 from θ. On the other hand, δ^* moves x_2 and x_4 closer to θ. Thus the correction appears to be beneficial for those X for which the direction of correction is close to perpendicular to $(X - \theta)$. (This type of picture was shown to me by Lawrence Brown.)

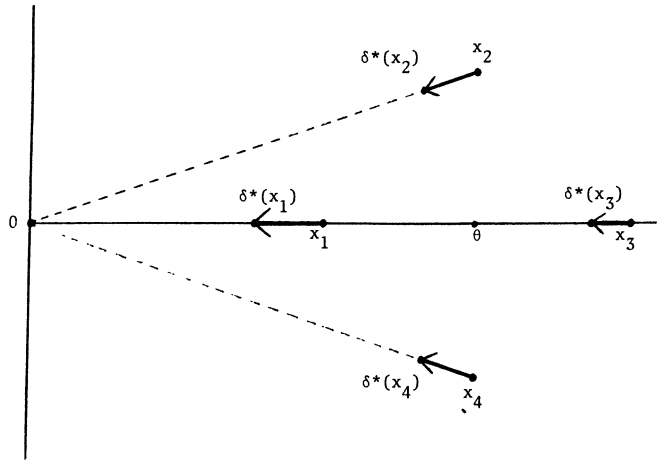

Figure 2.

Returning to the original situation, the heuristics in the above paragraph suggest that if θ lies in the same direction from X as z, then δ^* will not be doing too well. This harmful effect should be compensated for by widening the confidence region in that direction. This is precisely what $C^*(X)$ does.

Morris (1977) bases his confidence regions only upon the first two terms of $\Sigma^*(X)$ and the diagonal elements of the third term. The above argument indicates this may be undesirable.

We now proceed with a more rigorous analysis of the properties of $C^*(X)$. The two common criteria used in evaluating confidence regions are size and probability of coverage. Size will be considered first. (Many of the mathematical results which follow will be stated for general $\Sigma_n(X)$.)

3.2. *Size of* $C^n(X)$. There are a number of reasonable measures of the size of an ellipsoid. Virtually all are functions of the lengths of the semiaxes of the ellipsoid. For $C^n(X)$, the lengths of the semiaxes are proportional to the characteristic roots of $\Sigma_n(X)^{\frac{1}{2}}$. Actually, it is perhaps more appropriate to be concerned with the roots of $[Q\Sigma_n(X)]^{\frac{1}{2}}$, in order to take into account the relative importance of the various coordinates as reflected by Q. This is natural as can be seen by transforming the problem by $Q^{\frac{1}{2}}$ (i.e., define $Y = Q^{\frac{1}{2}}X$, $\eta = Q^{\frac{1}{2}}\theta$, etc.). In the transformed problem, the loss is sum of squares error loss so that all coordinates are of equal importance. It is easy to check that the posterior covariance matrix (given Y) in the transformed problem is

$$Q^{\frac{1}{2}}\Sigma_n\left(Q^{-\frac{1}{2}}Y\right)Q^{\frac{1}{2}} = Q^{\frac{1}{2}}\Sigma_n(X)Q^{\frac{1}{2}},$$

and hence it is natural to look at the characteristic roots of the square roots of this

matrix, or equivalently the roots of $[Q\Sigma_n(X)]^{\frac{1}{2}}$. For those who prefer to consider size of the original $\Sigma_n(X)$, merely set $Q = I$ in the results below.

The following three measures of size of $C^n(X)$ will be considered:

1. $\det[Q\Sigma_n(X)]^{\frac{1}{2}} = (\det Q)^{\frac{1}{2}}(\det \Sigma_n(X))^{\frac{1}{2}}$, which up to a multiplicative dimensional constant is the volume of the transformed confidence ellipsoid. Clearly it suffices to consider only $[\det \Sigma_n(x)]^{\frac{1}{2}}$, since Q occurs only in a multiplicative constant which will be the same for all transformed ellipsoids. Hence comparisons of volumes will be unaffected by Q.

2. $\operatorname{tr}[Q\Sigma_n(X)]^{\frac{1}{2}}$, which is the sum of the semiaxes of the transformed confidence ellipsoid.

3. $\operatorname{tr}[Q\Sigma_n(X)]$, which is the sum of the squares of the semiaxes of the transformed confidence ellipsoid. This measure of size is of additional interest since it is also the posterior expected loss of δ^n.

The results in this section will be concerned with comparing the size of $C^n(X)$ (and $C^*(X)$) to the size of $C^0(X)$, the usual confidence region. Note that for $C^0(X)$, the three measures of size that will be discussed are $\det(\Sigma^{\frac{1}{2}})$, $\operatorname{tr}(Q\Sigma)^{\frac{1}{2}}$, and $\operatorname{tr}(Q\Sigma)$ respectively.

The first result gives a condition on X under which $\Sigma_n(X) < \Sigma$, and hence $C^n(X)$ has smaller size than $C^0(X)$ under any reasonable measure of size. The notation in (3.9) will be used extensively from here on.

THEOREM 3.2.1. *If* $u(\|X\|^2) > w(\|X\|^2)$, *then* $\Sigma_n(X) < \Sigma$.

PROOF. This follows immediately from (3.5), noting that $\Sigma C^{-1}XX'C^{-1}\Sigma \leqslant \|X\|^2\Sigma C^{-1}\Sigma$. □

To investigate the measures of size, the following two lemmas are needed.

LEMMA 3.2.2.

$$\det \Sigma_n(X) = [\det \Sigma]\left[\det\left(I - u(\|X\|^2)C^{-1}\Sigma\right)\right] \times \left[1 + w(\|X\|^2)X'\left(C\Sigma^{-1}C - u(\|X\|^2)C\right)^{-1}X\right].$$

PROOF. Clearly

$$\det \Sigma_n(X) = [\det \Sigma]\left[\det(I - uC^{-1}\Sigma + wC^{-1}XX'C^{-1}\Sigma)\right] \tag{3.15}$$

$$= [\det \Sigma]\left[\det(I - uC^{-1}\Sigma)\right] \times \left[\det\left(I + wC^{-1}XX'C^{-1}\Sigma\{I - uC^{-1}\Sigma\}^{-1}\right)\right].$$

Note that $C^{-1}XX'B$ has rank one for any nonsingular $(p \times p)$ matrix B, and has characteristic roots 0 (with multiplicity $(p - 1)$) and $(X'BC^{-1}X)$. ($C^{-1}X$ is the characteristic vector of the nonzero root.) The characteristic roots of $[I + wC^{-2}XX'B]$ are hence 1 (with multiplicity $(p - 1)$) and $(1 + wX'BC^{-1}X)$. It

follows that

$$\det\bigl(I + wC^{-1}XX'C^{-1}\Sigma\{I - uC^{-1}\Sigma\}^{-1}\bigr)$$
$$= 1 + wX'C^{-1}\Sigma\{I - uC^{-1}\Sigma\}^{-1}C^{-1}X$$
$$= 1 + wX'(C\Sigma^{-1}C - uC)^{-1}X.$$

Together with (3.15) this gives the desired result. □

LEMMA 3.2.3. *Assume that* $a_i \geqslant 0$ *and* $b_i \geqslant 0$ $(i = 1, \cdots, p)$, *and that* $p \geqslant 2$, $\Sigma_{i=1}^p b_i = 1$, *and* $\Sigma_{i=1}^p a_i \geqslant 2 \max_{1 \leqslant i \leqslant p}\{a_i\}$. *Then*

$$\text{(3.16)}\quad \Pi_{i=1}^p(1 + ya_i[2b_i - 1]) \leqslant 1, \text{ for all } y \in \bigl[0, (\max_i\{a_i(1 - 2b_i)\})^{-1}\bigr].$$

PROOF. Without loss of generality assume that a_1 is the largest a_i. If $b_i \leqslant \frac{1}{2}$ for all i, the conclusion is obvious. Hence assume $b_j > \frac{1}{2}$ for some j. Note then that $b_i < \frac{1}{2}$ for all $i \neq j$. Examining (3.16), it is clear that the worst case to consider is $j = 1$ (since a_1 is the largest a_i). Thus assume $b_1 > \frac{1}{2}$.

Since $2a_1 \leqslant \Sigma_{i=1}^p a_i$ (or $a_1 \leqslant \Sigma_{i=2}^p a_i$), it is clear that

(3.7)

$$\Pi_{i=1}^p(1 + ya_i[2b_i - 1]) \leqslant (1 + y\{\Sigma_{i=2}^p a_i\}[2b_1 - 1])[\Pi_{i=2}^p(1 + ya_i[2b_i - 1])].$$

Denoting the right-hand side above by $\varphi(y)$, a calculation gives

$$\frac{d}{dy}\varphi(y) = \{\Sigma_{i=2}^p a_i\}[2b_1 - 1][\Pi_{i=2}^p(1 + ya_i[2b_i - 1])]$$
$$+ [\Pi_{i=2}^p(1 + ya_i[2b_i - 1])]\Sigma_{j=2}^p\left\{\frac{a_j[2b_j - 1](1 + y\{\Sigma_{i=2}^p a_i\}[2b_1 - 1])}{1 + ya_j[2b_j - 1]}\right\}$$
$$= [\Pi_{i=2}^p(1 + ya_i[2b_i - 1])]\Sigma_{j=2}^p$$
$$\times\left\{\frac{a_j[2(b_1 + b_j - 1) + y(2b_1 - 1)(2b_j - 1)(a_j + \Sigma_{i=2}^p a_i)]}{1 + ya_j[2b_j - 1]}\right\}.$$

Since $(2b_1 - 1) > 0$, $(2b_j - 1) < 0$, $a_i \geqslant 0$, $(b_1 + b_j - 1) \leqslant 0$, and $(1 + ya_j[2b_j - 1]) \geqslant 0$ (due to the domain of y), it is clear that $(d/dy)\varphi(y) \leqslant 0$. Hence $\varphi(y)$ is maximized at $y = 0$, which together with (3.17) establishes the result. □

Measure of size 1: volume. The following theorem gives conditions under which the volume of $C^n(X)$ is less than the volume of $C^0(X)$.

THEOREM 3.2.4. $[\det \Sigma_n(X)]^{\frac{1}{2}} \leqslant [\det \Sigma]^{\frac{1}{2}}$ *for all* X, *if and only if* $\operatorname{tr}(C^{-1}\Sigma) \geqslant 2\operatorname{ch}_{\max}(C^{-1}\Sigma)$.

PROOF. Using Lemma 3.2.2, it is clear that showing that $[\det \Sigma_n(X)]^{\frac{1}{2}} \leqslant [\det \Sigma]^{\frac{1}{2}}$ is equivalent to showing that

(3.18)

$$H = \bigl[\det\bigl(I - u(\|X\|^2)C^{-1}\Sigma\bigr)\bigr]\bigl[1 + w(\|X\|^2)X'\bigl(C\Sigma^{-1}C - u(\|X\|^2)C\bigr)^{-1}X\bigr] \leqslant 1.$$

For convenience, let T be orthogonal such that $T'C^{-\frac{1}{2}}\Sigma C^{-\frac{1}{2}}T = D$ is diagonal with diagonal elements $\{d_1, \cdots, d_p\}$, d_1 being the largest. Note that the condition $\operatorname{tr}(C^{-1}\Sigma) \geqslant 2\operatorname{ch}_{\max}(C^{-1}\Sigma)$ is simply

$$\Sigma_{i=1}^{p} d_i \geqslant 2d_1. \tag{3.19}$$

Also define $z = T'C^{-\frac{1}{2}}X$, so that $\|X\|^2 = X'C^{-1}X = |z|^2$. Then H can be rewritten

$$\begin{aligned} H &= \left[\Pi_{i=1}^{p}(1 - u(|z|^2)d_i)\right]\left[1 + w(|z|^2)z'(D^{-1} - u(|z|^2)I)^{-1}z\right] \\ &= \left[\Pi_{i=1}^{p}(1 - u(|z|^2)d_i)\right]\left[1 + w(|z|^2)\Sigma_{i=1}^{p}\{z_i^2 d_i/(1 - u(|z|^2)d_i)\}\right]. \end{aligned} \tag{3.20}$$

To prove the "only if" part of the theorem, choose $z = |z|(1, 0, \cdots, 0)'$. Then

$$\begin{aligned} H &= \left[\Pi_{i=1}^{p}(1 - u(|z|^2)d_i)\right]\left[1 + w(|z|^2)|z|^2 d_1/(1 - u(|z|^2)d_1)\right] \\ &= \left[\Pi_{i=2}^{p}(1 - u(|z|^2)d_i)\right]\left[1 + d_1\{w(|z|^2)|z|^2 - u(|z|^2)\}\right]. \end{aligned} \tag{3.21}$$

Letting $|z| \to \infty$ and using Lemma 2.1.1 (iii) and (ix) and Lemma 3.1.1 (iv), it is clear that

$$\begin{aligned} u(|z|^2) &= r_n(|z|^2)/|z|^2 = 2n/|z|^2 + o(|z|^{-2}), \\ w(|z|^2)|z|^2 - u(|z|^2) &= \frac{2r_n(|z|^2) - 2|z|^2 r_n'(|z|^2)}{|z|^2} - \frac{r_n(|z|^2)}{|z|^2} \\ &= 2n/|z|^2 + o(|z|^{-2}). \end{aligned}$$

Hence from (3.21)

$$\begin{aligned} H &= \left[1 - \frac{2n}{|z|^2}(\Sigma_{i=2}^{p} d_i) + o(|z|^{-2})\right]\left[1 + \frac{2nd_1}{|z|^2} + o(|z|^{-2})\right] \\ &= 1 + \frac{2n}{|z|^2}\{d_1 - \Sigma_{i=2}^{p} d_i\} + o(|z|^{-2}). \end{aligned}$$

Thus if (3.19) is violated, then $H > 1$ for large enough $|z|$ (and z in the given direction). This proves the "only if" part of the theorem.

To prove the "if" part, observe from (3.20) that

$$\begin{aligned} H &\leqslant \left[\Pi_{i=1}^{p}(1 - ud_i)\right]\left[\Pi_{i=1}^{p}(1 + wz_i^2 d_i/\{1 - ud_i\})\right] \\ &= \Pi_{i=1}^{p}(1 + d_i[wz_i^2 - u]) \\ &\leqslant \Pi_{i=1}^{p}\left(1 + ud_i\left[\frac{2z_i^2}{|z|^2} - 1\right]\right), \end{aligned} \tag{3.22}$$

the last step following from Lemma 3.1.1 (v). Letting $y = u$, $a_i = d_i$, $b_i = z_i^2/|z|^2$, and applying Lemma 3.2.3 gives that if (3.19) is satisfied, then $H \leqslant 1$, completing the proof. □

COROLLARY 3.2.5. *If* $C = \rho\Sigma(\rho \geqslant 1)$ *and* $p \geqslant 2$, *then* $[\det \Sigma_n(X)]^{\frac{1}{2}} \leqslant [\det \Sigma]^{\frac{1}{2}}$ *for all* X.

PROOF. $\operatorname{tr}(C^{-1}\Sigma) = p\tau \geqslant 2\tau = 2\mathrm{ch}_{\max}(C^{-1}\Sigma)$, so Theorem 3.2.4 gives the desired result. □

COROLLARY 3.2.6. *$C^*(X)$ has smaller volume than $C^0(X)$, for all X, if and only if*

$$\operatorname{tr}(I + A\Sigma^{-1})^{-1} \geqslant 2\mathrm{ch}_{\max}(I + A\Sigma^{-1})^{-1}.$$

PROOF. Obvious from Theorem 3.2.4, noting that $\Sigma(\Sigma + A)^{-1} = (I + A\Sigma^{-1})^{-1}$. □

Note in particular that for the symmetric problem where Σ and A are multiples of the identity, then $C^n(X)$ and $C^*(X)$ have smaller volumes than $C^0(X)$ for $p \geqslant 2$.

The question arises as to how significant an improvement in volume is obtainable using $C^*(X)$ instead of $C^0(X)$. Using Lemma 3.2.2 it is an easy matter to calculate

$$V^*(X) = \frac{\text{volume of } C^*(X)}{\text{volume of } C^0(X)} = \frac{[\det \Sigma^*(X)]^{\frac{1}{2}}}{[\det \Sigma]^{\frac{1}{2}}}.$$

Typical of the results obtained are those given in Tables 2 and 3 below. Table 2 considers the symmetric situation $\Sigma = I$ and $C = \rho^*(\Sigma + A) = 2I$, for $p = 6$ and $p = 12$. V_6^* and V_{12}^* are the volume ratios in 6 and 12 dimensions, respectively. $V^*(X)$ is a function of $|X|$ in this situation. It is somewhat easier to picture things in terms of

$$R^*(X) = [V^*(X)]^{1/p},$$

which is termed by Faith (1976) the ratio of the effective radii of $C^*(X)$ and $C^0(X)$. (The effective radius of a set is the radius of a p-sphere having the same volume as the set.) In Table 2, R_6^* and R_{12}^* stand for $R^*(X)$ in 6 and 12 dimensions. In the symmetric situation $C^*(X)$ is clearly significantly smaller than $C^0(X)$.

Table 3 deals with the nonsymmetric situation $p = 6$, $\Sigma = I$, and A diagonal with diagonal elements $\{.65, 3.5, 6.5, 9.5, 12.5, 45.5\}$. The entries V_i^* and $R_i^*(1 \leqslant i \leqslant 6)$ refer to the quantities $V^*(X)$ and $R^*(X)$ calculated along the ith axis. V_7^* and R_7^* are calculated along the line $|X|(1, 1, 1, 1, 1, 1)'/6^{\frac{1}{2}}$. Note that

$$\operatorname{tr}(I + A\Sigma^{-1})^{-1} = 1.729 < 1.818 = 2\mathrm{ch}_{\max}(I + A\Sigma^{-1})^{-1},$$

so by Corollary 3.2.6 it must be true that $V^*(X) > 1$ for some X. Indeed for large $|X|$ along the first axis, Table 3 shows that this is the case. Such X are very unlikely

TABLE 2
Volume ratio

$\|X\|$	0	1.0	2.0	4.0	6.0	8.0	10.0	20.0	50.0
V_6^*	.296	.309	.352	.561	.784	.877	.921	.980	.997
R_6^*	.816	.822	.840	.908	.960	.978	.986	.997	.999
V_{12}^*	.039	.041	.045	.075	.201	.422	.588	.881	.980
R_{12}^*	.764	.766	.772	.806	.875	.931	.957	.990	.998

TABLE 3
Volume ratio

$\|X\|$	0	1.0	3.0	5.0	7.0	9.0	11.0	20.0	50.0
V_1^*	.467	.514	.858	1.000	1.002	1.001	1.001	1.000	1.000
R_1^*	.881	.895	.975	1.000	1.000	1.000	1.000	1.000	1.000
V_2^*	.467	.476	.556	.722	.861	.919	.946	.984	.997
R_2^*	.881	.884	.907	.947	.975	.986	.991	.997	1.000
V_3^*	.467	.471	.513	.605	.732	.833	.889	.967	.995
R_3^*	.881	.882	.895	.920	.949	.970	.981	.994	.999
V_4^*	.467	.470	.498	.559	.654	.756	.833	.950	.992
R_4^*	.881	.882	.890	.908	.932	.955	.970	.991	.999
V_5^*	.467	.469	.490	.536	.608	.698	.781	.932	.989
R_5^*	.881	.882	.888	.901	.920	.942	.960	.988	.998
V_6^*	.467	.467	.473	.484	.502	.528	.560	.757	.959
R_6^*	.881	.881	.883	.886	.892	.899	.908	.955	.993
V_7^*	.467	.477	.573	.755	.901	.949	.967	.990	.998
R_7^*	.881	.884	.911	.954	.983	.991	.994	.998	1.000

to occur, however, if the prior information that θ_1 has mean 0 and variance .65 is even approximately correct.

Measure of size 2: sum of semiaxes. For the second measure of size, $\mathrm{tr}[Q\Sigma_n(X)]^{\frac{1}{2}}$, general results were not obtained. However, for the case $Q = \Sigma^{-1}$ and $C = \rho\Sigma(\rho \geq 1)$ (which includes the symmetric case where Q, Σ, and C are all multiples of the identity) the following result shows that $C^n(X)$ is smaller than $C^0(X)$ if $p \geq 2$.

THEOREM 3.2.7. *If* $Q = \Sigma^{-1}$, $C = \rho\Sigma(\rho \geq 1)$, *and* $p \geq 2$, *then* $\mathrm{tr}[Q\Sigma_n(X)]^{\frac{1}{2}} \leq \mathrm{tr}[Q\Sigma]^{\frac{1}{2}}$.

PROOF. Defining $a = u(\|X\|^2)/\rho$, it can be calculated that

$$\begin{aligned}\left[Q\Sigma_n(X)\right]^{\frac{1}{2}} &= \left[(1-u/\rho)I + wXX'/\rho^2\right]^{\frac{1}{2}} \\ &= (1-a)^{\frac{1}{2}}I + \Big[-(1-a)^{\frac{1}{2}} \\ &\quad + \left\{1 - a + (X'\Sigma^{-1}X)w/\rho^2\right\}^{\frac{1}{2}}\Big](X'\Sigma^{-1}X)\Sigma^{-1}XX'.\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{tr}\left[Q\Sigma_n(X)\right]^{\frac{1}{2}} &= (p-1)(1-a)^{\frac{1}{2}} + \left\{1 - a + (X'\Sigma^{-1}X)w/\rho^2\right\}^{\frac{1}{2}} \\ &\leq (p-1)(1-a)^{\frac{1}{2}} + (1+a)^{\frac{1}{2}} = h(a),\end{aligned}$$

the last step following from Lemma 3.1.1 (v). For $0 < a < 1$,

$$\frac{d}{da}h(a) = \frac{-(p-1)(1-a)^{-\frac{1}{2}}}{2} + \frac{(1+a)^{-\frac{1}{2}}}{2} < \frac{-(p-1)}{2} + \tfrac{1}{2} \leq 0.$$

Thus $h(a)$ is maximized at $h(0) = p = \mathrm{tr}(Q\Sigma)^{\frac{1}{2}}$ and the result follows. □

Numerical calculations will not be given for the above measure of loss since (at least for the symmetric situation) $\mathrm{tr}[Q\Sigma_n(X)]^{\frac{1}{2}}$ behaves like $p(1 - R^*)$.

Measure of size 3: sum of squares of semiaxes. The final measure of size is $L(X) = \text{tr}[Q\Sigma_n(X)]$, which is also the posterior expected loss. Clearly

$$L(X) = \text{tr}(Q\Sigma) - u(\|X\|^2)\text{tr}(Q\Sigma C^{-1}\Sigma) + w(\|X\|^2)X'C^{-1}\Sigma Q\Sigma C^{-1}X. \tag{3.23}$$

THEOREM 3.2.8. *$L(X) = \text{tr}[Q\Sigma_n(X)] \leqslant \text{tr}[Q\Sigma]$ for all X, if and only if $\text{tr}(\Sigma Q\Sigma C^{-1}) \geqslant 2\text{ch}_{\max}(\Sigma Q\Sigma C^{-1})$.*

PROOF. The "if" part follows immediately from (3.23) and the inequality

$$\begin{aligned} w(\|X\|^2)X'C^{-1}\Sigma Q\Sigma C^{-1}X &< \frac{2r_n(\|X\|^2)}{\|X\|^2}\left(\frac{X'C^{-1}\Sigma Q\Sigma C^{-1}X}{X'C^{-1}X}\right) \\ &\leqslant 2u(\|X\|^2)\text{ch}_{\max}(\Sigma Q\Sigma C^{-1}). \end{aligned}$$

The "only if" part is proved analagously to the "only if" part of Theorem 3.2.4. Choose X to be a multiple of the eigenvector corresponding to the largest characteristic root of $C^{-\frac{1}{2}}\Sigma Q\Sigma C^{-\frac{1}{2}}$, let $|X| \to \infty$ in (3.23), and use Lemma 2.1.1 (iii) and (ix) and Lemma 3.1.1 (iv). □

COROLLARY 3.2.9. *$C^*(X)$ has smaller size (measure 3) than $C^0(X)$ for all X, if and only if*

$$\text{tr}\left(\Sigma Q\left[I + A\Sigma^{-1}\right]^{-1}\right) \geqslant 2\text{ch}_{\max}\left(\Sigma Q\left[I + A\Sigma^{-1}\right]^{-1}\right).$$

PROOF. Obvious. □

An interesting observation can be made concerning the relationship between $R(\theta, \delta^n)$ and $E_\theta L(X)$.

THEOREM 3.2.10. *If $\text{tr}(\Sigma Q\Sigma C^{-1}) \geqslant (2n + 2)\text{ch}_{\max}(\Sigma Q\Sigma C^{-1})$, then $R(\theta, \delta^n) < E_\theta L(X)$ for all θ.*

PROOF. Integrating by parts as in Berger (1976c) (the technique was first noticed in the symmetric case by Stein (1973)) gives

$$\begin{aligned} R(\theta, \delta^n) = \text{tr}(Q\Sigma) + E_\theta\Bigg[&\frac{-2r_n}{\|X\|^2}\left\{\text{tr}(\Sigma Q\Sigma C^{-1}) - \frac{2X'C^{-1}\Sigma Q\Sigma C^{-1}X}{\|X\|^2}\right\} \\ &- \frac{4r_n'(\|X\|^2)X'C^{-1}\Sigma Q\Sigma C^{-1}X}{\|X\|^2} + \frac{r_n^2 X'C^{-1}\Sigma Q\Sigma C^{-1}X}{\|X\|^4}\Bigg]. \end{aligned}$$

Applying Lemma 3.1.1 (iv) and (3.23) to this expression gives

$$\begin{aligned} R(\theta, \delta^n) = E_\theta L(X) - E_\theta\Bigg[&\frac{r_n}{\|X\|^2}\left\{\text{tr}(\Sigma Q\Sigma C^{-1})\frac{(r_n + 2)X'C^{-1}\Sigma Q\Sigma C^{-1}X}{\|X\|^2}\right\} \\ &+ \frac{2r_n'(\|X\|^2)X'C^{-1}\Sigma Q\Sigma C^{-1}X}{\|X\|^2}\Bigg]. \end{aligned} \tag{3.24}$$

Since $r_n'(\|X\|^2) > 0$, $r_n(\|X\|^2) < 2n$, and

$$\frac{X'C^{-1}\Sigma Q\Sigma C^{-1}X}{\|X\|^2} \leqslant \mathrm{ch}_{\max}(\Sigma Q\Sigma C^{-1}),$$

the conclusion follows. □

COROLLARY 3.2.11. *If* $\Sigma Q\Sigma C^{-1} = \tau I$ *and* $n \leqslant (p-2)/2$, *then* $R(\theta, \delta^n) < E_\theta L(X)$ *for all* θ.

PROOF. Obvious. □

The above result was obtained for the situation $Q = \Sigma = C = I$ and $n = (p-2)/2$ by Morris (1977). Stein (1974) has related results in the symmetric situation.

Theorem 3.2.10 essentially says that, under the given condition, $L(X)$ is an overestimate (on the average) of the true expected loss for δ^n. In some sense, this indicates that the corresponding confidence sets $C^n(X)$ are larger than necessary, i.e., an error on the side of conservatism is being made. Note that for the symmetric situation, $C^*(X)$ satisfies the condition of Corollary 3.2.11.

Theorem 3.2.10 is somewhat puzzling in light of the fact that if $n > p/2$ (so that the priors g_n have finite mass) then

$$\int R(\theta, \delta^n) g_n(\theta)\, d\theta = \int [E_\theta L(X)]\, g_n(\theta)\, d\theta.$$

(Both sides are equal to the Bayes risk, up to the normalizing constant of g_n.) If $n \leqslant p/2$, the integrals above are infinite, making the result of Theorem 3.2.10 possible. The following result indicates what happens for $(p-2)/2 < n \leqslant p/2$. The proof will be omitted.

THEOREM 3.2.12. *If* $(p-2)/2 < n \leqslant p/2$, *then*

$$\int [R(\theta, \delta^n) - E_\theta L(X)]\, g_n(\theta)\, d\theta = 0.$$

Thus for $n > (p-2)/2$, $E_\theta L(X)$ is "on the average" equal to $R(\theta, \delta^n)$, and hence $L(X)$ is not an overestimate.

In conclusion, it can be noted that for the important symmetric problem (Q, Σ, and C multiples of the identity matrix), $C^n(X)$ is smaller than $C^0(X)$ for all measures of size considered and $p \geqslant 2$. Even for nonsymmetric problems, $C^n(X)$ tends to be smaller than $C^0(X)$ under quite weak conditions. For example, the conditions of Theorems 3.2.4, 3.2.7, and 3.2.8 tend to be considerably weaker than the minimax condition of Theorem 2.2.1.

3.3 *Probability of coverage of* C^n. The other major facet of the confidence region C^n which is of interest is its probability of covering the true value of θ, i.e.,

(3.25)

$$P_\theta(\theta \in C^n(X)) = \int_{\Omega_\theta} (2\pi)^{-p/2} (\det \Sigma)^{-\frac{1}{2}} \exp\{-(x-\theta)'\Sigma^{-1}(x-\theta)/2\}\, dx,$$

where $\Omega_\theta = \{x \in R^p : \theta \in C^n(x)\}$. Note that (3.25) is the usual (frequentist) probability of coverage, not a Bayesian probability.

Dealing with probability of coverage analytically is very difficult. It seems virtually impossible to theoretically obtain uniform (for all θ) dominance results as were obtained for size. Numerical studies of probability of coverage are very useful (and will be given), but they have the weakness in these high dimensional, many parameter settings of not being able to adequately cover the broad range of possible problems. When discussing $R(\theta, \delta^n)$ in Section 2.2, it was shown that a very useful analytical way of determining approximate risk behavior was to look at the "tail approximation" given in line (2.11). This suggests doing a similar thing for probability of coverage: obtain a large θ approximation for the probability of coverage of C^n. In looking at numerical studies, it will be seen that this approximation is a very good guide in determining the behavior of $P_\theta(\theta \in C^n(X))$.

THEOREM 3.3.1. *For the confidence ellipsoid*

$$C^n(X) = \{\theta : [\theta - \delta^n(X)]'\Sigma_n(X)^{-1}[\theta - \delta^n(X)] \leqslant k(\alpha)\},$$

$$P_\theta(\theta \in C^n(X)) = (1 - \alpha) + \frac{2n[k(\alpha)/2]^{p/2}\exp\{-k(\alpha)/2\}}{p\Gamma(p)\theta'C^{-1}\theta} \times \left\{\operatorname{tr}(\Sigma C^{-1}) - \frac{(2 + 2n)\theta'C^{-1}\Sigma C^{-1}\theta}{\theta'C^{-1}\theta}\right\} + 0(|\theta|^{-4}).$$

PROOF. Given in the Appendix. □

COROLLARY 3.3.2. *If* $2n < [\operatorname{tr}(\Sigma C^{-1})/\operatorname{ch}_{\max}(\Sigma C^{-1})] - 2$ *and* $0 < \alpha < 1$, *then* $P_\theta(\theta \in C_n(X)) > (1 - \alpha)$ *for large enough* $|\theta|$.

PROOF. Obvious from Theorem 3.3.1 and the fact that $(\theta'C^{-1}\Sigma C^{-1}\theta)/(\theta'C^{-1}\theta) \leqslant \operatorname{ch}_{\max}(\Sigma C^{-1})$. □

COROLLARY 3.3.3. *If* $C = \rho\Sigma$, *then*

$$P_\theta(\theta \in C^n(X)) = (1 - \alpha) + \frac{2n[k(\alpha)/2]^{p/2}\exp\{-k(\alpha)/2\}[p - (2 + 2n)]}{p\Gamma(p)\|\theta\|^2} + 0(|\theta|^{-4}).$$

PROOF. Obvious. □

Corollaries 3.3.2 and 3.3.3 show that $C^n(X)$ can possibly have probability of coverage greater than $(1 - \alpha)$ for all θ only if $n \leqslant (p - 2)/2$. Unfortunately, the estimator δ^n is inadmissible if $n < (p - 2)/2$. Thus to obtain a good estimator and a probability of coverage which is not seriously worse than $(1 - \alpha)$, it seems that the choice $n = (p - 2)/2$ should be made. In part, this is why δ^* and C^* were recommended with the choice $n = (p - 2)/2$.

For problems in which $C \neq \rho\Sigma$, Theorem 3.3.1 is useful in determining the directions in which $C^*(X)$ has greater or smaller probability of coverage than $(1 - \alpha)$. Indeed, since the error term in Theorem 3.3.1 is $0(|\theta|^{-4})$ while the second

term is $0(|\theta|^{-2})$, the approximation is fairly accurate for even moderate values of $|\theta|$. (Numerical studies showed this to be the case.)

As an example, the case $p = 6$, $\Sigma = I$, A diagonal with diagonal elements $\{.65, 3.5, 6.5, 9.5, 12.5, 45.5\}$, and $1 - \alpha = .90$ was considered. (This example was discussed in Section 3.2 with respect to the size of $C^*(X)$.) The probabilities of coverage, $P_\theta(\theta \in C^*(X))$, were calculated along the six axes and along the first quadrant diagonal. Table 4 gives the results for various values of $|\theta|$. (p_i stands for the probability of coverage along the ith axis ($1 \leqslant i \leqslant 6$), while p_d is for along the diagonal.) From Theorem 3.3.1 (with $C = \rho^*(\Sigma + A)$) it could be predicted that C^* would have a probability of coverage smaller than $(1 - \alpha)$ for large enough $|\theta|$ along the first two axes and the diagonal, and probability of coverage which eventually exceeds $(1 - \alpha)$ along the remaining axes. This behavior is exactly what is observed in Table 4. The coverage probability along the first axis appears particularly bad, but again the chance of being in this region is small (according to the prior information). (The probabilities in this and subsequent tables were computed by simulation, with M random vectors X^i being generated, and $p = P_\theta(\theta \in C^*(X))$ being estimated by

$$\hat{p} = M^{-1}(\text{the number of } i \text{ for which } \theta \in C^*(X^i)).$$

Clearly $\hat{p}$ is the sample proportion from a binomial sampling situation, so the standard deviation of $\hat{p}$ is

$$\sigma_{\hat{p}} = [p(1 - p)/M]^{\frac{1}{2}} \sim [\hat{p}(1 - \hat{p})/M]^{\frac{1}{2}}.$$

The entries in Table 4 were obtained with $M = 20{,}000$. Thus, for example, the standard deviation of p_3 when $|\theta| = 10$ is approximately $[(.908)(.092)/20{,}000]^{\frac{1}{2}} \sim .002$.)

For symmetric problems (or more generally those with $C = \rho\Sigma$), one would hope that $C^*(X)$ does have coverage probability greater than $(1 - \alpha)$. Unfortunately, Theorem 3.3.1 or Corollary 3.3.3 are no longer of any assistance, since $[p - (2 + 2n)] = 0$. It is thus the term of order $|\theta|^{-4}$ that is dominant, as the following theorem shows. (In one sense, this also helps justify the choice $n = (p - 2)/2$; the resulting confidence procedure is closer to the usual confidence procedure for "extreme" X.)

TABLE 4
Probabilities of coverage of $C^*(X)$.

$\lvert\theta\rvert$	0	1.0	1.5	2.0	3.0	4.0	5.0	6.0	10.0	15.0
p_1	.960	.935	.885	.820	.787	.821	.852	.870	.890	.895
p_2	.960	.957	.953	.947	.931	.915	.903	.897	.899	.899
p_3	.960	.959	.956	.955	.950	.943	.933	.923	.908	.904
p_4	.960	.959	.958	.957	.952	.948	.941	.935	.914	.905
p_5	.960	.960	.960	.959	.956	.953	.949	.944	.923	.910
p_6	.960	.960	.960	.960	.960	.959	.959	.958	.953	.943
p_d	.960	.956	.950	.941	.911	.873	.850	.848	.880	.892

THEOREM 3.3.4. *If* $C = \rho\Sigma$ *and* $n = (p-2)/2$, *then*

$$P_\theta(\theta \in C^n(X)) = (1-\alpha) + \frac{(p-2)[k(\alpha)/2]^{p/2}\exp\{-k(\alpha)/2\}}{4p\Gamma(p/2)(\theta'\Sigma^{-1}\theta)^2}\left\{4p(p-2) + \frac{k(\alpha)}{2(p+2)}[p^3+2p^2-32p-48]\right\} + 0(|\theta|^{-6}).$$

PROOF. Given in the Appendix. □

COROLLARY 3.3.5. *If* $A = \rho\Sigma$, *then for large* $|\theta|$, $P_\theta(\theta \in C^*(X)) > (1-\alpha)$ *providing*

(i) $0 < k(\alpha) < 1.212$ (i.e., $0 < (1-\alpha) < .25$) *when* $p = 3$;
(ii) $0 < k(\alpha) < 4.8$ (i.e., $0 < (1-\alpha) < .69$) *when* $p = 4$;
(iii) $0 < k(\alpha) < 25.45$ (i.e., $0 < (1-\alpha) < .9999$) *when* $p = 5$;
(iv) $0 < k(\alpha) < \infty$ (i.e., $0 < (1-\alpha) < 1$) *when* $p \geqslant 6$.

PROOF. The conditions on $k(\alpha)$ are simply those for which $\{4p(p-2) + k(\alpha)[2(p+2)]^{-1}[p^3+2p^2-32p-48]\} > 0$. Theorem 3.3.4 thus gives the desired result. □

For $p \geqslant 6$ (and virtually always for $p = 5$), the coverage probability of C^* is thus greater than $(1-\alpha)$ for large enough $|\theta|$. To determine the behavior for small $|\theta|$, numerical studies were conducted for $(1-\alpha) = .90$, $p = 4$, 6, and 12, and $C = 2\Sigma$. The results are given in Table 5 for various values of $(\theta'\Sigma^{-1}\theta)^{\frac{1}{2}}$. The coverage probability is never much worse than .90, and for small $(\theta'\Sigma^{-1}\theta)^{\frac{1}{2}}$ was considerably better. Note that as predicted by Corollary 3.3.5, the coverage probability fell below .90 for $p = 4$ and large $|\theta|$, but was above .90 for $p = 6$ and 12 and large $|\theta|$. The dip below .90 at $(\theta'\Sigma^{-1}\theta)^{\frac{1}{2}} = 8$ and $p = 12$ is somewhat surprising. The $0(|\theta|^{-4})$ term of Theorem 3.3.4 is apparently not yet dominant at this point. (The number, M, of random vectors used in the simulation was $M = 80{,}000$ for $p = 4$, $M = 60{,}000$ for $p = 6$, and $M = 40{,}000$ for $p = 12$. Note, therefore, that the standard deviation of the entry for $p = 12$ and $(\theta'\Sigma^{-1}\theta)^{\frac{1}{2}} = 8$ is $\sigma_{\hat{p}} \sim [(.895)(.105)/40{,}000]^{\frac{1}{2}} \sim .0015$.)

3.4 *Comparison with other confidence procedures.* As mentioned at the beginning of Section 3, several other multivariate confidence procedures have been proposed. For the most part they have been presented and studied only in the symmetric situation (Q, Σ, and A multiples of I), so the comparisons in this section

TABLE 5
Probabilities of coverage of $C^*(X)$.

p \ $(\theta'\Sigma^{-1}\theta)^{\frac{1}{2}}$	0	1	2	3	4	5	6	8	10	15
4	.971	.965	.945	.918	.902	.897	.898	.898	.898	.898
6	.993	.989	.976	.946	.916	.902	.900	.901	.901	.901
12	1.000	.999	.998	.988	.958	.921	.900	.895	.898	.900

will be restricted to that case. Along with $C^0(X)$ and $C^*(X)$, we will consider

$$C^{B-J}(X) = \{\theta : |\theta - \delta^*(X)|^2 \leqslant k(\alpha)\},$$

and

$$C^M(X) = \{\theta : [\theta - \delta^*(X)]^t \Sigma_M^{-1} [\theta - \delta^*(X)] \leqslant k(\alpha)\},$$

where $\Sigma_M(X)$ consists of the diagonal elements of $\Sigma^*(X)$. $C^{B-J}(X)$ is simply the usual confidence region centered at the improved estimator δ^* (in the spirit of the Brown (1966) and Joshi (1967) confidence sets). $C^M(X)$ is related to the region suggested by Morris (1977) in the symmetric situation. One difference is that his choice of C in the prior g_n was always $C = I$, not $C = \rho^*(\Sigma + A)$ as proposed here. (Some comments about both choices will be made.) The major difference is that confidence intervals, not confidence ellipsoids, are considered in Morris (1977). Hence overall probability of coverage is not the goal he pursues. To make meaningful comparisons, therefore, an ellipsoid using the variances in Morris (1977) is considered.

The other major proposed confidence regions, those of Stein (1962) and (1974) and Faith (1976), will not be discussed. Stein's regions are developed heuristically for large p and without modification are probably not suitable for small p. Faith's regions will not be considered for two reasons. First, as they are developed in a Bayesian fashion (though in the symmetric case), their performance is quite likely very similar to $C^*(X)$. On the other hand, they have a complicated shape and are hard to work with or evaluate. The relative simplicity of the other procedures makes them attractive.

In comparing sizes, only volume will be discussed, though similar conclusions hold for other measures of size. Since $C^0(X)$ and $C^{B-J}(X)$ have the same size, the results of Section 3.2 hold for both. (See in particular Corollary 3.2.5 and Table 2.) $C^*(X)$ clearly achieves a very significant reduction in size over $C^0(X)$ or $C^{B-J}(X)$.

Since $\Sigma_m(X)$ consists of the diagonal elements of $\Sigma^*(X)$, and $\Sigma^*(X)$ is positive definite, it follows that $\det[\Sigma^*(X)] \leqslant \det[\Sigma_M(X)]$. Hence $C^*(X)$ has smaller volume than $C^M(X)$ also. The difference in volume between $C^*(X)$ and $C^M(X)$ is, however, much less than that between $C^*(X)$ and $C^0(X)$. Indeed if X lies along an axis, it can be shown that $C^*(X)$ and $C^M(X)$ have the same volume. The volume ratio of $C^*(X)$ to $C^M(X)$ along the diagonals is given in Table 6 for various values of X, p, and C, when $\Sigma = I$. Morris always chooses $C = I$, while $C = 2I$ is more typical of $C = \rho^*(\Sigma + A)$ as suggested here.

TABLE 6
Volume ratio of $C^*(X)$ and $C^M(X)$ (along diagonals).

	$\lvert X\rvert$	1	2	3	4	5	6	7	9	15
p	$12(C = 2I)$	1.000	.999	.995	.981	.949	.921	.929	.970	.996
	$6(C = 2I)$	1.000	.997	.986	.971	.970	.980	.989	.996	.999
	$6(C = I)$	.990	.912	.848	.900	.955	.978	.989	.996	.999

To compare probabilities of coverage, numerical studies were conducted. Tables 7, 8 and 9 give results for $\Sigma = I$, $C = 2I$, and p equal 4, 6, and 12 respectively. Both C^* and C^{B-J} have probabilities of coverage which depend only on $|\theta|$. C^M, on the other hand, does not. Hence results for C^M are given for θ along the axes (C_a^M) and for θ along the diagonals (C_d^M). (Table 7 was done with $M = 80{,}000$, Table 8 with $M = 60{,}000$, and Table 9 with $M = 40{,}000$.)

Except for small $|\theta|$, C^{B-J} has better probability of coverage than C^*. On the other hand, C^* has significantly smaller volume than C^{B-J} (Table 2). In looking at the tradeoffs involved, the smaller size seems to more than offset the smaller probability of coverage. From an applications viewpoint, the confidence procedure C^* seems more appropriate also. It can be reported as a $(1 - \alpha)$ confidence region and will have a definitely reportable smaller size than $C^0(X)$. $C^{B-J}(X)$, on the other hand, has the same size as $C^0(X)$ and can also only be reported as a $(1 - \alpha)$ confidence region. The gains in probability of coverage if the true θ happens to be small are hard to report. C^{B-J} would, in a conservative sense, be more competitive in nonsymmetric situations, since its probability of coverage would be less likely to drop below $(1 - \alpha)$ than would the probability of coverage of C^*.

C^* and C^M have very similar probabilities of coverage. (Note that both are calculated at $C = 2I$ for comparison purposes. The choice $C = I$ gives less attractive results for both regions.) C^M is better along the axes, while C^* is better along the diagonals. The smaller size of $C^*(X)$ and its greater simplicity in

TABLE 7
Probabilities of coverage ($p = 4$).

$\|\theta\|$	0	1	2	3	4	5	6	8	10	15
C^*	.971	.965	.945	.918	.902	.897	.897	.898	.898	.899
C^{B-J}	.970	.967	.959	.945	.928	.914	.908	.903	.902	.900
C_a^M	.959	.954	.940	.921	.909	.904	.902	.900	.899	.899
C_d^M	.959	.954	.938	.916	.900	.895	.895	.898	.898	.899

TABLE 8
Probabilities of coverage ($p = 6$).

$\|\theta\|$	0	1	2	3	4	5	6	8	10	15
C^*	.993	.989	.976	.946	.916	.902	.900	.901	.901	.901
C^{B-J}	.990	.989	.985	.976	.962	.944	.930	.917	.912	.906
C_a^M	.981	.977	.965	.945	.927	.917	.912	.907	.904	.902
C_d^M	.981	.977	.964	.939	.912	.898	.895	.899	.901	.901

TABLE 9
Probabilities of coverage ($p = 12$).

$\|\theta\|$	0	1	2	3	4	5	6	8	10	15
C^*	1.000	.999	.998	.988	.958	.921	.900	.895	.898	.900
C^{B-J}	.999	.999	.999	.997	.995	.990	.978	.952	.936	.917
C_a^M	.995	.994	.991	.980	.961	.942	.933	.922	.912	.903
C_d^M	.995	.994	.991	.979	.951	.916	.893	.888	.894	.898

nonsymmetric situations make it attractive. Both procedures, however, should do quite well.

4. Incorporation of prior information. As mentioned in Section 1, prior input in the form of a prior mean vector μ and a prior covariance matrix A is envisaged. The use of A in the development of δ^* and C^* has already been discussed. To use μ, the estimator and confidence region should be centered at μ. Thus

$$\delta^*(X) = X - \frac{r^*(\|X-\mu\|^2)\Sigma(\Sigma+A)^{-1}(X-\mu)}{(X-\mu)'(\Sigma+A)^{-1}(X-\mu)} \tag{4.1}$$

is the recommended estimator. The definition of $\Sigma^*(X)$ is unchanged, except that X should be replaced by $X-\mu$ in all expressions. This shift changes none of the properties or results established in Sections 2 and 3.

It is sometimes desirable to choose C^{-1} to be singular. The only change which should then be made in the definitions of δ^n and C^n is to choose $n = ([\text{rank } C^{-1}] - 2)/2$ instead of $n = (p-2)/2$. The rank of C^{-1} is the effective dimensionality of the problem. This can be seen by diagonalizing Σ and C, and then noting that δ^n and Σ_n are the generalized Bayes estimator and posterior covariance matrix for a subproblem of rank C^{-1}. Thus all the results of Section 2 and 3 (with the exception of the admissibility of δ^n) hold with p replaced by [rank C^{-1}].

The reason for choosing C^{-1} singular would be that in some directions there is no prior information whatsoever (or alternatively, that A has infinite characteristic roots in these directions). The corresponding coordinates are then effectively excluded from the correction terms of the estimator δ^n and the posterior covariance matrix Σ_n.

An example of the use of singular C^{-1} is when shrinkage towards the common mean $\bar{X} = \sum_{i=1}^p X_i/p$ is desired. Defining (1) as the matrix of all ones, $\bar{1}$ as the column vector of ones, and letting $C^{-1} = I - (1/p)(1)$, it is easy to check that (for $\Sigma = I$)

$$\delta^*(X) = X - \frac{r_n(|X-\bar{X}\bar{1}|^2)(X-\bar{X}\bar{1})}{|X-\bar{X}\bar{1}|^2}, \tag{4.2}$$

an estimator which shrinks towards the common mean. Note that C^{-1} has rank $(p-1)$, so $n = (p-3)/2$ is the appropriate choice of n. Choosing C^{-1} as above is essentially a statement that the θ_i are felt to be similar (or their priors have a common mean or their prior is exchangeable), but that the common value that the θ_i are thought to be near is totally unknown. This last assumption seems somewhat extreme intuitively, and the following Bayesian considerations suggest a reasonable alternative.

Assume that the θ_i are thought to be a random sample from a normal distribution with mean θ_0 and variance σ^2. (It is convenient to develop μ and A through the assumption of normal priors due to the resulting ease in manipulation.) It is often assumed that θ_0 also has a normal distribution with mean μ_0 and variance σ_0^2. (This

problem is discussed in Lindley and Smith (1972), where earlier works on the model are also referenced.) As pointed out in Lindley and Smith (1972), this two stage prior is equivalent to assuming that θ has a p-variate normal distribution with mean $\mu = \mu_0\bar{1}$ and covariance matrix $A = (\sigma^2 I + \sigma_0^2(1))$. The common Bayesian technique is to use the linear Bayes estimator, letting $\sigma_0^2 \to \infty$. (The prior information at the second stage is deemed vague, so taking σ_0^2 to infinity results in a more robust estimator.) Due to the fact that δ^* is already quite robust, however, the best guesses μ and A can safely be used directly in δ^*. There is no need to let $\sigma_0^2 \to \infty$. Note that at the two extremes, letting $\sigma_0^2 \to \infty$ in δ^* would result in (4.2) (providing $\Sigma = I$ and $n = (p - 3)/2$ were used), while choosing $\sigma_0^2 = 0$ would simply result in an estimator shrinking towards the believed mean $\mu_0\bar{1}$.

As another example of the use of prior information, assume that the linear restriction

$$H(\theta - \theta_0) = 0$$

is thought to hold, where θ_0 is a p vector and H is a $(k \times p)$ matrix $(k \leqslant p)$ of rank $k (k \geqslant 3)$. Suppose a $(k \times k)$ positive definite matrix A is also determined, where A reflects the accuracy with which the linear restrictions are believed to hold. (A can be thought of as the estimated covariance matrix of the prior distribution of $H(\theta - \theta_0)$.)

The appropriate version of δ^* for this situation is

$$\delta^*(X) = X - \frac{r_n\big([X - \theta_0]'C[X - \theta_0]\big)\Sigma C(X - \theta_0)}{[X - \theta_0]'C[X - \theta_0]},$$

where $n = (k - 2)/2$ and

$$C = \frac{k}{(k - 2)} H'(H\Sigma H' + A)^{-1}H.$$

For C^*, the confidence region, Σ^* should be chosen to be Σ_n with X replaced by $[X - \theta_0]$ and n and C as above.

The rationale for the above choices arises from an analysis in which the null space of H (i.e., $N = \{\theta : H\theta = 0\}$) is given a prior distribution in which the variances are sent to infinity. This is a mathematical way of saying that there is no prior information about N (since no restrictions were specified for this space). The projection of X upon N should thus not be used in the correction terms of δ^* and Σ^* (and indeed it is not). The details of the analysis will be omitted. Note that k is used in place of p in the above estimator, since this is really the dimensionality of the prior information. It should also be emphasized that due to the robustness of δ^* and C^*, even quite uncertain linear restrictions can be usefully incorporated.

5. Unknown variance. In applications, it is important to consider the situation in which the covariance matrix of X is unknown. Attention will be restricted to the case where the covariance matrix is of the form $\sigma^2\Sigma$, Σ known but σ^2 unknown. (This is the common situation in regression problems.)

There are two possible approaches to dealing with this problem. The first is simply to replace σ^2 by an estimate in δ^* and C^* (with appropriate changes to $k(\alpha)$ in C^*). The second is to place a prior distribution upon σ^2 (in addition to θ), and to develop δ^* and C^* in terms of the combined prior information.

The second approach was used by Strawderman (1973) for the case $\Sigma = I$ (in g_n). (M. E. Bock (personal communication) has been able to explicitly evaluate the resulting estimator.) Unfortunately, the resulting estimator is extremely complex, even in this simple setting. The problems of constructing such an estimator for the nonsymmetric setting, and then of meaningfully analyzing it, seem considerable. Indeed the priors placed on σ^2 are rather unintuitive, and whether or not they have a beneficial effect on the estimator is unclear. It should be emphasized that δ^* and C^* were developed in a Bayesian fashion mainly because it appeared necessary to use prior information in the choice of a competitor to $\delta^0(X) = X$. There is no such compelling reason to use prior information on σ^2 in constructing δ^*. The approach that will be adopted is thus the first approach, merely replacing σ^2 by an estimate in σ^* and C^*. (Of course, if significant prior information about σ^2 were available, it would be reasonable to use this in the estimation of σ^2, but this could be left up to individual taste. Note that the effect upon δ^* would probably be slight, in the sense that δ^* would still probably be very robust, but the effect of wrong prior information about σ^2 on C^* could be considerable.)

When σ^2 is unknown, assume a random variable S^2 is observable (independent of X), where S^2/σ^2 has a chi-square distribution with m degrees of freedom. A suitable estimate of σ^2 for use in δ^* and C^* is $S^2/(m+2)$. Thus $[S^2/(m+2)]\Sigma$ and $C = \rho^*\{[S^2/(m+2)]\Sigma + A\}$ should be used in δ^* and C^* in place of the previous Σ and C. A reason for choosing $S^2/(m+2)$ as the estimator of σ^2 is that it is the natural estimator for certain minimax results. The following theorem is an example. For convenience, define

$$G(Q, \Sigma, A) = \lim_{t\to 0} \frac{\operatorname{tr}\left[(\Sigma Q \Sigma(t\Sigma + A)^{-1}\right]}{\operatorname{ch}_{\max}\left[\Sigma Q \Sigma(t\Sigma + A)^{-1}\right]}. \tag{5.1}$$

Note that if A is nonsingular, then $G(Q, \Sigma, A) = \operatorname{tr}(\Sigma Q \Sigma A^{-1})/\operatorname{ch}_{\max}(\Sigma Q \Sigma A^{-1})$.

THEOREM 5.1. *Assume Q^{-1}, Σ, and A are simultaneously diagonalizable, with resulting diagonal elements $\{q_i^{-1}\}$, $\{d_i\}$, and $\{A_i\}$, satisfying for $1 \leqslant i, j \leqslant p$*

$$\left[(A_j/d_j) - (A_i/d_i)\right](d_i q_i - d_j q_j) \geqslant 0. \tag{5.2}$$

Let $C = \rho(S^2)([S^2/(m+2)]\Sigma + A)$, where ρ is nondecreasing in S^2. Then

$$\delta^n(X, S^2) = \left(I - \frac{r_n(\|X\|^2)S^2\Sigma C^{-1}}{\|X\|^2(m+2)}\right)X$$

has smaller risk than $\delta^0(X) = X$, providing $n \leqslant G(Q, \Sigma, A) - 2$.

PROOF. Integrating by parts as in Theorem 3.2.10 gives

$$(5.3)\quad R(\vartheta, \sigma^2, \delta^n) - R(\theta, \sigma^2, \delta^0)$$

$$= E_{\theta,\sigma^2}\left[\frac{-2r_n S^2\sigma^2}{\|X\|^2(m+2)}\left\{\operatorname{tr}(\Sigma Q \Sigma C^{-1}) - \frac{2(X^tC^{-1}\Sigma Q\Sigma C^{-1}X)}{\|X\|^2}\right\}\right.$$

$$\left. - \frac{4\sigma^2 r_n'(\|X\|^2)S^2(X^tC^{-1}\Sigma Q\Sigma C^{-1}X)}{\|X\|^2(m+2)}\right] + E_{\theta,\sigma^2}\left[\frac{r_n^2(X^tC^{-1}\Sigma Q\Sigma C^{-1}X)S^4}{\|X\|^4(m+2)^2}\right].$$

Efron and Morris (1976) proved the identity

$$(5.4)\quad E_{\sigma^2}[g(S^2)S^2] = \sigma^2 m E_{\sigma^2}[g(S^2)] + 2\sigma^2 E_{\sigma^2}[S^2 g'(S^2)],$$

for any differentiable function g for which the expectations exist. Defining

$$h(S^2) = (X^tC^{-1}\Sigma Q\Sigma C^{-1}X)/\|X\|^4,$$

(recall C is a function of S^2) and setting

$$g(S^2) = r_n^2(\|X\|^2)S^2h(S^2)/(m+2)^2,$$

it follows from (5.4) that

$$(5.5)\quad E\left[\frac{r_n^2S^4h(S^2)}{(m+2)^2}\right] = \sigma^2 m E\left[\frac{r_n^2S^2h(S^2)}{(m+2)^2}\right]$$

$$+2\sigma^2 E\left[S^2\left\{\frac{2r_n\left(\frac{d}{dS^2}r_n(\|X\|^2)\right)S^2h(S^2)}{(m+2)^2} + \frac{r_n^2h(S^2)}{(m+2)^2} + \frac{r_n^2S^2h'(S^2)}{(m+2)^2}\right\}\right].$$

From the definition of C and the assumption that ρ is nondecreasing in S^2, it is clear that $\|X\|^2 = X^tC^{-1}X$ is nonincreasing in S^2. Hence

$$(5.6)\quad \frac{d}{dS^2}r_n(\|X\|^2) \leqslant 0.$$

Defining $Y = TX$, where T is a $(p \times p)$ matrix such that $TQ^{-1}T^t$, $T\Sigma T^t$, and TAT^t are all diagonal matrices. It is easy to check that

$$h(S^2) = \left[\Sigma_{i=1}^p Y_i^2 q_i d_i \Big/ \left(\frac{S^2}{m+2} + \frac{A_i}{d_i}\right)^2\right] \Big/ \left[\Sigma_{i=1}^p Y_i^2 \Big/ \left(\frac{S^2}{m+2} + \frac{A_i}{d_i}\right)\right]^2.$$

Defining $b_i = [S^2/(m+2)] + [A_i/d_i]$, a calculation gives that

$$\begin{aligned} h'(S^2) &= \frac{-2}{(m+2)\left(\Sigma_{i=1}^p Y_i^2/b_i\right)^3}\left\{\left(\Sigma_{i=1}^p Y_i^2 q_i d_i/b_i^3\right)\left(\Sigma_{j=1}^p Y_j^2/b_j\right)\right. \\ &\quad \left. - \left(\Sigma_{i=1}^p Y_i^2 q_i d_i/b_i^2\right)\left(\Sigma_{j=1}^p Y_j^2/b_j^2\right)\right\} \\ &= \frac{-2}{(m+2)\left(\Sigma_{i=1}^p Y_i^2/b_i\right)^3}\left\{\Sigma_{i=1}^p \Sigma_{j=1}^p \frac{Y_i^2 Y_j^2}{b_i^3 b_j^3}\left[b_j^2 q_i d_i - b_i b_j q_i d_i\right]\right\} \\ &= \frac{-2}{(m+2)\left(\Sigma_{i=1}^p Y_i^2/b_i\right)^3} \\ &\quad \times\left\{\Sigma_{i=1}^p \Sigma_{j=1}^{(i-1)} \frac{Y_i^2 Y_j^2}{b_i^3 b_j^3}\left[b_j^2 q_i d_i - b_i b_j q_i d_i + b_i^2 q_j d_j - b_i b_j q_j d_j\right]\right\} \\ &= \frac{-2}{(m+2)\left(\Sigma_{i=1}^p Y_i^2/b_i\right)^3}\left\{\Sigma_{i=1}^p \Sigma_{j=1}^{(i-1)} \frac{Y_i^2 Y_j^2}{b_i^3 b_j^3}\left[(b_j - b_i)(b_j q_i d_i - b_i q_j d_j)\right]\right\}. \end{aligned}$$

Using the definition of b_i, a calculation gives that

$$(5.7)\qquad (b_j - b_i)(b_j q_i d_i - b_i q_j d_j) = \left(\frac{A_j}{d_j} - \frac{A_i}{d_i}\right)(q_i d_i - q_j d_j)\frac{S^2}{(m+2)} + \left(\frac{A_j}{d_j} - \frac{A_i}{d_i}\right)\left(\frac{A_j q_i d_i}{d_j} - \frac{A_i q_j d_j}{d_i}\right).$$

The first term on the right-hand side of (5.7) is nonnegative by (5.2). The second term is nonnegative since (5.2) implies that the two factors of the second term have the same sign (or one is zero). It can thus be concluded that $h'(S^2) \leqslant 0$. Together with (5.5) and (5.6), this implies that

$$(5.8)\qquad E_{\sigma^2}\left[\frac{r_n^2 S^4 h(S^2)}{(m+2)^2}\right] \leqslant \sigma^2 E_{\sigma^2}\left[\frac{r_n^2 S^2 h(S^2)}{(m+2)}\right].$$

Using (5.3), (5.8) and the facts that $r_n'(\|X\|^2) > 0$, $r_n(\|X\|^2) < 2n$, and $(X'C^{-1}\Sigma Q \Sigma C^{-1}X)/\|X\|^2 \leqslant \mathrm{ch}_{\max}(\Sigma Q \Sigma C^{-1})$, it follows that

$$(5.9)\qquad R(\theta, \sigma^2, \delta^n) - R(\theta, \sigma^2, \delta^0) < E_{\theta,\sigma^2}\left[\frac{-2 r_n S^2 \sigma^2 \mathrm{ch}_{\max}(\Sigma Q \Sigma C^{-1})}{\|X\|^2(m+2)} \times\left\{\frac{\mathrm{tr}(\Sigma Q \Sigma C^{-1})}{\mathrm{ch}_{\max}(\Sigma Q \Sigma C^{-1})} - (2+n)\right\}\right].$$

Clearly

$$(5.10)\qquad \frac{\mathrm{tr}(\Sigma Q \Sigma C^{-1})}{\mathrm{ch}_{\max}(\Sigma Q \Sigma C^{-1})} = \Sigma_{i=1}^p \frac{(d_i q_i/b_i)}{\max_j\{d_j q_j/b_j\}} = 1 + \Sigma_{i \neq k} \frac{b_k d_i q_i}{b_i d_k q_k},$$

where k is the coordinate at which the maximum is attained. But if $d_k q_k / b_k \geqslant d_i q_i / b_i$ for $i \neq k$, then for (5.2) to hold it must be true that $b_k \leqslant b_i$, or equivalently that $A_k / d_k \leqslant A_i / d_i$. Hence

$$\frac{b_k}{b_i} = \frac{S^2/(m+2) + A_k/d_k}{S^2/(m+2) + A_i/d_i}$$

is nondecreasing in S^2. It follows that (5.10) is minimized at $S^2 = 0$, attaining the value $G(Q, \Sigma, A)$. Together with (5.9), this establishes that

$$R(\theta, \sigma^2, \delta^n) - R(\theta, \sigma^2, \delta^0) < -E_{\theta, \sigma^2}\left[\frac{2r_n S^2 \sigma^2 \mathrm{ch}_{\max}(\Sigma Q \Sigma C^{-1})}{\|X\|^2(m+2)}\{G(Q, \Sigma, A) - (2+n)\}\right].$$

By the condition on n, the argument of the expectation is positive and the conclusion follows. □

Two special cases of interest are given in the following corollaries.

COROLLARY 5.2. *If $Q = \tau\Sigma^{-1}$, then δ^n (chosen as in Theorem 5.1) has smaller risk than δ^0 if $n \leqslant G(Q, \Sigma, A) - 2$. (If A is nonsingular, $G(Q, \Sigma, A) = \mathrm{tr}(\Sigma A^{-1})/\mathrm{ch}_{\max}(\Sigma A^{-1})$.)*

PROOF. Clearly $Q^{-1} = \tau^{-1}\Sigma$, Σ, and A are simultaneously diagonalizable. Also, $d_i q_i = \tau$ for all i, so that (5.2) is satisfied. The conclusion follows from Theorem 5.1. □

Note that $Q = \Sigma^{-1}$ is an often considered choice of Q, as it gives rise to a loss which is invariant and, more importantly, is the natural loss for the prediction problem of linear regression. (Predict the value of a future observation arising from the same design matrix.)

COROLLARY 5.3. *If $A = \tau\Sigma$, then δ^n (chosen as in Theorem 5.1) has smaller risk than δ^0 if $n \leqslant [\mathrm{tr}(\Sigma Q)/\mathrm{ch}_{\max}(\Sigma Q)] - 2$.*

PROOF. Q^{-1}, Σ, and $A = \tau\Sigma$ are all simultaneously diagonalizable and $A_i/d_i = \tau$ for all i. Hence (5.2) is satisfied and Theorem 5.1 can be applied to give the desired result. □

The estimator δ^n is undoubtedly uniformly better than δ^0 in situations where (5.2) is not satisfied, but a more general proof was not found. Note, in any case, from the statement of Theorem 5.1, that m (the degrees of freedom of S^2) is not part of the condition of the theorem. This is why $S^2/(m+2)$ seemed the natural estimator of σ^2 to use in δ^*.

COROLLARY 5.4. *If Q^{-1}, Σ, and A are simultaneously diagonalizable and satisfy (5.1), then δ^* has smaller risk than δ^0 if $p \leqslant 2G(Q, \Sigma, A) - 2$.*

PROOF. Obvious from Theorem 5.1. □

The estimation of σ^2 does not affect the robust Bayesian properties of δ^* appreciably, so numerical studies (such as Table 1) will not be presented for this case.

When estimating σ^2 by $S^2/(m+2)$, the appropriate definition of the confidence region C^n is now

$$C^n(X, S^2) = \left\{\theta : \left[\theta - \delta^n(X, S^2)\right]' \Sigma_n(X, S^2)^{-1} \left[\theta - \delta^n(X, S^2)\right] \leqslant k(\alpha)\right\},$$

where δ^n and Σ_n are defined as earlier with Σ replaced by $[S^2/(m+2)]\Sigma$ and $k(\alpha) = (m+2)pF_{p,m}(1-\alpha)/m$, $F_{p,m}(1-\alpha)$ being the $100(1-\alpha)$th percentile of the F distribution with p and m degrees of freedom. Note that the usual confidence ellipsoid when σ^2 is unknown is

$$C^0(X, S^2) = \left\{\theta : (\theta - X)'\Sigma^{-1}(\theta - X) \leqslant \left(\frac{S^2}{m}\right) pF_{p,m}(1-\alpha)\right\}.$$

In considering the size of $C^n(X, S^2)$, the results in Section 3.2 all hold with Σ replaced by $[S^2/(m+2)]\Sigma$. The conditions of the theorems then depend on S^2, however, at least for $C^*(X, S^2)$ which chooses $C = \rho^*([S^2/(m+2)]\Sigma + A)$. Global theorems can be developed, if desired, an example of which is the following.

THEOREM 5.5. *$C^*(X, S^2)$ has smaller volume than $C^0(X, S^2)$ for all X and S^2 if $G(\Sigma^{-1}, \Sigma, A) \geqslant 2$. ($G$ is defined in* (5.1).)

PROOF. By Corollary 3.2.6, it is only necessary to show that for all $S^2 > 0$,

$$\frac{\operatorname{tr}\left[I + A\Sigma^{-1}/\left\{S^2/(m+2)\right\}\right]^{-1}}{\operatorname{ch}_{\max}\left[I + A\Sigma^{-1}/\left\{S^2/(m+2)\right\}\right]^{-1}} \geqslant 2. \tag{5.11}$$

Letting $\{b_i\}$ denote the roots of $\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}$, it is clear that (5.11) can be rewritten

$$\Sigma_{i=1}^{p} \frac{S^2/(m+2) + \min\{b_j\}}{S^2/(m+2) + b_i} \geqslant 2. \tag{5.12}$$

The expression on the left-hand side of (5.12) is clearly minimized as $S^2 \to 0$. But the limit as $S^2 \to 0$ is nothing but $G(\Sigma^{-1}, \Sigma, A)$, and the conclusion follows. □

Tables 2 and 3 still give typical volume ratios of $C^*(X, S^2)$ to $C^0(X, S^2)$ (when $S^2 = m + 2$ for example).

Numerical studies were performed to investigate $P_\theta(\theta \in C^*(X, S^2))$. The results turned out to be very similar to those in Section 3.3 and so will not be presented. The general tendency was for $C^*(X, S^2)$ to have (for all θ) probability of coverage closer to $(1-\alpha)$ than in the corresponding situation with known variance. (Essentially, the additional randomization over S^2 smooths out the more extreme probabilities of coverage for the known variance case.)

In conclusion, it appears that estimation of σ^2 does not reduce the benefits of using δ^* and C^*.

6. Generalizations and comments.

1. An interesting feature of δ^n can be observed using Lemma 2.1.1 (v), namely that

$$\lim_{n\to\infty}\delta^n(X) = (I - \Sigma C^{-1})X.$$

Hence if $C = (\Sigma + A)$, the limiting estimator is the optimal linear Bayes estimator. Larger than recommended values of n may, therefore, be useful when accurate information about the tail of the prior is available. For example, if it is thought that the prior has a normal tail, so that $(I - \Sigma(\Sigma + A)^{-1})X$ is being considered for use, it might pay instead to use δ^n with a large value of n. The resulting estimator will behave similarly to the linear estimator except that it will be more robust with respect to inaccurate prior information. Of course, the larger n is, the less robust δ^n will be.

2. More general classes of priors can be considered. Indeed it can be checked that $g_n(\theta)$ and $r_n(v)$ in Section 2 can be replaced by

$$g_n^*(\theta) = \int_0^1 [\det B(\lambda)]^{-\frac{1}{2}} \exp\{-\theta' B(\lambda)^{-1}\theta/2\}\, d\mu(\lambda),$$

$$r_n^*(v) = \frac{v\int_0^1 h(\lambda)^{(1+p/2)}\exp\{-vh(\lambda)/2\}\, d\mu(\lambda)}{\int_0^1 h(\lambda)^{p/2}\exp\{-vh(\lambda)/2\}\, d\mu(\lambda)},$$

where $B(\lambda) = [C/h(\lambda)] - \Sigma$ and $0 < h(\lambda) \leqslant 1$ for $0 < \lambda \leqslant 1$. For a wide variety of h and μ, $r_n^*(v)$ can be explicitly evaluated. For example, choosing $h(\lambda) = \lambda$ and $d\mu(\lambda) = I_{(\varepsilon,1)}(\lambda)\lambda^{(n-1-p/2)}\, d\lambda$ results in a calculable estimator which behaves like $\delta^n(X)$ for small and moderate values of $\|X\|^2$ (the region depending on ε), but behaves like a linear Bayes estimator for large values of $\|X\|^2$. As another example, if μ is chosen to put unit mass at a particular point, the resulting prior is simply a normal prior. The general class is clearly very rich. (See Efron and Morris (1973a) and Faith (1976) for related classes of estimators in the symmetric situation.) Of the various estimators we considered which arose from priors in this class, δ^n seemed the most attractive. Hence attention was restricted to δ^n.

3. Unfortunately, a problem does arise with δ^* (and with other estimators of the form (1.1)). The estimator definitely performs best when all coordinates are similar or can be transformed so they are similar. (More precisely, this occurs when $[\Sigma Q \Sigma(\Sigma + A)^{-1}]$ is close to a multiple of the identity.) Thus if, for example, there were two groups of similar coordinates, the groups being quite different from each other, it would probably pay to separately estimate each group. In terms of a prior, this could be interpreted as saying the θ_i's should not be forced to act dependently (as in g_n), but should be separated into similar groups, with independent prior distributions on each group. The question is—when and how should this separation take place? (Efron and Morris (1973b) give an interesting discussion of the problem in the symmetric situation.)

4. All results in the paper have been for quadratic loss, due to the relative ease of calculation. Numerical studies (such as in Berger (1976b)) have indicated

however, that estimators like δ^* tend to have risks which are quite robust with respect to the functional form (or more precisely the tail) of the loss. See Berger (1976b) for further discussion.

5. The well-known relationship between confidence sets and testing of hypotheses, indicates that in some sense the usual multivariate tests of a point null hypothesis can be improved upon by using as an acceptance region $A(\theta) = \{x : \theta \in C^*(x)\}$. Of course the usual multivariate tests are admissible when error probabilities are the criteria of evaluation, so no uniform improvement is possible. The improvement that could be obtained would thus be with respect to Bayesian criteria. The point is that the use of acceptance regions based on $C^*(X)$ could result in robust Bayesian tests. A complete discussion of this is outside the scope of this paper.

6. The procedures δ^n and C^n can also be used in one and two dimensions. Though their classical (frequentist) properties will not be as appealing in such low dimensional settings, their performance as robust Bayes procedures will still be extremely satisfactory. Since $n = (p-2)/2$ can no longer be chosen, $n = \frac{1}{2}$ seems appropriate. The procedures δ^* and C^* are recommended, with this change, for situations in which significant prior information is available.

APPENDIX

PROOF OF THEOREM 3.3.1. The proof is related to the theoretical proof in Brown (1966) of the inadmissibility of the usual confidence sets.

For simplicity, assume that $\Sigma = I$. (This can be assumed without loss of generality, as is seen by considering the linearly transformed problem $Z = \Sigma^{-\frac{1}{2}}X$, $\eta = \Sigma^{-\frac{1}{2}}\theta$, and $C' = \Sigma^{-\frac{1}{2}}C\Sigma^{-\frac{1}{2}}$.)

Define

$$\Omega_\theta = \{x \in R^p : \theta \in C^n(x)\} = \{x : [\theta - \delta^n(x)]'\Sigma_n(x)^{-1}[\theta - \delta^n(x)] \leqslant k(\alpha)\}.$$

Using Lemmas 2.1.1, 3.1.1, and 3.1.2, a fairly lengthy Taylors series argument shows that if $x \in \Omega_\theta$, then

$$\text{(A.1)}\quad \Sigma_n(x)^{-\frac{1}{2}}(\theta - \delta^n(x)) = (\theta - x) + 2n\|\theta\|^{-2}C^{-1}\theta - n\|\theta\|^{-2}C^{-1}(\theta - x) + 2n\|\theta\|^{-4}[\theta'C^{-1}(\theta - x)]C^{-1}\theta + 0(|\theta|^{-3}).$$

Define

$$\text{(A.2)}\quad Y = (\theta - X) + 2n\|\theta\|^{-2}C^{-1}\theta - n\|\theta\|^{-2}C^{-1}(\theta - X) + 2n\|\theta\|^{-4}[\theta'C^{-1}(\theta - X)]C^{-1}\theta.$$

Letting J denote the Jacobian of this transformation from X to Y, a calculation shows that

$$\text{(A.3)}\quad |\det J|^{-1} = 1 + n\|\theta\|^{-2}\text{tr}(C^{-1}) + 2n\|\theta\|^{-4}\theta'C^{-2}\theta + 0(|\theta|^{-4}).$$

Observe from (A.1) and (A.2) that if $x \in \Omega_\theta$, then

$$[\theta - \delta^n(x)]'\Sigma_n(x)^{-1}[\theta - \delta^n(x)] = |y|^2 + 0(|\theta|^{-3}) \tag{A.4}$$

and

$$(\theta - x) = y - 2n\|\theta\|^{-2}C^{-1}\theta + n\|\theta\|^{-2}C^{-1}y - 2n\|\theta\|^{-4}(\theta'C^{-1}y)C^{-1}\theta + 0(|\theta|^{-3}). \tag{A.5}$$

From (A.5) it follows that if $x \in \Omega_\theta$, then

$$\begin{aligned}\exp\{-|\theta - x|^2/2\} = \exp\{-|y|^2/2\}\big[1 &+ 2n\|\theta\|^{-2}y'C^{-1}\theta - n\|\theta\|^{-2}y'C^{-1}y \\ &- 2n^2\|\theta\|^{-4}\theta'C^{-2}\theta + 2n\|\theta\|^{-4}(\theta'C^{-1}y)^2 \\ &+ 2n^2\|\theta\|^{-4}(y'C^{-1}\theta)^2 + 0(|\theta|^{-3})\big].\end{aligned} \tag{A.6}$$

Defining $S(k) = \{y \in R^p : |y|^2 < k(\alpha)\}$, it follows from (A.3), (A.4) and (A.6) that a change of variables from X to Y gives

$$\begin{aligned}P_\theta(\theta \in C_n(X)) &= \int_{\Omega_\theta}(2\pi)^{-p/2}\exp\{-|x - \theta|^2/2\}\,dx \\ &= \int_{S(k)}(2\pi)^{-p/2}\exp\{-|y|^2/2\}\big\{1 + 2n\|\theta\|^{-2}y'C^{-1}\theta + n\|\theta\|^{-2}\mathrm{tr}(C^{-1}) \\ &\quad - n\|\theta\|^{-2}\|y\|^2 - (2n^2 + 2n)\|\theta\|^{-4}\big[\theta'C^{-2}\theta - (\theta'C^{-1}y)^2\big] + 0(|\theta|^{-3})\big\}\,dy.\end{aligned} \tag{A.7}$$

Define

$$\begin{aligned}h(\alpha) &= (2\pi)^{-p/2}\int_{S(k)}y_i^2\exp\{-|y|^2/2\}\,dy \\ &= (2\pi)^{-p/2}\int_{S(k)}p^{-1}|y|^2\exp\{-|y|^2/2\}\,dy \\ &= (1 - \alpha) - \frac{[k(\alpha)]^{p/2}\exp\{-k(\alpha)/2\}}{2^{(p-2)/2}\Gamma(p/2)}.\end{aligned} \tag{A.8}$$

It is easy to check that

$$\int_{S(k)}\exp\{-|y|^2/2\}(y'C^{-1}\theta)\,dy = 0,$$

$$\int_{S(k)}(2\pi)^{-p/2}\exp\{-|y|^2/2\}\|y\|^2\,dy = h(\alpha)\mathrm{tr}(C^{-1}),$$

and

$$\int_{S(k)}(2\pi)^{-p/2}\exp\{-|y|^2/2\}(y'C^{-1}\theta)^2\,dy = h(\alpha)\theta'C^{-2}\theta.$$

It follows from (A.7) that

$$P_\theta(\theta \in C^n(X)) = (1 - \alpha) + \frac{n[1 - \alpha - h(\alpha)]}{\|\theta\|^2}\left\{\mathrm{tr}(C^{-1}) - \frac{(2n + 2)\theta'C^{-2}\theta}{\|\theta\|^2}\right\} + 0(|\theta|^{-3}). \tag{A.9}$$

This gives the desired result except that the error term is $0(|\theta|^{-3})$ instead of $0(|\theta|^{-4})$. It can be checked, however, that due to the symmetry of the problem, the terms which are $0(|\theta|^{-m})$ for m odd must always integrate to zero (as did the term $[2n\|\theta\|^{-2}y'C^{-1}\theta]$). Hence the next nonzero term of the expansion of $P_\theta(\theta \in C^n(X))$ will be $0(|\theta|^{-4})$.

PROOF OF THEOREM 3.3.4. A very laborious calculation exactly paralleling the proof of Theorem 3.3.1 (but including all terms up to $0(|\theta|^{-4})$) gives in place of (A.9)

$$
\begin{aligned}
P_\theta(\theta \in C^n(X)) &= (1-\alpha) + \left[(p-2)/8\right](\theta'\Sigma^{-1}\theta)^{-2} \\
&\quad \times \left\{4p(p-2)(1-\alpha) - 2(p-2)(3p+4)h(\alpha)\right. \\
&\quad \left. + \left[p^3 + 3p^2 - 30p + 16\right]l(\alpha)/3 - (p-1)\left[p^2 + 2p - 32\right]g(\alpha)\right\} \\
&\quad + 0(|\theta|^{-6}),
\end{aligned} \tag{A.10}
$$

where $h(\alpha)$ is given in (A.8),

$$
\begin{aligned}
l(\alpha) &= (2\pi)^{-p/2}\int_{S(k)} y_1^4 \exp\{-|y|^2/2\}\, dy \\
&= 3(1-\alpha) - \frac{6(k/2)^{p/2}\exp\{-k/2\}}{p\Gamma(p/2)}\left(1 + \frac{k}{2(p+2)}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
g(\alpha) &= (2\pi)^{-p/2}\int_{S(k)} y_1^2 y_2^2 \exp\{-|y|^2/2\}\, dy \\
&= (1-\alpha) - \frac{2(k/2)^{p/2}\exp\{-k/2\}}{p\Gamma(p/2)}\left(1 + \frac{k}{(p-1)}\left[1 - \frac{3}{2(p+2)}\right]\right).
\end{aligned}
$$

Inserting these expressions in (A.10) and collecting terms gives the desired result. □

Acknowledgments. I am grateful for the very good suggestions of the referees and the Associate Editor.

REFERENCES

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

[2] BERGER, J. (1976b). Tail minimaxity in location vector problems and its applications. *Ann. Statist.* **4** 33–50.

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

[4] BERGER, J. (1979). Multivariate estimation with nonsymmetric loss functions. In *Optimizing Methods in Statistics*, J. S. Rustagi (Ed.). Academic Press, New York.

[5] BERGER, J. (1980). *Statistical Decision Theory: Foundations, Concepts, and Methods.* Springer, New York.

[6] BERGER, J. and SRINIVASAN, C. (1978). Generalized Bayes estimators in multivariate problems. *Ann. Statist.* **6** 783–801.

[7] BOX, G. E. P. and TIAO, G. C. (1968). Bayesian estimation of means for the random effects model. *J. Amer. Statist. Assoc.* **63** 174–181.

[8] BROWN, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.* **37** 1087–1136.

[9] BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–904.

[10] BROWN, L. D. (1979). An heuristic method for determining admissibility of estimators—with applications. *Ann. Statist.* **7** 960–994.

[11] BROWN, P. J. and ZIDEK, J. V. (1978). Multivariate ridge regression with unknown covariance matrix. Technical report, Univ. British Columbia.

[12] BROWN, P. J. and ZIDEK, J. V. (1980). Adaptive multivariate ridge regression. *Ann. Statist.* **8** 64–74

[13] CASELLA, G. (1977). Minimax ridge regression estimation. Ph.D. thesis, Purdue Univ.

[14] DEMPSTER, A. P., SCHATZOFF, M. and WERMUTH, N. (1976). A simulation study of alternatives to ordinary least squares (with discussion). *J. Amer. Statist. Assoc.* **72** 77–106.

[15] DICKEY, J. M. (1974). Bayesian alternatives to the F-test and least squares estimate in the normal linear model. In *Studies in Bayesian Econometrics and Statistics.* (S. E. Feinberg and A. Zellner, eds.). North Holland, Amsterdam.

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

[17] EFRON, B. and MORRIS, C. (1973b). Combining possibly related estimation problems. *J. Roy. Statist. Soc., Ser. B* **35** 379–421.

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

[19] FAITH, R. E. (1976). Minimax Bayes set and point estimators of a multivariate normal mean. Technical Report No. 66, Univ. Michigan.

[20] HILL, BRUCE M. (1974). On coherence, inadmissibility and inference about many parameters in the theory of least squares. In *Studies in Bayesian Econometrics and Statistics.* (S. Feinberg and A. Zellner, eds.). North Holland, Amsterdam.

[21] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12** 55–68.

[22] HOERL, A. E., KENNARD, R. W. and BALDWIN, K. R. (1975). Ridge regression: some simulations. *Comm. Statist.* **4** 105–123.

[23] HUDSON, M. (1974). Empirical Bayes estimation. Technical Report #58, Stanford Univ.

[24] JOSHI, V. M. (1967). Inadmissibility of the usual confidence sets for the mean of a multivariate normal population. *Ann. Math. Statist.* **38** 1868–1875.

[25] JUDGE, G. and BOCK, M. E. (1977). Implications of pre-test and Stein rule estimators in econometrics. In *The Series Studies in Mathematical and Managerial Economics.* North Holland, Amsterdam.

[26] LEONARD, T. (1976). Some alternative approaches to multiparameter estimation. *Biometrika* **63** 69–75.

[27] 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.

[28] LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41.

[29] MORRIS, C. (1977). Interval estimation for empirical Bayes generalizations of Stein's estimator. The Rand Paper Series, Rand Corp., California.

[30] OLSHEN, A. (1977). Comment on "A note on a reformulation of the S-Method of multiple comparison" by H. Scheffe. *J. Amer. Statist. Assoc.* **72** 144–146.

[31] RAO, C. R. (1977). Simultaneous estimation of parameters—a compound decision problem. In *Statistical Decision Theory and Related Topics II.* (S. S. Gupta and D. S. Moore, eds.). Academic Press.

[32] ROLPH, J. E. (1976). Choosing shrinkage estimators for regression problems. *Comm. Statist.* **A5(9)** 789–802.

[33] RUBIN, H. (1977). Robust Bayesian estimation. In *Statistical Decision Theory and Related Topics II.* (S. S. Gupta and D. S. Moore, eds.). Academic Press.

[34] STEIN, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probability* **1** 197–206. Univ. California Press.

[35] STEIN. C. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. Ser. B* **24** 265–296.

[36] STEIN, C. (1973). Estimation of a mean of a multivariate distribution. *Proc. Prague Symp. Asymptotic Statist.* 345–381.

[37] STEIN, C. (1974). Estimation of the parameters of a multivariate normal distribution—I. Estimation of the means. Stanford Univ., Depart. Statist., Technical Report No. 63.

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

[39] STRAWDERMAN, W. E. (1973). Proper Bayes minimax estimators of the multivariate normal mean vector for the case of common unknown variances. *Ann. Statist.* **1** 1189–1194.

[40] THISTED, R. A. (1976). Ridge regression, minimax estimation, and empirical Bayes methods. Ph.D. Thesis, Stanford Univ.

[41] ZELLNER, A. and VANDAELE, W. (1971). Bayes-Stein estimators for K-means, regression and simultaneous equation models. In *Studies in Bayesian Econometrics and Statistics*. (S. Feinberg and A. Zellner, eds.). North-Holland, Amsterdam.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907