SUBJECTIVE HIERARCHICAL BAYES ESTIMATION
OF A MULTIVARIATE NORMAL MEAN: ON THE
FREQUENTIST INTERFACE*

by

James O. Berger          Christian Robert
Purdue University        Purdue University and
                         Université de Rouen

Technical Report #88-24C

Department of Statistics
Purdue University

June 1988
Revised April 1989

* Research Supported by the National Science Foundation (U.S.), Grant DMS#8702620, and by l'Institut National de la Statistique et des Etudes Economiques (France).
ABSTRACT

In shrinkage estimation of a multivariate normal mean, the two dominant approaches to construction of estimators have been the hierarchical or empirical Bayes approach and the minimax approach. The first has been most extensively used in practice, because of its greater flexibility in adapting to varying situations, while the second has seen the most extensive theoretical development. In this paper we consider several topics on the interface of these approaches, concentrating, in particular, on the interface between hierarchical Bayes and frequentist shrinkage estimation.

The hierarchical Bayes setup considered is quite general, allowing (and encouraging) utilization of subjective second stage prior distributions to represent knowledge about the actual location of the normal means. (The first stage of the prior is used, as usual, to model suspected relationships among the means.) We begin by providing convenient representations for the hierarchical Bayes estimators to be considered, as well as formulas for their associated posterior covariance matrices and unbiased estimators of matricial mean square error; these are typically proposed by Bayesians and frequentists, respectively, as possible “error matrices” for use in evaluating the accuracy of the estimators. These two measures of accuracy are extensively compared in a special case, to highlight some general features of their differences.

Risks and various estimated risks or losses (with respect to quadratic loss) of the hierarchical Bayes estimators are also considered. Some rather surprising minimax results are established (such as one in which minimaxity holds for any subjective second stage prior on the mean), and the various risks and estimated risks are extensively compared.

Finally, a conceptually trivial (but often calculationally difficult) method of verifying minimaxity is illustrated, based on numerical maximization of the unbiased estimator of risk (using certain convenient calculational formulas for hierarchical Bayes estimators), and applied to an illustrative example.

Key Words and Phrases: Hierarchical Bayes estimation; Subjective second stage prior distributions; Posterior covariance matrix; Unbiased estimator of matricial mean square error; Frequentist risk; Posterior expected loss; Unbiased estimator of risk; Superharmonic distributions.

Short Title: Hierarchical Bayes Shrinkage Estimation

1980 AMS Subject Classification: Primary 62F15; secondary 62C20, 62J07.
1. INTRODUCTION

Suppose we observe

$$X = (X_1, X_2, \ldots, X_p)^t \sim N_p(\theta, \Sigma), \quad \Sigma \text{ known},$$

and desire to estimate the unknown $\theta = (\theta_1, \ldots, \theta_p)^t$. We will consider both "inference" and decision-theoretic estimation; for the latter we will utilize the usual quadratic loss for an estimator $\delta(x) = (\delta_1(x), \ldots, \delta_p(x))^t$, namely

$$L(\theta, \delta(x)) = (\theta - \delta(x))^t Q(\theta - \delta(x)),$$

(1.1)

where $Q$ is positive definite. (Some attention will also be paid to the matrix loss $(\theta - \delta(x))(\theta - \delta(x))^t$.)

When $\Sigma = \sigma^2 I_p$ and the $\theta_i$ are thought to be "similar" or exchangeable, an often recommended estimator for $\theta$ (cf. Efron and Morris (1972)) is (when $p > 3$)

$$\delta(x) = x - \frac{(p - 3)\sigma^2}{\sum_{i=1}^p (x_i - \overline{x})^2} (x - \overline{x} 1),$$

(1.2)

where $\overline{x} = \frac{1}{p} \sum_{i=1}^p x_i$ and $1 = (1, 1, \ldots, 1)^t$. The usual derivation of this estimator follows from assuming that the $\theta_i$ are i.i.d. $N(\beta, \sigma^2_n)$, calculating the corresponding Bayes estimator, estimating $\beta$ and $\sigma^2_n$ from the data, and finally inserting these estimates in the Bayes estimator.

This standard empirical Bayes approach has a number of well-documented difficulties, especially when $p$ is small or moderate or when confidence intervals are desired (cf. Berger (1985)). These difficulties are most easily overcome by using the hierarchical Bayesian approach to the problem. Instead of estimating $\beta$ and $\sigma^2_n$ directly, one simply places a "second stage" prior distribution, $\pi_2(\beta, \sigma^2_n)$, on them, and then performs a Bayesian analysis (e.g., calculation of the posterior mean). When prior information about $\beta$ (or $\sigma^2_n$) is available, the hierarchical Bayes estimator can be substantially better than (1.2) when $p$ is small or moderate (cf. Berger (1982b) and Berger and Chen (1987)). And even when the noninformative second stage prior $\pi_2(\beta, \sigma^2_n) \equiv 1$ is used, the hierarchical Bayes approach will typically equal or outperform the empirical Bayes approach. (Note that the modified empirical Bayes approach of Morris (1983), which is itself quite successful, is patterned after the hierarchical Bayes approach.)

A recent discovery in Brown (1987) also pertains to this issue. Brown has shown that (1.2) is inadmissible (in a nontrivial sense) and can be improved upon by additionally incorporating
shrinkage to a specified point. Such additional shrinkage is precisely what subjective hierarchical Bayes estimators tend to produce, providing further frequentist motivation for their study.

From the Bayesian perspective, there are also purely subjective reasons for utilizing the hierarchical Bayesian approach. Here are two examples from Berger (1985) that emphasize the richness of the structures that can be modeled within the hierarchical Bayesian framework. (These examples will be utilized later.)

**Example 1.** For years $1, 2, \ldots, 7$, the IQ of a child is tested. Letting $\theta_i$ be the true IQ in year $i$, suppose that $\theta_i$ is measured by a test score $X_i \sim N(\theta_i, 100)$. Here, it is quite natural to treat the $\theta_i$ as being i.i.d. $N(\beta, \sigma_\theta^2)$, allowing for year-to-year variation in IQ, but recognizing that the IQs should be similar.

Another available piece of information here, assuming that the child is a "random" member of the population (i.e., that he has not been identified as belonging to some special group having a strong correlation with IQ), is that the overall population distribution of IQs is $N(100, 225)$. To incorporate this information, one could assign $\beta$ a $N(100, 225)$ prior distribution.

To complete the hierarchical Bayesian description of the problem, a second stage prior distribution for $\sigma_\theta^2$ is needed. Although an expert might well have subjective knowledge about $\sigma_\theta^2$, which could certainly then be incorporated, it will probably be more common to be quite vague about this parameter, and choose, say, $\pi(\sigma_\theta^2) = 1$.

**Example 2.** Consider a variation on Example 1. Suppose a linear trend in the $\theta_i$ is suspected. This could be modeled as

$$\theta_i = \beta_1 + \beta_2 \cdot i + \epsilon_i,$$

where $\beta_1$ and $\beta_2$ are unknown, and the $\epsilon_i$ are i.i.d. $N(0, \sigma_\epsilon^2)$. This fits into the hierarchical Bayesian framework by defining the first stage prior of $\theta$ to be $N_p(y \beta, \Phi_\theta I_p)$, where

$$y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{pmatrix}^t$$

and $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$.

It is then necessary to also choose a second stage prior $\pi_2(\beta_1, \beta_2, \sigma_\theta^2)$. The prior for $(\beta_1, \beta_2)$ could be chosen in a similar fashion to that for $\beta$ in Example 1.

A third reason to consider the hierarchical Bayesian approach is the need for conditional measures of accuracy. To either construct error estimates or confidence sets, there is considerable evidence that conditional (i.e. data-dependent) measures must be used. (The recent literature on this issue includes Johnstone (1988) and Lu and Berger (1988a,b)). The hierarchical Bayesian
approach produces accuracy measures, based on the posterior distribution, that are automatically conditional. The major competitor to the hierarchical Bayesian approach is the conditional frequentist approach based on unbiased estimators of accuracy (see, e.g., Stein (1981), Johnstone (1988), and Lu and Berger (1988a,b)). We will be partly concerned with comparison of these alternative approaches.

A final motivation for the paper is to consider minimaxity of various hierarchical Bayes estimators. While it has been recognized that minimaxity and "good" shrinkage patterns are often incompatible (cf. Morris (1983), Berger (1985), and Casella (1985)), they are sometimes simultaneously achievable. Here we are only considering estimators developed through Bayesian hierarchical modelling designed to reflect actual beliefs about \( \theta \), so that "good" shrinkage patterns are automatically obtained. If minimaxity is also present, one has a very attractive situation.

Two minimaxity results are discussed. The first, based on ideas of Stein (1981), Zheng (1982), and George (1986a,b,c), is quite surprising, in that it establishes minimaxity of certain hierarchical Bayes estimators simultaneously for all second stage priors on the first stage mean. For instance, one can model exchangeability but also incorporate any subjective information about the location of the common mean, all while staying minimax (and hence satisfactory to a frequentist).

Unfortunately, the result is applicable only in rather special cases. Thus we also discuss a conceptually trivial numerical method of verifying minimaxity of a given estimator, namely numerically maximize the unbiased estimator of risk, and see if it is less than the minimax risk. Although there can be formidable computational problems involved in this verification, the approach is much more general and typically much easier than analytic verification of minimaxity. This will be further discussed in Section 4.

The organization of the paper is as follows. In Section 2, the general model being considered is developed, and useful expressions for the hierarchical Bayes estimators are given. Section 3 considers the determination of estimation accuracy (e.g., estimated variances and estimated risks), from both Bayesian and estimated frequentist perspectives, and compares the two approaches. Section 4 presents the minimaxity results.


2. THE HIERARCHICAL BAYES ESTIMATOR

2.1 THE HIERARCHICAL PRIOR DISTRIBUTION

The prior distribution that will be considered is a mixture of a “first stage” distribution on \( \theta \) w.r.t. hyperparameters \( \mu \) and \( \mathcal{D}_\pi \); in particular, we consider

\[
\pi(\theta) = \int \pi_1(\theta|\mu, \mathcal{D}_\pi) \pi_2(\mu, \mathcal{D}_\pi) d\mu \ d\mathcal{D}_\pi,
\]

where the first stage prior

\[
\pi_1(\theta|\mu, \mathcal{D}_\pi) \quad \text{is} \quad \mathcal{N}_\mu(\mu, \mathcal{D}_\pi)
\]

and the second stage prior is \( \pi_2(\mu, \mathcal{D}_\pi) \), which will always be assumed to have a density w.r.t. Lebesgue measure on the domains of \( \mu \) and \( \mathcal{D}_\pi \). (The theoretical Sections, 2.3.2, 3.1.2, and 4.1, do not require assumption (2.2).) The following two examples indicate the diverse possibilities for choice of \( \pi_2 \); these examples will also form the basis of our later developments. Two important generalizations of these examples are given in Appendix 1.

**Example 3 (Regression structured means).** Suppose

\[
\mu = y\beta,
\]

where \( y \) is a \((p \times \ell)\) matrix of known regressors (such that \( y^t y \) is positive definite) and \( \beta \) is an \((1 \times \ell)\) vector of regression coefficients. Thus \( \theta \) is modelled as having the regression structure

\[
\theta = y\beta + \epsilon,
\]

where \( \epsilon \sim \mathcal{N}_\mu(0, \mathcal{D}_\pi) \). An important special case is that of exchangeable means, defined by

\[
y = 1, \quad \beta \in \mathbb{R}^1, \quad \text{and} \quad \mathcal{D}_\pi = \sigma^2 \mathcal{I}_p.
\]

The second stage prior density will be assumed to be of the form

\[
\pi_2(\beta, \mathcal{D}_\pi) = \pi_2^1(\beta) \pi_2^2(\mathcal{D}_\pi),
\]

where either

**Case 1:** \( \pi_2^1(\beta) \) is \( \mathcal{N}_\epsilon(\beta^0, A) \), or
Case 2: \( \pi_2^0(\beta) \) is \( T_\ell(\alpha, \beta^0, A) \);

here \( \beta^0, A, \) and \( \alpha \) are given, and \( T_\ell(\alpha, \beta^0, A) \) denotes the \( \ell \)-variate \( t \)-distribution with \( \alpha \) degrees of freedom, location vector \( \beta^0 \), and scale matrix \( A \). Usually \( \beta^0 \) can be thought of as a subjectively specified "guess" for \( \beta \), while \( A \) is typically a subjectively specified "accuracy" matrix corresponding to this guess (cf. Example 1). When \( p \) is small (or \( \ell \) is a substantial fraction of \( p \)) it can be quite important to utilize such subjective information about \( \beta \) (cf. Berger, 1982b). Note, however, that it is typically possible to be "noninformative" about \( \beta \) if desired, by letting \( A \to \infty \) in \( \pi_2^0 \) (which can be shown to correspond to choosing \( \pi_2^1(\beta) \equiv 1 \)).

Case 1, the choice of a normal distribution for \( \pi_2^1 \), is calculationally easiest. Using a \( t \)-distribution, as in Case 2, adds one dimension of numerical integration to the calculations but results in additional robustness with respect to the subjective input \( \beta^0 \) (cf. Angers, 1987).

Finally, we will allow \( A^{-1} \) to have eigenvalues that are zero. (All expressions will be in terms of \( A^{-1} \), so there is no need to define \( A \) in this case.) Let

\[ m = \text{Rank} \left( A^{-1} \right), \quad (2.6) \]

and let \( \Omega_0 \) denote the null space of \( A^{-1} \). For \( \beta - \beta^0 \in \Omega_0, \pi_2^1(\beta) \) is constant, implying that the prior is noninformative on that part of the parameter space of \( \beta \). Note that \( m = 0 \) corresponds to a constant (noninformative) prior for the entire parameter space of \( \beta \).

Example 4. The second example that will be utilized for illustrative purposes is based on Berger (1980) (see also Strawderman (1971), Berger (1976a, 1985), and Lu and Berger (1988a)). The example has the virtue of often yielding essentially closed form expressions for most quantities of interest, allowing for easier comparisons of various proposed methodologies.

Take, as the first stage prior,

\[ \pi_1(\theta | \mu, \xi): N_p(\mu, B(\xi)), \quad (2.7) \]

where \( B(\xi) = \xi C - \xi, C \) being a given positive definite matrix and \( \xi \) a scalar. (Thus, \( \xi_x = B(\xi) \) in (2.1).) The domain of \( \xi \) is taken to be a subset of

\[ (\text{ch}_\text{max} C^{-1} \xi, \infty), \quad (2.8) \]

where \( \text{ch}_\text{max} \) stands for maximum characteristic root, so that \( B(\xi) \) is always positive definite. This form of \( \beta(\xi) \) is used because it allows for closed form calculation, while resulting in "robust" Bayesian shrinkage estimators; this is discussed further below.
Various scenarios will prove to be of interest in this example. For instance, the minimax theorems in Section 4 will be established under the assumption that the second stage prior for \((\mu, \xi)\) has conditional densities \(\pi_2(\xi|\mu)\) that are nondecreasing in \(\xi\) for each \(\mu\). The calculational simplifications that were alluded to earlier arise in the following special case.

**Special Case of Example 4:** Let \(H = \{\mu = y\beta: \beta \in \mathbb{R}^\ell\}\), where \(y\) is a given matrix of covariates, as in Example 3, and \([yC^{-1}y']\) has full rank \(\ell\). Suppose further that

\[
\lambda_0 \equiv \text{ch}_{\text{max}}(C^{-1}Y) \leq 1, \tag{2.9}
\]

and that the second stage prior for \((\mu, \xi)\) is supported on \(H \times (1, \infty)\) and has constant density therein.

The assumption about \(\mu\) above is equivalent to placing a noninformative prior on \(\beta\), as mentioned in Example 3 (there, setting \(A^{-1} = 0\)). The case \(\ell = 0\) is allowed, and will be defined by \(H = \{\mu^0\}, \mu^0\) given.

When \(\ell = 0\), the usual choice of \(C\) is

\[
C = \tau(Y^2 + A), \quad \tau = (p - 2)/p, \tag{2.10}
\]

where \(A\) is a specified positive definite matrix satisfying (2.9), i.e.

\[
\text{ch}_{\text{max}}(A^{-1}Y) \leq \frac{1}{2}(p - 2). \tag{2.11}
\]

This choice of \(C\) and \(H\) results in a prior that is similar to the usual conjugate \(N_p(\mu, A)\) prior, in that it is unimodal with subjectively specified mode \(\mu\) while \(A\) can be thought of as a subjectively specified accuracy matrix (for the best guess \(\mu\)). The reason for building a two stage prior (i.e., introducing the random \(\xi\)) is that this robustifies the usual conjugate prior, resulting in familiar robust shrinkage estimators. See Berger (1980, 1985) for general discussion (though Berger, 1985, uses a slightly different prior).

When \(C = Y = I_p\), then this prior can be seen to specify shrinkage towards the subspace \(H\). When \(\ell = 0\), one then has shrinkage towards the point \(\mu^0\). Indeed, defining \(\sigma_x^2 = \xi - 1\), the prior reduces to the Example 3 scenario with \(Y^2 = \sigma_x^2 I_p\) and a noninformative prior on \(\beta\). Note that we will, therefore, also be providing closed form expressions for the hierarchical Bayes estimator in that case.

\[\square\]

### 2.2 EXISTENCE OF THE BAYES ESTIMATOR

The Bayes estimator that we will consider is the posterior mean (optimal for quadratic losses). Since we will often be working with improper second stage prior distributions, \(\pi_2(\mu, Y^2)\), it is
important to keep track of when the Bayes estimator actually exists (i.e., when the posterior distribution has a mean). The following lemma gives such a result. (We use \( f(x|\theta) \) to denote the \( \mathcal{N}_p(\theta, \Sigma) \) density of \( X \).)

**Lemma 2.1.** If, for all \( x \in \mathbb{R}^p \), the marginal distribution

\[
m(x) = \int f(x|\theta)\pi(\theta)d\theta \\
= \int f(x|\theta)\pi_1(\theta|\mu, \Sigma_\pi)\pi_2(\mu, \Sigma_\pi)d\theta d\mu d\Sigma_\pi
\]

is finite, then the posterior mean and covariance matrix exist.

**Proof.** The reason for this result is essentially the analyticity of the Laplace transform on its domain of definition. In fact, we can write

\[
m(x) \propto e^{-\|x\|^2/2} \int e^{\theta^T x} e^{-\|\theta\|^2/2} \pi(\theta)d\theta \\
\propto e^{-\|x\|^2/2} h(x),
\]

where \( h(x) \) is the Laplace transform of \( e^{-\|\theta\|^2/2} \pi(\theta) \). As \( m(x) \) is finite for every \( x \), it follows from Corollary 2.6 of Brown (1986, p. 38) that all derivatives of \( m \) exist at every \( x \in \mathbb{R}^p \). And the posterior mean and the posterior covariance matrix can be expressed in terms of derivatives of \( m \) (see Sections 2.3.2 and 3.1.2). \( \square \)

The following lemmas give conditions under which \( m(x) \) is finite, for the situations of Examples 3 and 4. The proof of Lemma 2.2 is given in Appendix II.

**Lemma 2.2.** Consider the situation of Example 3 (Case 1 or 2) when \( \Sigma_\pi = \sigma_\pi^2 I \). If, for some \( K \),

\[
\int_0^K \pi_2^2(\sigma_\pi^2)d\sigma_\pi^2 < \infty,
\]

and

\[
\int_K^\infty \frac{1}{(\sigma_\pi^2)^{(p-\ell+m)/2}} \cdot \pi_2^2(\sigma_\pi^2)d\sigma_\pi^2 < \infty,
\]

then \( m(x) < \infty \) for all \( x \).

**Note:** The conditions of Lemma 2.2 are satisfied if, for some \( \varepsilon > 0, K_1 > 0, \) and \( K_2 > 0, \)

\[
\pi_2^2(\sigma_\pi^2) \leq \frac{K_1}{K_2 + (\sigma_\pi^2)^{1-\frac{1}{2}(p-\ell+m)+\varepsilon}}.
\]

In particular, the conditions are satisfied by \( \pi_2^2(\sigma_\pi^2) \equiv 1 \) if

\[
p > 2 + \ell - m.
\]
Lemma 2.3. Consider the situation of Example 4, "Special Case." Then, if $p > 2 + \ell$, $m(x) < \infty$ for all $x$.

Proof. A straightforward calculation yields

$$m(x) \propto \int_1^\infty \xi^{-(p-\ell)/2} e^{-\frac{1}{2} \xi} \text{d} \xi,$$

where

$$D = C^{-1} - C^{-1} y' [yC^{-1} y']^{-1} yC^{-1}.$$

The conclusion is immediate.

2.3. Expressions for the Hierarchical Bayes Estimator

There are two quite different representations for the hierarchical Bayes estimator (the posterior mean), $\delta^{HB}$. One is useful for calculation and relies upon the normality of the first stage of the prior distribution; the other will be used for theoretical purposes and is based on a representation in terms of the marginal distribution (2.12). Explicit formulae will be presented when $\Sigma = \sigma_x^2 I_p$.

2.3.1 Calculational Formulae

We have (cf. Lindley and Smith (1972) or Berger (1985))

$$\delta^{HB}(x) = E_\pi(\theta|x)[\theta] = E_{\pi_2}(x, \Sigma_\pi|x)[\delta(x|\mu, \Sigma_\pi)]$$

where, letting $W = (\Sigma + \Sigma_\pi)^{-1}$,

$$\delta(x|\mu, \Sigma_\pi) = x - \Sigma W (x - \mu)$$

and

$$\pi_2(\mu, \Sigma_\pi|x) \propto (\det W)^{1/2} \exp\left\{-\frac{1}{2} (x - \mu)^t W (x - \mu)\right\} \pi_2(\mu, \Sigma_\pi).$$

Note that $\delta(x|\mu, \Sigma_\pi)$ is the conditional mean of $\theta$ given $\mu$ and $\Sigma_\pi$. This decomposition can be calculationally advantageous when $\mu$ and $\Sigma_\pi$ have low dimensional distributions; in that case, the calculation of (2.19) requires only low dimensional integration. Also, when $\mu$ has a normal distribution or a $t$-distribution, the computation of (2.19) simplifies further, as indicated in the following.
examples. For motivational purposes, we begin with the exchangeable scenario of Example 3, as defined by (2.5).

Example 3 (exchangeable means case). Here, $\mu = \beta 1$ and $\beta$ has a normal distribution. Then (2.19) becomes

$$\delta^{HB}(x) = E_{\pi_2}(\sigma_\pi^2 | x)[\delta(x | \sigma_\pi^2)],$$

with

$$\delta(x | \sigma_\pi^2) = x - \frac{\sigma^2}{\sigma^2 + \sigma_\pi^2}(x - \bar{x}1) - \frac{\sigma^2}{(pA + \sigma^2 + \sigma_\pi^2)(x - \beta 0)1} \tag{2.22}$$

and

$$\pi_2(\sigma_\pi^2 | x) \propto \frac{\exp\{-\frac{1}{2} \left[ \frac{||x - \bar{x}1||^2}{\sigma^2 + \sigma_\pi^2} + \frac{[(\bar{x} - \beta 0)^2]}{(pA + \sigma^2 + \sigma_\pi^2)^{1/2} A^{-1/2}} \right] \}}{(\sigma^2 + \sigma_\pi^2)^{(p-1)/2}(pA + \sigma^2 + \sigma_\pi^2)^{1/2} A^{-1/2}} \pi_2(\sigma_\pi^2) \tag{2.23}$$

(see Berger (1985, pp. 183–184)). If one chooses the noninformative second stage prior distribution $\pi_2(\mu, \sigma_\pi^2) = 1$, i.e., if $A = \infty$ and $\pi_2(\sigma_\pi^2) \equiv 1$, then $\delta^{HB}$ is given by

$$\delta^{HB}(x) = x - E_{\pi_2}(\sigma_\pi^2 | x)\frac{\sigma^2}{\sigma^2 + \sigma_\pi^2}(x - \bar{x}1), \tag{2.24}$$

where

$$\pi_2(\sigma_\pi^2 | x) \propto (\sigma^2 + \sigma_\pi^2)^{-(p-1)/2} \exp\{-\frac{||x - \bar{x}1||^2}{2(\sigma^2 + \sigma_\pi^2)}\}. \tag{2.25}$$

This estimator is the hierarchical Bayes version of the estimator (1.2) given in the introduction.

Note that $\delta^{HB}$ is defined even for $p = 3$, as long as $A < \infty$; when $A = \infty$, so $m = 0$, $\delta^{HB}$ does not exist for $p = 3$ (see Lemma 2.2). Thus, when $A < \infty$, $\delta^{HB}$ defines an exchangeable shrinkage estimator when $p = 3$, while (1.2) requires $p \geq 4$. Furthermore, $\delta^{HB}$ will be shown to be minimax even when $p = 3$; thus a frequentist who desires to use an exchangeability-based minimax shrinkage estimator when $p = 3$ must in addition incorporate subjective prior information about the location of the $\theta_i$ (see also the discussion in the introduction concerning Brown (1987)). Of course, if $A$ is very large and $p = 3$, there will be very little shrinkage. Indeed, for large $A$ and $p = 3$, it can be shown that

$$\delta^{HB}(x) \cong x - \frac{2\sqrt{p}}{(\log A)||x - \bar{x}1||^2}(1 - e^{-||x - \bar{x}1||^2/(2\sigma^2)})(x - \bar{x}1). \tag{2.26}$$

Thus, significant practical gains when $p = 3$ will only be available if subjective information about $\beta$ is not too vague. In contrast, when $p \geq 4$ even $A = \infty$ (yielding (2.24)) will result in significant practical gains.

Example 3 (Case 1, continued). If $X \sim N_p(\theta, \Sigma)$, $\pi_1(\theta | \beta, \sigma_\pi^2)$ is $N_p(y \beta, \sigma_\pi^2 I_p)$, and $\beta$ is $N(\beta 0, A)$, it is shown in Berger (1985, pp. 190–192) that

$$\delta^{HB}(x) = E_{\pi_2}(\sigma_\pi^2 | x)[\delta(x | \sigma_\pi^2)],$$
where
\[
\delta(x|\sigma^2_x) = x - \Sigma W (x - \hat{\beta}) - \Sigma W y U A^{-1}(\hat{\beta} - \beta^0),
\]
\[
W = (\Sigma + \sigma^2_x I_p)^{-1}, \quad \hat{\beta} = (y^t W y)^{-1} y^t W x,
\]
\[
U = (y^t W y + A^{-1})^{-1},
\]

\[
\pi_2^1(\sigma^2_x|x) \propto m(x|\sigma^2_x) \pi_2^2(\sigma^2_x),
\]

\[
m(x|\sigma^2_x) = \frac{\exp\left\{-\frac{1}{2}[(x - y\hat{\beta})^t W (x - y\hat{\beta}) + (\hat{\beta} - \beta^0)^t (y^t W y) U A^{-1}(\hat{\beta} - \beta^0)]\right\}}{[\det W]^{-1/2} [\det U]^{1/2}}.
\]

Recall that setting $A^{-1} = 0$ corresponds to choice of a noninformative prior ($\pi_2^1(\beta) = 1$) on $\beta$.

\[\square\]

**Example 3 (Case 2, continued).** Consider the situation above, except that now $\beta \sim T_\alpha(\alpha, \beta^0, A)$. As in Generalization II of Appendix I, we can use the representation of the $T_\alpha(\alpha, \beta^0, A)$ distribution as a Gamma ($\frac{\alpha}{2}, \frac{\alpha}{2}$) (denoted $\pi_3$) mixture of normals, to derive analogous expressions for $\delta^{HB}$. Indeed, one need only replace $A^{-1}$ by $\lambda A^{-1}$ in (2.27) and (2.28) (call the resulting expressions $\delta(x|\lambda, \sigma^2_x)$ and $m(x|\lambda, \sigma^2_x)$, respectively), and define

\[
\pi_2^3(\lambda, \sigma^2_x|x) = m(x|\lambda, \sigma^2_x) \lambda^{m/2} \pi_3(\lambda) \pi_2^2(\sigma^2_x)
\]

(recall that $m$ is the rank of $A^{-1}$). Then

\[
\delta^{HB}(x) = E_{\pi_2^3(\lambda, \sigma^2_x|x)}[\delta(x|\lambda, \sigma^2_x)].
\]

Angers (1987) gives a related expression for $\delta^{HB}$ in this case.

\[\square\]

**Example 4 (Special Case, continued).** As in Berger (1980) (see also Berger (1985) and Lu and Berger (1988a)), it can be shown that

\[
\delta^{HB}(x) = x - h_{(p - l - 2)}(||x||^2_x)^{\epsilon} C^{-1}(x - Px),
\]

where

\[
P = y^t[y C^{-1} y^t]^{-1} y C^{-1},
\]

\[
||x||^2_x = x C^{-1}(I_p - P)x,
\]

and $h_m(v)$ is a closed form expression defined in Appendix III.
2.3.2 Theoretical formulae

A general expression for the posterior mean, when \( X \sim \mathcal{N}_p(\theta, \mathbf{\Sigma}) \), is

\[
\delta^{HB}(x) = x + \mathbf{\Sigma} \nabla \log m(x),
\]  

(2.34)

where \( m(x) \) is the marginal distribution of \( X \). (For a proof when \( \mathbf{\Sigma} = I_p \), see, e.g., Berger and Srinivasan (1978)). Another representation that will be useful follows from defining

\[
m(x|\mu) = \int f(x|\theta) \pi_1(\theta|\mu, \mathbf{\Sigma}_\pi) \pi_2^2(\mathbf{\Sigma}_\pi|\mu) d\theta d\mathbf{\Sigma}_\pi,
\]

so that

\[
m(x) = \int m(x|\mu) \pi_2^1(\mu) d\mu.
\]

Then

\[
\delta^{HB}(x) = x + \mathbf{\Sigma} \frac{\nabla m(x)}{m(x)}
\]

\[
= x + \mathbf{\Sigma} \int \frac{\nabla m(x|\mu) \pi_1^1(\mu)}{m(x)} d\mu
\]

\[
= \int \delta(x|\mu) \pi_1^1(\mu|x) d\mu,
\]

(2.35)

where

\[
\delta(x|\mu) = x + \mathbf{\Sigma} \nabla \log m(x|\mu),
\]

(2.36)

\[
\pi_1^1(\mu|x) = \frac{m(x|\mu) \pi_2^1(\mu)}{m(x)}.
\]

(2.37)

This decomposition will allow us to work conditionally on \( \mu \) (see Section 4.1). For other uses of this type of representation for \( \delta^{HB} \), see Haff (1988).

3. Estimated accuracy and loss

What error measures are to be associated with the hierarchical Bayes estimator \( \delta^{HB} \)? Two types of measures that are often considered are (i) Bayesian posterior measures and (ii) unbiased estimators of loss or variance. The use of Bayesian posterior measures is well-established, while consideration of unbiased estimators of loss is increasing (cf. Stein (1973, 1981), Judge and Bock (1978), Berger (1985), Johnstone (1988), Brown (1988), Bock (1988), and Lu and Berger (1988a,b)). Section 3.1 gives standard posterior measures for our scenario, while Section 3.2 presents unbiased estimators of loss and accuracy. Both “calculational” and “theoretical” versions are given. In Section 3.3, the two types of measures are compared.
3.1 POSTERIOR MEASURES

3.1.1 CALCULATIONAL FORMULAE

For the model developed in previous sections, the posterior mean $\delta^{HB}$ is given by (2.19) and the posterior covariance matrix is

$$V^{HB}(x) = E^{\pi_2(\theta|x)}[\Sigma - \Sigma W \Sigma^t + (\delta(x|\mu, \Sigma_{\theta}) - \delta^{HB}(x))(\delta(x|\mu, \Sigma_{\theta}) - \delta^{HB}(x))^t]$$  \hspace{1cm} (3.1)

where $\delta(x|\mu, \Sigma_{\theta})$ is given by (2.20) and $\pi_2(\mu, \Sigma_{\theta}|x)$ by (2.21) (see Berger (1985, pp. 139–140)). When the quadratic loss (1.1) is being considered, the posterior expected loss is given by

$$\rho^{HB}(x) = E^{\pi_2(\theta|x)}[(\theta - \delta^{HB}(x))^t Q(\theta - \delta^{HB}(x))]$$

$$= \text{tr}(V^{HB}(x)Q).$$  \hspace{1cm} (3.2)

In the various examples, we will explicitly give only the formulae for $V^{HB}$; the formulae for $\rho^{HB}(x)$ follow immediately from (3.1).

Example 3 (continued). For Case 1, the posterior covariance matrix is (Berger (1985, p. 190))

$$V^{HB}(x) = E^{\pi_2(\sigma^2|x)}[\Sigma - \Sigma W \Sigma^t + \Sigma W y U y^t W \Sigma^t$$

$$+ (\delta(x|\sigma^2) - \delta^{HB}(x))(\delta(x|\sigma^2) - \delta^{HB}(x))^t],$$  \hspace{1cm} (3.3)

where $\delta(x|\sigma^2)$, $W$, $U$, and $\pi_2(\sigma^2|x)$ are given by (2.27) through (2.29).

For Case 2, the same formula holds, but with $\lambda A^{-1}$ replacing $A^{-1}$ in $U$ (and elsewhere), $\delta(x|\lambda, \sigma^2)$ and $\pi_2(\sigma^2, \lambda|x)$ replacing $\delta(x|\sigma^2)$ and $\pi_2(\sigma^2|x)$, and $\delta^{HB}(x)$ given by (2.30). □

Example 4 (Special Case, continued). As in Berger (1980, 1985), it can be shown that the posterior covariance matrix is given by

$$V^{HB}(x) = \Sigma - h_{(p-\ell-2)}(||x||^2) \Sigma C^{-1} \Sigma$$

$$+ g_{(p-\ell-2)}(||x||^2) \Sigma C^{-1} (I_p - P)x x^t(I - P)^t C^{-1} \Sigma;$$  \hspace{1cm} (3.5)

here $||x||_2^2$ and $P$ are defined in (2.33) and (2.32), while $h_m$ and $g_m$ are defined in Appendix III. □
3.1.2 Theoretical formulae

Proposition 3.1. If $H_m(x)$ is the Hessian matrix of $m(x)$ (i.e. the matrix with $(i,j)$ element $(\frac{\partial^2}{\partial x_i \partial x_j} m(x)))$, the posterior covariance matrix can be written

$$V^{HB}(x) = \Sigma + \Sigma \frac{H_m(x)}{m(x)} \Sigma - \Sigma (\nabla \log m(x))(\nabla \log m(x))^t \Sigma. \quad (3.6)$$

Proof. Straightforward, using (2.34) and differentiating inside the first integral representation for $m(x)$ in (2.12). \qed

Corollary 3.2. Under quadratic loss, the posterior expected loss of $\delta^{HB}$ is

$$\rho^{HB}(x) = \text{tr}(Q \Sigma) + \frac{1}{m(x)} \text{tr}(H_m(x)\tilde{Q}) - (\nabla \log m(x))^t \tilde{Q}(\nabla \log m(x)), \quad (3.7)$$

where

$$\tilde{Q} = \Sigma Q \Sigma. \quad (3.8)$$

3.2 Unbiased Estimators of Accuracy

For quadratic loss, the usual frequentist measure of performance of $\delta$ is the risk function

$$R(\theta, \delta) = E_{\theta}(\theta - \delta(X))^t Q(\theta - \delta(X)),$$

$E_{\theta}$ denoting expectation with respect to the distribution of $X$ conditionally on $\theta$. Stein (1973, 1981) introduced the unbiased estimator of risk (for the normal problem), which is an expression $D\delta$ satisfying

$$R(\theta, \delta) = E_{\theta}[D\delta(X)]; \quad (3.9)$$

here $D$ is a certain differential operator. The concept has been mainly used to establish minimaxity results, though it is being increasingly used for other purposes (cf. Berger (1982), Spruill (1986), Chen (1988), Bock (1988), Johnstone (1988) and Brown (1988)).

A useful related concept follows from consideration of the matricial mean square error of $\delta$, defined as

$$V(\theta, \delta) = E_{\theta}[(\theta - \delta(X))(\theta - \delta(X))^t]. \quad (3.10)$$
While dominance of one estimator over another according to this criterion is rare, an unbiased estimator of \( V(\theta, \delta^{HB}) \) can be used as a frequentist version of \( V^{HB}(x) \); i.e., it can be used as an estimated “accuracy matrix” and to calculate the unbiased estimate of risk.

**Proposition 3.3.** For \( \delta^{HB}(x) \) in (2.34), assume \( m(x) \) satisfies \( E_\theta |\nabla \log m(X)|^2 < \infty \), \( E_\theta |H_{i,j}(X)/m(X)| < \infty \) for all \( i, j \) (where \( H_{i,j} \) is the \( (i, j) \) entry of \( H_m \)), and

\[
\lim_{|x_i| \to \infty} |\nabla \log m(x)| \exp\{-\frac{1}{2} (x - \theta)^t \Sigma^{-1} (x - \theta)\} = 0
\]

for all \( i \). Then

\[
V(\theta, \delta^{HB}) = E_\theta [\hat{V}_{\delta^{HB}}(X)], \tag{3.11}
\]

where \( \hat{V}_{\delta^{HB}}(\cdot) \), the unbiased estimator of the matricial MSE of \( \delta^{HB} \), is given by

\[
\hat{V}_{\delta^{HB}}(x) = \Sigma + 2 \Sigma \frac{H_m(x)}{m(x)} \Sigma - \Sigma (\nabla \log m(x))(\nabla \log m(x))^t \Sigma. \tag{3.12}
\]

**Proof.** A standard “integration by parts” argument; see Stein (1981) or Berger (1985) for similar proofs.

**Corollary 3.4.** Under the conditions of Proposition 3.3, an unbiased estimator of \( R(\theta, \delta^{HB}) \) is given by

\[
\hat{R}_{\delta^{HB}}(x) = \text{tr}(Q \Sigma) + \frac{2}{m(x)} \text{tr}(H_m(x) \tilde{Q}) - (\nabla \log m(x))^t \tilde{Q}(\nabla \log m(x)), \tag{3.13}
\]

where \( \tilde{Q} = \Sigma Q \Sigma \).

**Proof.** Follows immediately from Proposition 3.3, since

\[
\hat{R}(\theta, \delta^{HB}) = \text{tr}(QV(\theta, \delta^{HB})). \tag*{\Box}
\]

Note the considerable similarity between the results of Propositions 3.1 and 3.3, and between the results of Corollaries 3.2 and 3.4. Indeed, it follows immediately that

\[
\hat{V}_{\delta^{HB}}(x) = 2V^{HB}(x) - \Sigma + (x - \delta^{HB}(x))(x - \delta^{HB}(x))^t, \tag{3.14}
\]

and

\[
\hat{R}_{\delta^{HB}}(x) = 2\rho^{HB}(x) - \text{tr}(Q \Sigma) + (x - \delta^{HB}(x))Q(x - \delta^{HB}(x))^t. \tag{3.15}
\]
These expressions are quite convenient for calculation of $\hat{V}_{\delta HB}$ and $\hat{R}_{\delta HB}$ (see Section 3.1.1).

3.3 COMPARISONS OF THE MEASURES OF ACCURACY

The posterior covariance matrix, $V^{HB}(x)$, and the unbiased estimator of the matricial MSE, $\hat{V}_{\delta HB}(x)$, are natural candidates for an "error matrix" to use in the evaluation of $\delta^{HB}$. (Of course, $V^{HB}$ would likely be preferred by Bayesians, while $\hat{V}_{\delta HB}$ might often be preferred by non-Bayesians.) The possible uses of $V^{HB}$ or $\hat{V}_{\delta HB}$ are many; the diagonal elements give "estimated variances" for the $\delta_i^{HB}$, and "confidence" ellipses or rectangles, based on these matrices (and a normal approximation), are easy to construct (see Berger (1980, 1985) for examples).

Not surprisingly, $V^{HB}$ and $\hat{V}_{\delta HB}$ can be very different. The purpose of this section is to give some indication as to the types of differences that can be expected, so as to allow a more informed choice between $V^{HB}$ and $\hat{V}_{\delta HB}$.

In Section 3.3.1, $V^{HB}(x)$ and $\hat{V}_{\delta HB}(x)$ are compared in a hopefully representative special case. Analogous comparisons between the posterior expected loss, $\rho^{HB}(x)$, and the unbiased estimator of risk, $\hat{R}_{\delta HB}(x)$, are given in Section 3.3.2. Section 3.3.3 contains some discussion.

Some might argue that comparing $V^{HB}$ and $\hat{V}_{\delta HB}$ (or $\rho^{HB}$ and $\hat{R}_{\delta HB}$) is meaningless; after all they are derived from completely different statistical perspectives and mean very different things. Furthermore, since $\delta^{HB}$ is derived using a prior distribution, it might seem odd to some statisticians to even consider using an unbiased estimator of accuracy. Our rationales for this comparison include the following:

(i) In practice, $V^{HB}$ and $\hat{V}_{\delta HB}$ (or $\rho^{HB}$ and $\hat{R}_{\delta HB}$) will be used in exactly the same way: to convey the possible error in $\delta^{HB}$. That they are derived from different perspectives will not mean much to a practitioner; in particular, if they are very different numbers, the natural question will be "which one is a better reflection of accuracy?" It is a conceit of theoreticians to believe that practitioners will be intimately aware of delicate theoretical differences in esoteric situations. To most practitioners, a standard error is a standard error.

(ii) Although $\delta^{HB}$ is derived using a prior distribution, the prior distribution may be viewed by a frequentist as simply a technical device. Very strong arguments can be made that, if one desires to use a shrinkage estimator for frequentist reasons, it should still be developed in a hierarchical Bayesian fashion (to properly direct the shrinkage and possibly to ensure admissibility). In this case the prior would be viewed simply as an artifact, and the frequentist would not necessarily desire to use the posterior measures of accuracy. Much of empirical Bayes analysis (cf., Morris, 1983) can also be viewed in this light.
(iii) Related to (ii), we feel that it is wrong to argue that the unbiased estimators of accuracy are “more robust” or require “less assumptions” than the posterior measures of accuracy. If the prior distribution is viewed simply as a helpful technical device, then the posterior measures of accuracy should start out on an equal footing with the unbiased estimators. Each prior just yields a different accuracy procedure, and it is fair to simply consider and compare such procedures. We have always found it rather curious that non-Bayesian will often consider and compare a variety of different procedures, but will not include procedures that happen to arise as Bayes procedures because “then you must believe in the prior.” This is an unfair double standard. Of course, Bayesians will argue that it is valuable to treat the prior seriously, but our argument is that frequentists will do better if they develop procedures in a Bayesian way, even if they do not take the prior seriously.

In this section we will only consider Example 4, Special Case, of Section 2.1, because the closed form expressions for $V^{HB}$ and $\hat{V}_{\delta HB}$ will allow for easier comparison. We also restrict attention to the $\ell = 0$ case, with $C$ as in (2.10) and (2.11). Again, therefore, the prior is to be thought of as a “robust” alternative to use of the conjugate $N_p(\mu, A)$ prior, $\mu$ and $A$ being subjectively specified location and “scale” factors for $\theta$.

For this situation, it is notationally convenient to define (recalling that $\tau = (p - 2)/p$)

$$B = \Sigma (\Sigma + A)^{-1/2}, \quad z = (\Sigma + A)^{-1/2} (x - \mu),$$

$$||x||^2 = (x - \mu)^t (\Sigma + A)^{-1} (x - \mu) = z^t z = |z|^2,$$

$$h(v) = \tau^{-1} h_{p-2}(\tau^{-1} v), \quad g(v) = \tau^{-2} g_{p-2}(\tau^{-1} v),$$

so that

$$\delta^{HB}(x) = x - h(||x||^2) \Sigma (\Sigma + A)^{-1} (x - \mu),$$

$$V^{HB}(x) = \Sigma - h(||x||^2) BB^t + g(||x||^2) Bzz^t B^t,$$

$$\hat{V}_{\delta HB}(x) = \Sigma - 2h(||x||^2) BB^t - (h^2(||x||^2) + 2g(||x||^2)) Bzz^t B^t,$$

$$\rho^{HB}(x) = \text{tr}(Q \Sigma) - h(||x||^2) \text{tr}(QBB^t) + g(||x||^2) z^t B^t QBz,$$

$$\hat{\rho}_{\delta HB}(x) = \text{tr}(Q \Sigma) - 2h(||x||^2) \text{tr}(QBB^t) + (h^2(||x||^2) + 2g(||x||^2)) z^t B^t QBz.$$

3.3.1 Comparison of variances

I. Small values of $||x||^2$

As $v \to 0$, it is shown in Berger (1980) that $h(v) \to 1$ and $g(v) \to 4/(p^2 - 4)$. Hence, if $||x||^2$ is small,

$$V^{HB}(x) \cong \Sigma - B B^t = \Sigma - \Sigma (\Sigma + A)^{-1} \Sigma,$$

$$\hat{V}_{\delta HB}(x) \cong \Sigma - 2BB^t = \Sigma - 2\Sigma (\Sigma + A)^{-1} \Sigma.$$
This exposes a potential problem with $\hat{V}_{\delta HB}$, since $\hat{V}_{\delta HB}$ will have negative eigenvalues unless $\mathcal{S} \leq A$. One might thus need some type of positive part fix for $\hat{V}_{\delta HB}$. Even then, however, $\hat{V}_{\delta HB}$ can be accused of being too small. To see this note that, for small $||x||^2$,

$$\delta^{HB}(x) \equiv x - \mathcal{S}(\mathcal{S} + A)^{-1}(x - \mu),$$

which happens to be the posterior mean w.r.t. a conjugate $\mathcal{N}_p(\mu, A)$ prior. For this conjugate prior, $\mathcal{S} - \mathcal{S}(\mathcal{S} + A)^{-1}\mathcal{S}$ is the posterior covariance matrix, and is often considered to be an optimistic assessment of the accuracy of the posterior mean (because of possible prior uncertainty). The often substantially smaller $\hat{V}_{\delta HB}$ might strike many as definitely too small, therefore.

II. LARGE VALUES OF $||x||^2$

As $v \to \infty$, it is shown in Berger (1980) that $vh(v) \to (p - 2)$ and $v^2g(v) \to 2(p - 2)$. Hence, for large values of $||x||^2$,

$$V^{HB}(x) \approx \mathcal{S} - \frac{(p - 2)}{||x||^2}BB^t + \frac{2(p - 2)}{||x||^4}Bzz^tB^t,$$

$$\hat{V}_{\delta HB}(x) \approx \mathcal{S} - \frac{2(p - 2)}{||x||^2}BB^t + \frac{(p^2 - 4)}{||x||^4}Bzz^tB^t.$$

Note first that both $V^{HB}$ and $\hat{V}_{\delta HB}$ converge to $\mathcal{S}$ (at a rate proportional to $||x||^{-2}$). This is natural, since it can also be shown that $\delta^{HB}(x) \to x$, and lends credence to the analysis being “robust” w.r.t. possible misspecification of $\mu$ and $A$. (If $\mu$ and/or $A$ is misspecified, $||x||^2$ will tend to be large.)

Note next that

$$\hat{V}_{\delta HB}(x) - V^{HB}(x) \approx \frac{(p - 2)}{||x||^2} \left[ -BB^t + pB \begin{pmatrix} z \\ |z| \end{pmatrix} \begin{pmatrix} z \\ |z| \end{pmatrix}^t B^t \right].$$

The interest here is that the difference clearly has a comparatively large eigenvalue (at least when $p$ is large) in the $Bz = \mathcal{S}(\mathcal{S} + A)^{-1}(x - \mu)$ direction. Thus $\hat{V}_{\delta HB}$ seems to assess the accuracy in this direction to be less than does $V^{HB}$. This behavior will be seen to also hold for moderate $||x||^2$, and will be discussed further in Section 3.3.3.

III. MODERATE $||x||^2$ AND LARGE $p$

Recall that $\mu$ and $A$ are roughly to be thought of as the prior mean and covariance matrix for $\theta$. Hence $\mu$ and $(\mathcal{S} + A)$ are roughly the marginal mean and covariance matrix of $X$, so that we would “expect” to have

$$||x||^2 = (x - \mu)^t(\mathcal{S} + A)^{-1}(x - \mu) \approx p.$$
Indeed, as \( p \to \infty, \frac{||x||^2}{p} \) would then converge to 1.

In Appendix III, it is shown that, if \( \frac{||x||^2}{p} \to 1 \) as \( p \to \infty \), then \( \eta(||x||^2) \to 1 \) and \( pg(||x||^2) \to (2 - 4/\pi) \). Hence, for large \( p \) and \( ||x||^2 \leq p \),

\[
V^{HB}(x) \cong \mathcal{S} - BB^t + (2 - \frac{4}{\pi})B\tilde{z}^tB^t,
\]

\[
\tilde{V}_{\delta HB}(x) \cong \mathcal{S} - 2BB^t + pB\tilde{z}^tB^t,
\]

where \( \tilde{z} = z/||z| = (\mathcal{S} + A)^{-1/2}(x - \mu)/||x - \mu|| \).

Interestingly, this exhibits features of both the “small \( ||x||^2 \)” and “large \( ||x||^2 \)” cases simultaneously. To see this, let

\[
w_{(1)}^{(1)} = \frac{B\tilde{z}}{|B\tilde{z}|} = \frac{\mathcal{S}(\mathcal{S} + A)^{-1}(x - \mu)}{|\mathcal{S}(\mathcal{S} + A)^{-1}(x - \mu)|}
\]

and \( \{w_{(1)}, w_{(2)}, \ldots, w_{(p)}\} \) be an orthonormal basis. Then the “variances” of the “contrasts” \( w_{(i)}(\theta - \delta^{HB}(x)) \) are, for \( i \geq 2 \),

\[
w_{(i)}^{(1)}V^{HB}(x)w_{(i)} = w_{(i)}^{(1)}(\mathcal{S} - BB^t)w_{(i)},
\]

\[
w_{(i)}^{(1)}\tilde{V}_{\delta HB}(x)w_{(i)} = w_{(i)}^{(1)}(\mathcal{S} - 2BB^t)w_{(i)},
\]

and, for \( i = 1 \),

\[
w_{(1)}^{(1)}V^{HB}(x)w_{(1)} = w_{(1)}^{(1)}(\mathcal{S} - BB^t)w_{(1)} + (2 - \frac{4}{\pi})\tilde{z}^tB^tB\tilde{z},
\]

\[
w_{(1)}^{(1)}\tilde{V}_{\delta HB}(x)w_{(1)} = w_{(1)}^{(1)}(\mathcal{S} - 2BB^t)w_{(1)} + p\tilde{z}^tB^tB\tilde{z}.
\]

For \( i \geq 2 \), the variances arising from \( \tilde{V}_{\delta HB} \) might seem “too small,” much as in the “small \( ||x||^2 \)” situation. On the other hand, for large \( p \) and \( i = 1 \), the variance arising from \( \tilde{V}_{\delta HB} \) can be huge, much larger than that arising from \( V^{HB} \); this is related to the difference between \( \tilde{V}_{\delta HB} \) and \( V^{HB} \) that was noted for large \( ||x||^2 \).

IV. A NUMERICAL EXAMPLE

To indicate that the insights gained from the previous “limiting” cases can hold for “normal” situations, consider the example \( p = 6, \mu = 0, \mathcal{S} = \text{diag.} \{0.1, 1.0, 1.0, 1.0, 1.0, 1.0, 10.0\}, \) and \( A = \text{diag.} \{1.55, 2.0, 2.0, 2.0, 2.0, 2.0, 6.5\}. \) (Note that \( \text{ch}_{\text{max}} A^{-1} \mathcal{S} = 10/6.5 < 2 = \frac{1}{2}(p - 2) \), so that (2.11) is satisfied.)

We will investigate the behavior of \( V^{HB} \) and \( \tilde{V}_{\delta HB} \) when \( z = |z|e_i, \ e_i \) being the unit vector on the \( i^{th} \) axis; thus we assume that \( y \) (and hence \( x \)) lies on a coordinate axis. It is then easy to
see that $V^{HB}_i$ and $\hat{V}^{HB}_i$ are both diagonal matrices, with diagonal elements

$$V^{HB}_i(x) = \sigma_i^2 - \frac{\sigma_i^4}{(\sigma_i^2 + A_i)} \left[ h(|z|^2) - |z|^2 g(|z|^2) \right],$$

$$V^{HB}_j(x) = \sigma_j^2 - \frac{\sigma_j^4}{(\sigma_j^2 + A_j)} h(|z|^2) \quad \text{if } j \neq i;$$

$$\hat{V}_i(x) = \sigma_i^2 - \frac{\sigma_i^4}{(\sigma_i^2 + A_i)} [2h(|z|^2) - \{h^2(|z|^2) + 2g(|z|^2)\} |z|^2],$$

$$\hat{V}_j(x) = \sigma_j^2 - \frac{2\sigma_j^4}{(\sigma_j^2 + A_j)} h(|z|^2) \quad \text{if } j \neq i,$$

$\sigma_i^2$ and $A_i$ being the diagonal elements of $\Sigma$ and $A$, respectively. Here $h$ and $g$ have the comparatively simple forms (see Appendix III)

$$h(v) = \frac{4}{v} - \frac{9v}{2(4e^{3v/4} - 4 - 3v)},$$

$$g(v) = \frac{8}{v^2} + \frac{9(e^{3v/4}[4 - 3v] - 4)}{(4e^{3v/4} - 4 - 3v)^2}.$$

Figure 1 graphs $h(v)$, $h_1(v) = [h(v) - vg(v)]$, and $h_2(v) = [2h(v) - \{h^2(v) + 2g(v)\}v]$ as functions of $v$. It is then easy to compare the $V^{HB}_k(x)$ and $\hat{V}_k(x)$ for any value of $|z|^2$. For instance, if $|z|^2 = 6$ (recall that $|z|^2 = p$ is what one “expects” to observe), then $h(6) = 0.587$, $h_1(6) = -0.149$, and $h_2(6) = -2.36$, so that the $V^{HB}_j(x)$ for $j \neq i$ are the “conservative”

$$V^{HB}_j(x) = \sigma_j^2 - (0.587) \frac{\sigma_j^4}{(\sigma_j^2 + A_j)}$$

(compared with the conjugate prior variances $\sigma_j^2 - \sigma_j^4/(\sigma_j^2 + A_j)$), while the $\hat{V}_j(x)$ are the “optimistic”

$$\hat{V}_j(x) = \sigma_j^2 - (1.17) \frac{\sigma_j^4}{(\sigma_j^2 + A_j)}.$$

On the other hand, the variances $V^{HB}_i(x)$ and $\hat{V}_i(x)$ for $|z|^2 = 6$ are given by

$$V^{HB}_i(x) = \sigma_i^2 + (0.149) \frac{\sigma_i^4}{(\sigma_i^2 + A_i)},$$

$$\hat{V}_i(x) = \sigma_i^2 + (2.36) \frac{\sigma_i^4}{(\sigma_i^2 + A_i)}.$$

Interestingly, both $V^{HB}_i$ and $\hat{V}_i$ are larger than $\sigma_i^2$, but $\hat{V}_i(x)$ is dramatically larger. For instance,

$$V^{HB}_6(x) = 10 + (0.149) \frac{100}{16.5} = 10.90,$$

$$\hat{V}_6(x) = 10 + (2.36) \frac{100}{16.5} = 24.30.$$
To emphasize: in this example if one observes $z = (0,0,0,0,0,\sqrt{6})^t$ (i.e., $x = (0,0,0,0,0,9.95)^t$), then the estimated variance of $\theta_6$ obtained from $V^{HB}$ is 10.90, while that obtained from $\hat{V}_{\delta HB}$ is 24.30. Note that, here,

$$
\hat{\sigma}^2_{\delta HB}(x) = x_6 - h(6)\sigma^2_6(\sigma^2_6 + A_6)^{-1}x_6
$$

$$
= 9.95 - (.587)\frac{10}{16.5}(9.95) = 6.41,
$$

so that $\hat{\sigma}^2_{\delta HB}$ shifts $x_6 = 9.95$ about one sample standard deviation ($\sqrt{10}$).

Of some interest is the observation that

$$
\sup_{x} \sup_{w: ||w||=1} w^t(\hat{V}_{\delta HB} - V^{HB})w

= \sup_{|x|} e_6^t(\hat{V}_{\delta HB} - V^{HB})e_6

= \frac{\hat{\sigma}^2_6}{(\sigma^2_6 + A_6)} \sup_v [-h(v) + (h^2(v) + g(v)v]

= (6.0606) \sup_v \left[ \frac{20}{v} - \frac{9v(10 + 3v)}{4(4e^{3v/4} - 3v)} \right]

= 13.44
$$

(the maximum occurring for $v = 6.24$, which is near the "expected value" of 6 for $||x||^2$).

### 3.3.2 Comparison of Risks

For illustrative purposes here we take $Q = I_p$ in (1.1). The formulae for $\rho^{HB}(x)$, analogous to those in Section 3.3.1, are as follows: we only give the analogs of Parts III and IV.

### III. Moderate $||x||^2$ and Large $p$

Under the condition $||x||^2/p \to 1$ and $p$ large,

$$
\rho^{HB}(x) \approx \text{tr} \Sigma - \text{tr} BB^t + (2 - \frac{4}{\pi})\hat{z}^t B^t B \hat{z},
$$

$$
\hat{R}_{\delta HB}(x) \approx \text{tr} \Sigma - 2\text{tr} BB^t + p\hat{z}^t B^t B \hat{z}.
$$

To highlight the differences, consider $\hat{z}_i$, the unit eigenvector corresponding to the characteristic root $\lambda_i$ of $B^t B$. Then (writing $x^i$ for the corresponding value of $x$, and $\Lambda = (\text{tr} BB^t - \lambda_i)$)

$$
\rho^{HB}(x^i) = \text{tr} \Sigma - \Lambda - \left(\frac{4}{\pi} - 1\right)\lambda_i,
$$

and

$$
\hat{R}_{\delta HB}(x^i) = \text{tr} \Sigma - 2\Lambda + (p - 2)\lambda_i.
$$

Note first that the $\rho^{HB}(x^i)$ are always less than $\text{tr} \Sigma$, while $\hat{R}_{\delta HB}(x^i)$ can be much larger (if $p$ is large, and $\lambda_i$ is large compared to the average of the other characteristic roots). On the
other hand, \( \rho^{HB} \) is bounded below by \( \text{tr} \; \mathcal{Y} - \text{tr} \; BB^t > 0 \), while \( \hat{R}_{\delta HB} \) can be much smaller (even negative) when \( \lambda_i \) is a small characteristic root. (This is true even for \( |z|^2 = p \); for smaller \( |z| \), it will often be the case that \( \hat{R}_{\delta HB} \) is negative.)

IV. THE NUMERICAL EXAMPLE

For \( z = |z|e_i \),

\[
\rho^{HB}(x) = \sum_j \sigma_j^2 - h(|z|^2) \sum_j \frac{\sigma_j^4}{(\sigma_j^2 + A_j)} + |z|^2 g(|z|^2) \frac{\sigma_i^4}{(\sigma_i^2 + A_i)},
\]

\[
\hat{R}_{\delta HB}(x) = \sum_j \sigma_j^2 - 2h(|z|^2) \sum_j \frac{\sigma_j^4}{(\sigma_j^2 + A_j)} + |z|^2 [h^2(|z|^2) + 2g(|z|^2)] \frac{\sigma_i^4}{\sigma_i^2 + A_i}.
\]

For \( z = |z|e_6 \), these become

\[
\rho^{HB}(x) = (14.1) - (7.4)h(|z|^2) + (6.061)|z|^2 g(|z|^2),
\]

\[
\hat{R}_{\delta HB}(x) = (14.1) - (14.8)h(|z|^2) + (6.061)|z|^2 [h^2(|z|^2) + 2g(|z|^2))].
\]

At \( z = \sqrt{6}e_6 \), \( \rho^{HB} = 14.23 \) and \( \hat{R}_{\delta HB} = 26.83 \). This is quite a discrepancy, \( \hat{R}_{\delta HB} \) estimating the risk as being almost twice \( \rho^{HB} \).

At the other extreme, for \( z = |z|e_1 \),

\[
\rho^{HB}(x) = (14.1) - (7.4)h(|z|^2) + (.00606)|z|^2 g(|z|^2),
\]

\[
\hat{R}_{\delta HB}(x) = (14.1) - (14.8)h(|z|^2) + (.00606)|z|^2 [h^2(|z|^2) + 2g(|z|^2))],
\]

which at \( z = \sqrt{6}e_1 \) become, \( \rho^{HB} = 9.76 \) and \( \hat{R}_{\delta HB} = 5.43 \). Here \( \hat{R}_{\delta HB} \) evaluates the risk as being only about half of \( \rho^{HB} \). (Again, we have chosen \( |z|^2 = 6 \) to make the comparison because it is what we "expect" to observe.)

For "intermediate" \( z \), \( \rho^{HB}(x) \) and \( \hat{R}_{\delta HB}(x) \) can be much closer. For instance, if \( z = |z| (1, 1, 1, 1, 1, 1)^t/\sqrt{6} \),

\[
\rho^{HB}(x) = (14.1) - (7.4)[h(|z|^2) - |z|^2 g(|z|^2)/6],
\]

\[
\hat{R}_{\delta HB}(x) = (14.1) - (7.4)[2h(|z|^2) - |z|^2 [h^2(|z|^2) + 2g(|z|^2)))/6],
\]

which at \( z = (1, 1, \ldots, 1)^t \) become \( \rho^{HB} = 10.67 \) and \( \hat{R}_{\delta HB} = 9.78 \).

Graphs of \( \rho^{HB}(x) \) and \( \hat{R}_{\delta HB}(x) \) for the three cases \( z = |z|e_6 \), \( z = |z|e_1 \), and \( z = |z| (1,1,1,1,1,1)^t/\sqrt{6} \), are given as functions of \( |z| \) in Figure 2. They are labelled \( \rho_1, \rho_2, \rho_3 \) and \( \hat{R}_1, \hat{R}_2, \hat{R}_3 \), respectively. Note that the \( \hat{R}_i \to -0.7 \) as \( |z| \to 0 \), and are always substantially smaller than the corresponding \( \rho_i \) for small \( |z| \).
3.3.3 Discussion

The differences between \( \hat{V}_{\delta HB} \) or \( \hat{R}_{\delta HB} \) and \( V^{HB} \) or \( \rho^{HB} \) can be partly explained by the differences between frequentist and Bayesian evaluations of error. For instance, in the example of Section 3.3.2 IV, the actual frequentist risk at \( \theta = (0,0,0,0,0,10)^t \) is about 21 (see Berger (1980), Figure 1), while the posterior Bayes risk for \( x = (0,0,0,0,0,10)^t \) is about 14. The large \( \hat{R}_{\delta HB}(x) \cong 27 \) for this \( x \) is thus partly due to its estimating an inherently larger quantity.

Whether the frequentist risk of 21 or the Bayesian posterior risk of 14 is a better measure of accuracy when \( x \) is near \( (0,0,0,0,0,10)^t \) is an issue we will sidestep. Note, however, that there are arguments both ways. For instance, on the frequentist side one might argue that a situation of possible nonrobustness w.r.t. the prior has been identified; in particular, the “great” fit of \((x_1, \ldots, x_6)^t \) to the prior beliefs about \((\theta_1, \ldots, \theta_6)^t \) overcomes the “bad” fit of \( x_6 \) to the prior belief about \( \theta_6 \) (recall that \( \mu_6 = 0 \) and \( \sqrt{A_6} = \sqrt{6.5} = 2.55 \)), so that the Bayesian estimator will substantially shrink towards \( \mu = 0 \). But one might worry about the bad fit of \( x_6 \), especially upon observing that much less shrinkage would result from utilization of a prior for which the \( \theta_i \) were independent. (An alternative type of “fix” for individual extreme coordinates is discussed in Berger and Dey (1985) — see also Berger (1985) — based on an idea in Stein (1981).) In general, a frequentist risk that is substantially larger than \( \rho^{HB}(x) \) would cause us to investigate the robustness of \( \delta^{HB} \) more carefully.

Of course, we are not considering the report of \( R(\theta, \delta^{HB}) \), but instead the report of \( \hat{R}_{\delta HB}(x) \) (or \( \hat{V}_{\delta HB}(x) \)). And we have identified a seemingly systematic problem with the latter: when \( ||x|| \) is small, \( \hat{R}_{\delta HB} \) or \( \hat{V}_{\delta HB} \) seem themselves to be too small (even sometimes negative) while if \( ||x||^2 \) is moderate or large (in certain directions), \( \hat{R}_{\delta HB} \) or \( \hat{V}_{\delta HB} \) will be too large (such as in the previously discussed example in which \( \hat{R}_{\delta HB} ((0,0,0,0,0,10)^t) \cong 27 \) while the risk function in the vicinity of \( (0,0,0,0,0,10)^t \) is no more than 21 and \( \rho^{HB} \) is only about 14).

Upon reflection, the reason for \( \hat{R}_{\delta HB} \) or \( \hat{V}_{\delta HB} \) being “extreme” is clear. Consider \( \hat{R}_{\delta HB} \), for instance, recalling that

\[
E_{\theta} \hat{R}_{\delta HB}(X) = R(\theta, \delta^{HB}). \tag{3.16}
\]

Let \( \theta_m \) and \( \theta_M \) be values of \( \theta \) minimizing and maximizing \( R(\theta, \delta^{HB}) \). (In the numerical example, \( \theta_m = 0 \) and \( \theta_M \cong (0,0,0,0,0,12)^t \).) For \( x \) in the immediate vicinity of \( \theta_m \), it must be the case that \( \hat{R}_{\delta HB}(x) \) is generally less than \( R(\theta_m, \delta^{HB}) \) or (3.16) will not hold when \( \hat{R}_{\delta HB}(x) \) is averaged over all \( x \). Similarly, for \( x \) in the immediate vicinity of \( \theta_M \), \( \hat{R}_{\delta HB}(x) \) must typically exceed \( R(\theta_M, \delta^{HB}) \) for (3.16) to hold. This systematic tendency toward extremes is troubling, especially at the lower
end. Our opinion is that having errors (or estimated risks) less than \( V_H^B(x) \) (or \( \rho_H^B(x) \)) is very hard to justify, and is the most serious potential failing of \( \hat{V}_H^B \) and \( \hat{R}_H^B \).

In conclusion, our preference is to use \( V_H^B(x) \) and \( \rho_H^B(x) \) as the estimates of accuracy with, however, the qualification that if \( \hat{V}_H^B(x) \) or \( \hat{R}_H^B(x) \) are much larger, then investigation of robustness with respect to the prior assumption (in particular w.r.t. the strong implied dependence of the \( \theta_i \)) should be undertaken.

4. MINIMAXITY OF \( \delta_H^B \)

4.1 ANALYTIC SUFFICIENT CONDITIONS

To show that \( \delta_H^B = x + \hat{\mathcal{Q}} \nabla \log m(x) \) is minimax, it is sufficient to show that (see (3.13))

\[
\hat{R}_H^B(x) \leq \text{tr} \left( Q \hat{\mathcal{Q}} \right) \quad \text{for all } x, \quad (4.1)
\]

since then \( R(\theta, \delta_H^B) = E_\theta \hat{R}_H^B(X) \leq \text{tr} \left( Q \hat{\mathcal{Q}} \right) \), the minimax risk for the problem. It is straightforward to show that (4.1) can be rewritten as

\[
\nabla(\hat{Q} \nabla \sqrt{m(x)}) \leq 0 \quad \text{for all } x, \quad (4.2)
\]

where \( \hat{Q} = \mathcal{Q} \hat{\mathcal{Q}} \), and

\[
\nabla(v(x)) = \sum_{i=1}^{p} \frac{\partial}{\partial x_i} v_i(x).
\]

When \( \hat{Q} = I_p \), (4.2) is the celebrated "superharmonicity" minimax condition of Stein (1981); see also Zheng (1982), George (1986a,b,c), Haff and Johnson (1986), and Haff (1988).

In general, analytic verification of (4.1) (or (4.2)) can be very difficult, especially for complicated estimators such as \( \delta_H^B \). In one circumstance, however, verification is relatively easy. The following proposition, generalizing results of Stein (1981), Zheng (1982), and George (1986a), provides the needed tool.

**Proposition 4.1.** For the situation of Proposition 3.3, \( \delta_H^B \) is minimax if

\[
\nabla(\hat{Q} \nabla m(x|\mu)) \leq 0 \quad \text{for all } x \text{ and } \mu. \quad (4.3)
\]

(See Section 2.3.2 for definition of \( m(x|\mu) \).)

**Proof.** Clearly,

\[
\nabla(\hat{Q} \nabla m(x)) = \int_{\mathbb{R}_p} \nabla(\hat{Q} \nabla m(x|\mu)) \pi_\mu \mu \, d\mu,
\]
so $\hat{V}(\tilde{Q} \nabla m(x)) \leq 0$. But it can be easily verified that this implies (4.2), proving the result. 

The great simplification in use of (4.3) is that one can work conditionally on $\mu$. Furthermore, if (4.3) is satisfied, then $\delta^{HB}$ is minimax regardless of the distribution, $\pi^1_2$, chosen for $\mu$ (subject to the mild conditions of Section 2.2 and Proposition 3.3). This is startling, not only because of its generality, but also because it is an instance in which essentially any subjective prior information about a parameter ($\mu$) can be utilized while maintaining complete frequentist justification (minimaxity). In the next section we will discuss conditions on $\pi^2_2(\mathfrak{X}_{\pi} | \mu)$ under which (4.3) holds.

4.2 MINIMAXITY OF $\delta^{HB}$ IN THE EXAMPLES

Consider first the scenario of Example 4, in which the first stage prior is $\mathcal{N}_p(\mu, \xi C - \mathfrak{X})$, $C$ given, and $(\mu, \xi)$ has a second stage prior density

$$\pi_2(\mu, \xi) = \pi^1_2(\mu)\pi^2_2(\xi | \mu).$$

We will choose

$$Q = \mathfrak{X}^{-1} C \mathfrak{X}^{-1}. \tag{4.4}$$

Minimaxity results for other choices of $Q$ can be given but are of less interest, in that for other $Q$ the condition (4.3) can only be satisfied by inadmissible estimators. Furthermore, if $Q$ differs substantially from (4.4), then $\delta^{HB}$ will not be minimax; basically, minimaxity and Bayesian shrinkage patterns are compatible only for rather special $Q$.

**Theorem 4.2** If, for all $\mu$, $\pi^2_2(\xi | \mu)$ is non-decreasing on $(0, \infty)$, then (4.3) is satisfied.

**Proof.** Since (see Section 2.3.1)

$$W = (\mathfrak{X} + \mathfrak{X}_{\pi})^{-1} = \xi^{-1} C^{-1}$$

and

$$\tilde{Q} = \mathfrak{X} Q \mathfrak{X} = C,$$

calculation yields

$$m(x | \mu) = K \int \xi^{-p/2} \exp \left\{ -\frac{(x - \mu)^t C^{-1} (x - \mu)}{2\xi} \right\} \pi^2_2(\xi | \mu) d\xi$$

and

$$\hat{V}(\tilde{Q} \nabla m(x | \mu)) = K \int_0^\infty \left\{ -\frac{p}{\xi} + \frac{(x - \mu)^t C^{-1} (x - \mu)}{\xi^2} \right\} \times \exp \left\{ -\frac{(x - \mu)^t C^{-1} (x - \mu)}{2\xi} \right\} \xi^{-p/2} \pi^2_2(\xi | \mu) d\xi.$$
Defining \( a = (x - \mu)^t C^{-1} (x - \mu) / 2 \), condition (4.3) is equivalent to

\[
\psi(a) \equiv \int (2a - p\xi) \xi^{-(p+4)/2} e^{-a/\xi} \pi_2^2(\xi|\mu) d\xi \leq 0, \quad \text{for all } a > 0.
\]

Letting \([t_0, \infty)\) denote the support of \( \pi_2^2(\xi|\mu) \), integration by parts yields

\[
\psi(a) = -2e^{-a/t_0} t_0^{-p/2} \pi_2^2(t_0|\mu) - 2 \int_{t_0}^\infty \xi^{-p/2} e^{-a/\xi} \pi_2'(\xi|\mu) d\xi,
\]

where \( \pi_2' \) denotes the (almost everywhere existing) derivative of \( \pi_2^2(\xi|\mu) \). (Note that the monotonicity condition on \( \pi_2^2 \) ensures that this integration by parts is valid.) But, since \( \pi_2^2(\xi|\mu) \) is nondecreasing, \( \pi_2'(\xi|\mu) \geq 0 \), and the right hand side of (4.5) is clearly negative, completing the proof.

A natural choice for \( \pi_2^2 \) is \( \pi_2^2(\xi|\mu) \equiv 1 \) (on a subset \([t_0, \infty)\) of (2.8)). This clearly is nondecreasing, and so (4.3) is satisfied and the resulting \( \delta^{HB} \) will be minimax. Note that this covers the “Special Case” of Example 4 that we have frequently discussed.

In Berger (1980), the second stage prior distribution for \( \xi \) that was considered was (with \( \mu \) being given)

\[
\pi_2^2(\xi|\mu) \propto \xi^{-(n+1-p/2)} \quad \text{on } (1, \infty),
\]

where any \( n \leq (p - 2)/2 \) could be selected. These are all nondecreasing, but only \( n = (p - 2)/2 \) (corresponding to the uniform prior on \((1, \infty))\) yields an admissible estimator. Indeed, it is unlikely that one would ever want an unbounded increasing \( \pi_2^2(\xi|\mu) \). (Note that, for fixed \( \mu \), minimaxity theorems based on hierarchical priors of this type were given in Strawderman (1971) and Berger (1976a, 1980).)

It might, on the other hand, be desired to use decreasing \( \pi_2^2(\xi|\mu) \). Unfortunately, (4.3) cannot be satisfied for such \( \pi_2^2 \), as the following lemma shows.

**Lemma 4.3.** If there exists \( t_1 \) such that, on \((t_1, \infty)\), \( \pi_2^2(\xi|\mu) \) is continuous, nonincreasing and nonconstant, then (4.3) cannot hold for all \( x \).

**Proof.** In the proof of Theorem 4.2, it was shown that (4.3) is equivalent to showing that (4.5) is negative. Now, by the assumptions on \( \pi_2^2 \), there exists an interval \((b, c)\), \( b > t_0 \), such that \( \pi_2'(\xi|\mu) < -\varepsilon < 0 \) on \((b, c)\). Clearly

\[
\int_{t_0}^\infty \xi^{-p/2} e^{-a/\xi} \pi_2'(\xi|\mu) d\xi < (-\varepsilon) \int_b^c \xi^{-p/2} e^{-a/\xi} d\xi \\
\leq (-\varepsilon) c^{-p/2} e^{-a/c} (c - b).
\]
Thus (see (4.5))

$$\psi(a) > -2e^{-a/t_0}t_0^{-p/2} \pi_2^2(t_0|\mu) + 2\varepsilon e^{-p/2} e^{-a/b}(e - b).$$

Letting $a \to \infty$ in this expression, it becomes clear that (4.5) can be positive.

Although Lemma 4.3 rules out decreasing priors, a variety of non-monotonic priors will also satisfy (4.3) for every $x$ and $\mu$. For instance, certain $\pi_2^2$ which decrease for a while, then increase, and then either are constant or continue to increase, can be shown to satisfy the condition. Oscillating priors (that finish on an increase) also might work. We have not attempted to determine which of these more general priors satisfy the condition, because they do not seem natural in practice.

Although we have presented the results in this section in terms of Example 4, they also apply to the Example 3 scenario, providing one wants to choose $\Sigma_x = \sigma_x^2 \mathcal{N}$; then simply set $C = \mathcal{N}$ in Example 4, so that $\sigma_x^2 = (\xi - 1)$.

### 4.3 NUMERICAL VERIFICATION OF MINIMAXITY OF $\delta^{HB}$

Because of the special choice of $Q$ and the special nature of $\pi_2^2(\xi|\mu)$ required for the analytic minimaxity proof in Section 4.2, an alternative general method for verifying minimaxity of $\delta^{HB}$ is clearly desirable. An obvious method exists: for a given estimator, simply numerically verify (4.1) or (4.2). In this regard, (3.15) provides the most useful calculational formula for $R_{\delta^{HB}}$, so that the numerical problem can be rewritten as showing that

$$\Delta(x) = 2[\text{tr} (Q \mathcal{N}) - \rho^{HB}(x)] - [x - \delta^{HB}(x)]^t Q [x - \delta^{HB}(x)] \geq 0. \quad (4.6)$$

(Haff and Johnson (1986) give a related expression.) Thus, simply have a computer minimize $\Delta(x)$, and check to see if the minimum is nonnegative.

Numerically minimizing $\Delta(x)$ is not necessarily trivial. First of all, calculation of $\delta^{HB}$ and $\rho^{HB}$ will often involve numerical integration, and inaccuracies in the integration can cause instabilities in the minimization routine. Second, as always in high dimensions, one needs to worry about local minima. Third, if $\delta^{HB}$ is minimax, $\Delta(x)$ will converge to its minimum (of zero) as $|x| \to \infty$, so that one has to truncate the minimization algorithm when $\Delta(x)$ gets within $\varepsilon$ of 0 and $|x|$ is large. (Strictly speaking, one has then only shown that $\delta^{HB}$ is probably $\varepsilon$-minimax; a tail-minimax argument, as in Berger (1976b), could be employed to complete a proof of minimaxity, but from a practical perspective this would hardly seem necessary if $\varepsilon$ were small.)

**Example 1 (continued).** In the notation of Example 4 (continued) in Section 2.3.1, (2.5) holds and $p = 7$, $\sigma^2 = 100$, $\beta^0 = 100$, $A = 225$, and $\pi_2^2(\sigma_x^2) = 1$. For sum of squares error loss ($Q = I_7$,
in (1.1)), the results in Section 4.2 (and Lemma 2.2) show that $\delta^{HB}$ is minimax. Here, however, the current IQ, $\theta_7$, might be of substantially more importance than the previous IQs, so that $Q = \text{diag.}\{1,1,1,1,1,1,q\}$ (with $q > 1$) might be deemed to be more reasonable. We will investigate the minimaxity of $\delta^{HB}$ for such $Q$, using the numerical method.

For this example, algebra yields

$$
(100)^{-2} \Delta(x) = 2(E_1 - 225E_3)(6 + q) + (E_1^2 - 2E_3)[s^2 + (q - 1)(x_7 - \overline{x})^2] \\
+ (E_1^2 - 2E_4)(\overline{x} - 100)^2(6 + q) + 2(E_1E_2 - 2E_5)(x_7 - \overline{x})(\overline{x} - 100)(q - 1), \tag{4.7}
$$

where $s^2 = ||x - \overline{x}||^2$ and the $E_i$ are the expectations with respect to $\pi_x^2(\sigma^2_x|x)$ of, respectively, $(100 + \sigma^2_x)^{-1}$, $(1875 + \sigma^2_x)^{-1}$, $(100 + \sigma^2_x)^{-2}$, $(1875 + \sigma^2_x)^{-2}$, and $(100 + \sigma^2_x)^{-1}(100 + \sigma^2_x)^{-1}$. From (4.7) (and (2.23)) it is not hard to show that $\Delta(x)$ actually depends only on the three quantities $x_7, \overline{x} = \frac{1}{6} \sum_{i=1}^{6} x_i$, and $s^2 = \frac{1}{6} \sum_{i=1}^{6} (x_i - \overline{x})^2$, and that these quantities vary independently. The minimization of (4.7) was thus done in only three dimensions; IMSL minimization and integration routines were used throughout.

Figure 3 presents the minimum of $\Delta(x)/100$, as a function of $q$. (The accuracy of the minima is about 0.05.) For $q \leq 1.7$, the minimum is zero, indicating that $\delta^{HB}$ is minimax for such $q$. For $q > 1.7$, however, $\delta^{HB}$ is clearly not minimax.

The simplicity of the above numerical verification of minimaxity, compared with analytic verification in general, should arguably make it the preferred technique (unless the analytic technique simultaneously handles a wide range of useful estimators). This is especially so because analytic verification is only occasionly possible (and then typically only in simple situations), while the numerical approach is always available (though not necessarily always doable computationally). A bonus that is obtained from the numerical method is a bound (the minimum of $\Delta(x)$) on the degree of non-minimaxity (since $R(\theta, \delta) - \text{tr} (Q \mathcal{X}) \leq -\inf \Delta(x)$).

Finally, note that minimization of $\Delta(x)$ is considerably simpler than maximization of $R(\theta, \delta^{HB})$ over $\theta$, since

$$R(\theta, \delta^{HB}) = \text{tr} (Q \mathcal{X}) - E_\theta \Delta(X);$$

the presence of the additional expectation over $X$, in calculation of $R(\theta, \delta^{HB})$, so complicates the numerical problem as to make it unmanageable on a routine basis for complicated $\delta^{HB}$. Thus the existence of an unbiased estimator of risk, and the availability of relatively simple expressions for it, are crucial elements of the numerical method.
BIBLIOGRAPHY


Report, Department of Mathematics, Cornell University, Ithaca.


APPENDIX I

GENERALIZATIONS OF EXAMPLES 3 AND 4

I. Frequently, we consider the case $\mathcal{H}_x = \sigma_\pi^2 I_p$ in the examples. An apparently more general case of considerable interest is that in which correlation among the $\theta_i$ is allowed, i.e. in which

$$\mathcal{H}_x = \sigma_\pi^2 I_p + \rho 11^t.$$

It is of interest to note that this case can be reduced to the $\mathcal{H}_x = \sigma_\pi^2 I_p$ case by defining

$$\mu^* = \mu + \sqrt{\rho} Z 1,$$

where $Z \sim \mathcal{N}(0, 1)$, and observing that

$$\pi(\theta) = \int \pi_1(\theta|\mu, \rho, \sigma_\pi^2) \pi_2(\mu, \rho, \sigma_\pi^2) d\mu d\rho d\sigma_\pi^2$$

$$= \int \pi_1^*(\theta|\mu^*, \sigma_\pi^2) \pi_2^*(\mu^*, \sigma_\pi^2) d\mu^* d\sigma_\pi^2;$$

where $\pi_1^*(\theta|\mu^*, \sigma_\pi^2)$ is $\mathcal{N}_p(\mu^*, \sigma_\pi^2 I_p)$ and

$$\pi_2^*(\mu^*, \sigma_\pi^2) = \int \pi_2(\mu^* - \sqrt{\rho} Z 1, \rho, \sigma_\pi^2) (2\pi)^{-1/2} e^{-z^2/2} dz d\rho.$$

II. A related apparent generalization (cf. Dickey (1968, 1974)) is that in which the first stage prior, $\pi_1(\theta|\mu, \mathcal{H}_x)$, is chosen to be $\mathcal{T}_p(\alpha, \mu, \mathcal{H}_x)$. Note, however, that this distribution is the mixture of a normal distribution w.r.t. a gamma distribution:

$$\pi_1(\theta|\mu, \mathcal{H}_x) = \int \pi_1^*(\theta|\mu, \mathcal{H}_x, \lambda) d\hat{s}_3(\lambda),$$

where $\pi_1^*(\theta|\mu, \mathcal{H}_x, \lambda)$ is $\mathcal{N}_p(\mu, \lambda^{-1} \mathcal{H}_x)$ and $\hat{s}_3(\lambda)$ is $\mathcal{G}(\frac{\alpha}{2}, \frac{\lambda}{2}, \frac{2}{\alpha})$. Therefore, this case can also be reduced to the canonical form in (2.1) and (2.2), by writing

$$\pi(\theta) = \int \pi_1^*(\theta|\mu, \mathcal{H}_x) \pi_2^*(\mu, \mathcal{H}_x) d\mu d\mathcal{H}_x,$$

where

$$\pi_2^*(\mu, \mathcal{H}_x) = \int \pi_2(\mu, \frac{1}{\rho} \mathcal{H}_x) \rho \hat{s}_3(\rho) d\rho.$$
APPENDIX II

Proof of Lemma 2.2.

Case 1. From Berger (1985, Section 4.6) one obtains that (up to a multiplicative constant)

\[ m(x) = \int_0^\infty m(x|\sigma_x^2)\pi_2^2(\sigma_x^2) d\sigma_x^2, \]

where \( m(x|\sigma_x^2) \) is given by (2.29). The exponential part of (2.29) is clearly bounded by 1, so to establish the finiteness of \( m(x) \) it is only necessary to verify that

\[ \int_0^\infty (\det W)^{1/2}[\det (y^tWy + A^{-1})]^{-1/2}\pi_2^2(\sigma_x^2) d\sigma_x^2 < \infty. \]  
(A1)

For \( 0 < \sigma_x^2 < K \),

\[ (\det W) \leq \det (\mathcal{S}^{-1}), \]  
(A2)

and

\[ \det (y^tWy + A^{-1}) \geq \det (y^t(\mathcal{S} + K I_p)^{-1}y). \]  
(A3)

It follows immediately from (A2), (A3), and (2.13) that

\[ \int_0^K (\det W)^{1/2}[\det (y^tWy + A^{-1})]^{-1/2}\pi_2^2(\sigma_x^2) d\sigma_x^2 < \infty. \]

For \( \sigma_x^2 > K \),

\[ \det (W) \leq \det (\sigma_x^2 I_p)^{-1} = \sigma_x^{-2p}, \]  
(A4)

and

\[ \det (y^tWy + A^{-1}) \geq \det (K'\sigma_x^{-2}y^t y + A^{-1}) \]
\[ = \det (K'\sigma_x^{-2}y^t y) \det (I_p + \frac{\sigma_x^2}{K'} A^{-1}(y^t y)^{-1}) \]
\[ = \det (K'\sigma_x^{-2}y^t y) \prod_{i=1}^m (1 + \frac{\sigma_x^2}{K'} \rho_i) \]
\[ \geq K^*(\sigma_x^2)^{-\ell_m}(\prod_{i=1}^m \rho_i); \]  
(A5)

where \( \{\rho_1, \ldots, \rho_m\} \) are the nonzero eigenvalues of \( A^{-1}(y^t y)^{-1} \). It follows immediate from (A4), (A5), and (2.14) that

\[ \int_K^\infty (\det W)^{1/2}[\det (y^tWy + A^{-1})]^{-1/2}\pi_2^2(\sigma_x^2) d\sigma_x^2 < \infty. \]
Case 2. Using the representation discussed in Section 2.3.1, Example 3 — Case 2, one has

\[ m(x) = \int_0^\infty \int_0^\infty m(x|a, \sigma^2) \pi_2^2(\sigma^2) \lambda^{m/2} \pi_3(\lambda) d\lambda d\sigma^2, \]

where \( m(x|a, \sigma^2) \) is given by (2.29) with \( A^{-1} \) replaced by \( \lambda A^{-1} \), and \( \pi_3 \) is a Gamma \( \left( \frac{a}{2}, \frac{\sigma^2}{2} \right) \) density. Again, the exponential part of \( m(x|a, \sigma^2) \) is bounded by 1, so it suffices to show that

\[ \int_0^\infty \int_0^\infty (\det W)^{1/2} [\det (y^t W y + \lambda A^{-1})]^{-1/2} \pi_2^2(\sigma^2) \lambda^{m/2} \pi_3(\lambda) d\lambda d\sigma^2 < \infty. \]

For \( K < \sigma^2 < \infty \), the bounds (A4) and (A5) that were given in Case 1 are still valid, with \( \prod_{i=1}^m \rho_i \) replaced by \( \lambda^m \prod_{i=1}^m \rho_i \). It then follows immediately, using also (2.14), that

\[ \int_0^K \int_0^\infty (\det W)^{1/2} [\det (y^t W y + \lambda A^{-1})]^{-1/2} \pi_2^2(\sigma^2) \lambda^{m/2} \pi_3(\lambda) d\lambda d\sigma^2 < \infty. \]

For \( 0 < \sigma^2 < K \), one needs to replace (A3) by

\[ \det (y^t W y + \lambda A^{-1}) \geq \det (y^t (\mathcal{I} + K I_p)^{-1} y) \lambda^m \prod_{i=1}^m \rho_i^*, \]

where now \( \{\rho_1^*, \ldots, \rho_m^*\} \) are the nonzero eigenvalues of \( A^{-1}(y^t (\mathcal{I} + K I_p)^{-1} y)^{-1} \). Together with (A2) and (2.13), this directly implies that

\[ \int_0^K \int_0^\infty (\det W)^{1/2} [\det (y^t W y + \lambda A^{-1})]^{-1/2} \pi_2^2(\sigma^2) \lambda^{m/2} \pi_3(\lambda) d\lambda d\sigma^2 < \infty, \]

completing the proof.

\[ \square \]

APPENDIX III

Define

\[ h_m(v) = \frac{m}{v} \left[ 1 - H_m\left( \frac{v}{2} \right) \right], \]

\[ g_m(v) = \frac{2m}{v^2} \left[ 1 + \left\{ \frac{v}{2} (h_m(v) - 1) - 1 \right\} H_m\left( \frac{v}{2} \right) \right], \]

where

\[ H_m(v) = \begin{cases} 
  v^{m/2} \left[ \frac{\Gamma\left( \frac{m}{2} + 1 \right)}{\Gamma\left( \frac{m}{2} \right)} \{ e^v - \sum_{i=1}^{\frac{m-3}{2}} \frac{v^i}{i!} \} \right]^{-1} & \text{if } m \text{ is even} \\
  v^{m/2} \left[ \Gamma\left( \frac{m}{2} + 1 \right) \{ e^v [2 \Phi(\sqrt{2}v) - 1] - \sum_{i=0}^{\frac{m-3}{2}} \frac{v^{i+1/2}}{\Gamma(i+3/2)} \} \right]^{-1} & \text{if } m \text{ is odd}; 
\end{cases} \]
where \( \Phi \) is the standard normal c.d.f. and the summation in the last expression is defined to be zero when \( m = 1 \).

**Lemma A1.** In the situation of Section 3.3.1, part III, suppose that \( \|x\|^2/p \rightarrow 1 \) as \( p \rightarrow \infty \). Then \( h(\|x\|^2) \rightarrow 1 \) and \( pg(\|x\|^2) \rightarrow (2 - \frac{4}{\pi}) \).

**Proof.** The result that \( h(\|x\|^2) \rightarrow 1 \) follows easily from Lemma 2.1.1 \((vi)\) of Berger (1980). To show that \( pg(\|x\|^2) \rightarrow (2 - \frac{4}{\pi}) \), note first that it is easy to show that any \( \|x\|^2 \) such that \( \|x\|^2/p \rightarrow 1 \) will give the same limiting result. Hence, for convenience, we will choose \( \|x\|^2 = 2n = p - 2 \). Note next that (see Berger (1980) for definitions)

\[
v g(v) = \left[ \frac{r_n(v) - r_n^2(v)}{v} \right] - 2 \frac{r_n(v)}{v} + (2n - r_n(v)) \left( \frac{r_n(v)}{v} - 1 \right),
\]

which, at \( v = 2n = p - 2 \), equals

\[
(2n)g(2n) = 2 \frac{r_n(2n)}{2n} - \frac{(2n - r_n)^2}{2n}.
\]

Again, Lemma 2.1.1 \((vi)\) shows that \( (2n)^{-1}r_n(2n) \rightarrow 1 \) as \( n \rightarrow \infty \), so that we need only show that (see Berger (1980))

\[
\frac{(2n - r_n)^2}{2n} = 2n \left[ \sum_{i=0}^{\infty} \frac{n^i \cdot i!}{(n + i)!} \right]^{-2} \rightarrow \frac{4}{\pi}
\]

as \( n \rightarrow \infty \). Stirling’s formula gives

\[
\frac{n^i \cdot i!}{(n + i)!} = \frac{n^i e^{-n} (n + 1)^{(n + 1) \frac{i}{2}} \sqrt{2\pi} (1 + O(\frac{1}{n}))}{e^{-n} (n + i + 1) (n + i + 1)^{(n + i + 1) \frac{i}{2}} \sqrt{2\pi} (1 + O(\frac{1}{n}))} = (1 - \frac{i + 1}{n + i + 1})^i (1 - \frac{i}{n + i + 1})^{n + i + \frac{1}{2}} e^{-i} (1 + O(\frac{1}{n})).
\]  \( \text{(A5)} \)

(Within this proof, \( O(\cdot) \) and \( o(\cdot) \) are to be understood to be uniform in \( 0 \leq i < \infty \).) Now

\[
\left( 1 - \frac{i + 1}{n + i + 1} \right)^i = \left( 1 - \frac{i}{n + i + 1} \right)^i \left( 1 - \frac{1}{n + 1} \right)^i,
\]

so, for \( i \leq n^\alpha \) where \( \alpha < 1 \),

\[
\frac{n^i \cdot i!}{(n + i)!} = \left( 1 - \frac{i}{n + i + 1} \right)^{n + i + 1} e^{i} (1 + o(1)).
\]  \( \text{(A6)} \)

Next note that, for \( i \leq n^\alpha \) where \( \alpha < \frac{2}{3} \),

\[
\log \left( 1 - \frac{i}{n + i + 1} \right)^{n + i + 1} = -i - \frac{i^2}{2(n + i + 1)} + o(1),
\]
so that (A6) becomes

\[ \frac{n^i n!}{(n + i)!} = \exp \left\{ -\frac{i^2}{2(n + i + 1)} \right\} (1 + o(1)). \]

Thus, for \( \alpha < \frac{2}{3} \),

\[ \sum_{i=0}^{n^\alpha} \frac{n^i n!}{(n + i)!} = \sum_{i=0}^{n^\alpha} \exp \left\{ -\frac{i^2}{2(n + i + 1)} \right\} (1 + o(1)) \]

\[ = \sum_{i=0}^{n^\alpha} \exp \left\{ -\frac{i^2}{2n} \right\} (1 + o(1)). \]

Finally, if \( \alpha > \frac{1}{3} \),

\[ \frac{1}{\sqrt{n}} \sum_{i=0}^{n^\alpha} e^{-i^2/(2n)} = \frac{1}{\sqrt{n}} \int_0^{n^\alpha} e^{-x^2/(2n)} dx (1 + o(1)) \]

\[ = \sqrt{\frac{\pi}{2}} (1 + o(1)). \]

Thus, for \( \frac{1}{3} < \alpha < \frac{2}{3} \),

\[ \lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{i=0}^{n^\alpha} \frac{n^i n!}{(n + i)!} = \sqrt{\frac{\pi}{2}}. \]

To deal with \( i > n^\alpha \), note that for all \( i \)

\[ \left( 1 - \frac{(i+1)}{(n+i+1)} \right)^i \leq \left( 1 - \frac{i}{(n+i+1)} \right)^i \]

and

\[ \log \left( 1 - \frac{i}{(n+i+1)} \right)^{n+i+1} \leq -i - \frac{i^2}{2(n+i+1)}. \]

Together with (A5) these imply that

\[ \frac{n^i n!}{(n + i)!} \leq \exp \left\{ -\frac{i^2}{2(n + i + 1)} \right\} (1 + O(\frac{1}{n})). \]

It is straightforward to check that, for \( \frac{1}{2} < \alpha < \frac{2}{3} \),

\[ \frac{1}{\sqrt{n}} \sum_{i=n^\alpha}^{\infty} \exp \left\{ -\frac{i^2}{2(n + i + 1)} \right\} \to 0, \]

completing the proof. \( \square \)
Figure 1. Graphs of \( h(v) \), \( h_1(v) \), and \( h_2(v) \).
Figure 2. Graphs along various rays of the posterior expected losses, $\rho_i(|z|)$, and unbiased estimators of risk, $\hat{R}_i(|z|)$; here $i=1, 2, \text{ and } 3$ denote the $|z| (0,0,0,0,1)^t$, $|z| (1,0,0,0,0)^t$, and $|z| (1,1,1,1,1)^t/\sqrt{6}$ rays, respectively.
Figure 3. The value of $\inf_{x} \Delta(x)/100$, as a function of the loss $q$. 