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

by

James O. Berger
Purdue University

Christian Robert
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

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

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

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}(\boldsymbol{x})) = (\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{x}))^t \boldsymbol{Q}(\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{x})), \tag{1.1}$$

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

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

$$\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{x} - \frac{(p-3)\sigma^2}{\sum_{i=1}^{p} (x_i - \overline{x})^2} (\boldsymbol{x} - \overline{x}\, \boldsymbol{1}), \tag{1.2}$$

where $\overline{x} = \frac{1}{p} \sum_{i=1}^{p} x_i$ and $\boldsymbol{1} = (1, 1, \ldots, 1)^t$. The usual derivation of this estimator follows from assuming that the θ_i are i.i.d. $\mathcal{N}(\beta, \sigma_\pi^2)$, calculating the corresponding Bayes estimator, estimating β and σ_π^2 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 β and σ_π^2 directly, one simply places a "second stage" prior distribution, $\pi_2(\beta, \sigma_\pi^2)$, on them, and then performs a Bayesian analysis (e.g., calculation of the posterior mean). When prior information about β (or σ_π^2) 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 $\pi_2(\beta, \sigma_\pi^2) \equiv 1$ 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 modelled 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 θ_i be the true IQ in year i, suppose that θ_i is measured by a test score $X_i \sim \mathcal{N}(\theta_i, 100)$. Here, it is quite natural to treat the θ_i as being i.i.d. $\mathcal{N}(\beta, \sigma_\pi^2)$, 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 $\mathcal{N}(100, 225)$. To incorporate this information, one could assign β a $\mathcal{N}(100, 225)$ prior distribution.

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

Example 2. Consider a variation on Example 1. Suppose a linear trend in the θ_i is suspected. This could be modelled as

$$\theta_i = \beta_1 + \beta_2 \, i + \varepsilon_i,$$

where β_1 and β_2 are unknown, and the ε_i are i.i.d. $\mathcal{N}(0, \sigma_\pi^2)$. This fits into the hierarchical Bayesian framework by defining the first stage prior of $\boldsymbol{\theta}$ to be $\mathcal{N}_p(\boldsymbol{y}\ \boldsymbol{\beta}, \sigma_\pi^2 \boldsymbol{I}_p)$, where

$$\boldsymbol{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{pmatrix}^t \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

It is then necessary to also choose a second stage prior $\pi_2(\beta_1, \beta_2, \sigma_\pi^2)$. The prior for (β_1, β_2) could be chosen in a similar fashion to that for β 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 $\boldsymbol{\theta}$, 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.

Among the huge literature on hierarchical Bayesian methodology, works that consider estimators similar to those in this paper include Lindley and Smith (1972), Box and Tiao (1973), Smith (1973), Deely and Lindley (1981), DuMouchel and Harris (1983), Berger (1985), and Angers (1987). Works that discuss minimaxity of Bayes estimators include Brown (1971), Strawderman (1971, 1973), Efron and Morris (1972), Berger (1976a, 1980, 1982a,b, 1985), Faith (1978), Judge and Bock (1978), Stein (1981), Li (1982), Zheng (1982), Chen (1983, 1988), Cooley and Lin (1983), George (1986a,b,c), Haff and Johnson (1986), Spruill (1986), Berger and Chen (1987), DasGupta and Rubin (1988), and Haff (1988).

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 $\boldsymbol{\theta}$ w.r.t. hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\pi$; in particular, we consider

$$\pi(\boldsymbol{\theta}) = \int \pi_1(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi)\pi_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi)d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}_\pi, \tag{2.1}$$

where the first stage prior

$$\pi_1(\boldsymbol{\theta}|\boldsymbol{\mu},\ \boldsymbol{\Sigma}_\pi) \quad \text{is} \quad \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi) \tag{2.2}$$

and the second stage prior is $\pi_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi)$, which will always be assumed to have a density w.r.t. Lebesgue measure on the domains of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\pi$. (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 π_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

$$\boldsymbol{\mu} = \boldsymbol{y}\boldsymbol{\beta}, \tag{2.3}$$

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

$$\boldsymbol{\theta} = \boldsymbol{y}\boldsymbol{\beta} + \varepsilon, \tag{2.4}$$

where $\varepsilon \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\pi)$. An important special case is that of exchangeable means, defined by

$$\boldsymbol{y} = \mathbf{1}, \qquad \boldsymbol{\beta} \in R^1, \quad \text{and} \quad \boldsymbol{\Sigma}_\pi = \sigma_\pi^2 \boldsymbol{I}_p. \tag{2.5}$$

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

$$\pi_2(\boldsymbol{\beta}, \boldsymbol{\Sigma}_\pi) = \pi_2^1(\boldsymbol{\beta})\pi_2^2(\boldsymbol{\Sigma}_\pi),$$

where either

Case 1: $\pi_2^1(\boldsymbol{\beta})$ is $\mathcal{N}_\ell(\boldsymbol{\beta}^0, \boldsymbol{A})$, or

Case 2: $\pi_2^1(\boldsymbol{\beta})$ is $\mathcal{T}_\ell(\alpha, \boldsymbol{\beta}^0, \boldsymbol{A})$;

here $\boldsymbol{\beta}^0$, $\boldsymbol{A}$, and α are given, and $\mathcal{T}_\ell(\alpha, \boldsymbol{\beta}^0, \boldsymbol{A})$ denotes the ℓ-variate t-distribution with α degrees of freedom, location vector $\boldsymbol{\beta}^0$, and scale matrix $\boldsymbol{A}$. Usually $\boldsymbol{\beta}^0$ can be thought of as a subjectively specified "guess" for $\boldsymbol{\beta}$, while $\boldsymbol{A}$ is typically a subjectively specified "accuracy" matrix corresponding to this guess (cf. Example 1). When p is small (or ℓ is a substantial fraction of p) it can be quite important to utilize such subjective information about $\boldsymbol{\beta}$ (cf. Berger, 1982b). Note, however, that it is typically possible to be "noninformative" about $\boldsymbol{\beta}$ if desired, by letting $\boldsymbol{A} \to \infty$ in π_2^1 (which can be shown to correspond to choosing $\pi_2^1(\boldsymbol{\beta}) \equiv 1$).

Case 1, the choice of a normal distribution for π_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 $\boldsymbol{\beta}^0$ (cf. Angers, 1987).

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

$$m = \text{ Rank } (\boldsymbol{A}^{-1}), \tag{2.6}$$

and let Ω_0 denote the null space of $\boldsymbol{A}^{-1}$. For $\boldsymbol{\beta} - \boldsymbol{\beta}^0 \in \Omega_0$, $\pi_2^1(\boldsymbol{\beta})$ is constant, implying that the prior is noninformative on that part of the parameter space of $\boldsymbol{\beta}$. Note that $m = 0$ corresponds to a constant (noninformative) prior for the entire parameter space of $\boldsymbol{\beta}$.

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,

$$\pi_1(\boldsymbol{\theta}|\boldsymbol{\mu}, \xi): \ \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{B}(\xi)), \tag{2.7}$$

where $\boldsymbol{B}(\xi) = \xi \boldsymbol{C} - \boldsymbol{\Sigma}$, $\boldsymbol{C}$ being a given positive definite matrix and ξ a scalar. (Thus, $\boldsymbol{\Sigma}_\pi = \boldsymbol{B}(\xi)$ in (2.1).) The domain of ξ is taken to be a subset of

$$(\text{ch}_{\max} \boldsymbol{C}^{-1} \boldsymbol{\Sigma}, \infty), \tag{2.8}$$

where $\text{ch}_{\max}$ stands for maximum characteristic root, so that $\boldsymbol{B}(\xi)$ 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 $(\boldsymbol{\mu}, \xi)$ has conditional densities $\pi_2^2(\xi|\boldsymbol{\mu})$ that are nondecreasing in ξ for each $\boldsymbol{\mu}$. The calculational simplifications that were alluded to earlier arise in the following special case.

Special Case of Example 4: Let $H = \{\boldsymbol{\mu} = \boldsymbol{y}\boldsymbol{\beta}\colon\ \boldsymbol{\beta} \in \mathbb{R}^{\ell}\}$, where $\boldsymbol{y}$ is a given matrix of covariates, as in Example 3, and $[\boldsymbol{y}\boldsymbol{C}^{-1}\boldsymbol{y}^t]$ has full rank ℓ. Suppose further that

$$\lambda_0 \equiv \operatorname{ch}_{\max}(\boldsymbol{C}^{-1}\boldsymbol{\Sigma}) \leq 1, \tag{2.9}$$

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

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

When $\ell = 0$, the usual choice of $\boldsymbol{C}$ is

$$\boldsymbol{C} = \tau(\boldsymbol{\Sigma} + \boldsymbol{A}), \quad \tau = (p-2)/p, \tag{2.10}$$

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

$$\operatorname{ch}_{\max}(\boldsymbol{A}^{-1}\boldsymbol{\Sigma}) \leq \frac{1}{2}(p-2). \tag{2.11}$$

This choice of $\boldsymbol{C}$ and $\boldsymbol{H}$ results in a prior that is similar to the usual conjugate $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{A})$ prior, in that it is unimodal with subjectively specified mode $\boldsymbol{\mu}$ while $\boldsymbol{A}$ can be thought of as a subjectively specified accuracy matrix (for the best guess $\boldsymbol{\mu}$). The reason for building a two stage prior (i.e., introducing the random ξ) 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 $\boldsymbol{C} = \boldsymbol{\Sigma} = \boldsymbol{I}_p$, then this prior can be seen to specify shrinkage towards the subspace H. When $\ell = 0$, one then has shrinkage towards the point $\boldsymbol{\mu}^0$. Indeed, defining $\sigma_\pi^2 = \xi - 1$, the prior reduces to the Example 3 scenario with $\boldsymbol{\Sigma}_\pi = \sigma_\pi^2 \boldsymbol{I}_p$ and a noninformative prior on $\boldsymbol{\beta}$. Note that we will, therefore, also be providing closed form expressions for the hierarchical Bayes estimator in that case. □

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, $\pi_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi)$, 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 $f(\boldsymbol{x}|\boldsymbol{\theta})$ to denote the $\mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ density of $\boldsymbol{X}$.)

LEMMA 2.1. *If, for all $\boldsymbol{x} \in \mathbb{R}^p$, the marginal distribution*

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$
$$= \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi)\pi_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi)d\boldsymbol{\theta}d\boldsymbol{\mu}d\boldsymbol{\Sigma}_\pi \tag{2.12}$$

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

$$m(\boldsymbol{x}) \propto e^{-||\boldsymbol{x}||^2/2} \int e^{\boldsymbol{\theta}\cdot\boldsymbol{x}} e^{-||\boldsymbol{\theta}||^2/2}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$
$$\propto e^{-||\boldsymbol{x}||^2/2} h(\boldsymbol{x}),$$

where $h(\boldsymbol{x})$ is the Laplace transform of $e^{-||\boldsymbol{\theta}||^2/2}\pi(\boldsymbol{\theta})$. As $m(\boldsymbol{x})$ is finite for every $\boldsymbol{x}$, it follows from Corollary 2.6 of Brown (1986, p. 38) that all derivatives of m exist at every $\boldsymbol{x} \in \mathbb{R}^p$. 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). □

The following lemmas give conditions under which $m(\boldsymbol{x})$ 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 $\boldsymbol{\Sigma}_\pi = \sigma_\pi^2 \boldsymbol{I}$. If, for some K,*

$$\int_0^K \pi_2^2(\sigma_\pi^2)d\sigma_\pi^2 < \infty, \tag{2.13}$$

and

$$\int_K^\infty \frac{1}{(\sigma_\pi^2)^{(p-\ell+m)/2}} \cdot \pi_2^2(\sigma_\pi^2)d\sigma_\pi^2 < \infty, \tag{2.14}$$

then $m(\boldsymbol{x}) < \infty$ for all $\boldsymbol{x}$.

NOTE: The conditions of Lemma 2.2 are satisfied if, for some $\varepsilon > 0$, $K_1 > 0$, and $K_2 > 0$,

$$\pi_2^2(\sigma_\pi^2) < \frac{K_1}{K_2 + (\sigma_\pi^2)^{[1-\frac{1}{2}(p-\ell+m)+\varepsilon]}}. \tag{2.15}$$

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

$$p > 2 + \ell - m. \tag{2.16}$$

LEMMA 2.3. *Consider the situation of Example 4, "Special Case." Then, if $p > 2+\ell$, $m(\boldsymbol{x}) < \infty$ for all $\boldsymbol{x}$.*

Proof. A straightforward calculation yields

$$m(\boldsymbol{x}) \propto \int_1^\infty \xi^{-(p-\ell)/2} e^{-\frac{1}{2\xi}\boldsymbol{x}^t \boldsymbol{D} \boldsymbol{x}} d\xi, \tag{2.17}$$

where

$$\boldsymbol{D} = \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{y}^t[\boldsymbol{y}\boldsymbol{C}^{-1}\boldsymbol{y}^t]^{-1}\boldsymbol{y}\boldsymbol{C}^{-1}. \tag{2.18}$$

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), $\boldsymbol{\delta}^{HB}$. 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 $\boldsymbol{\Sigma} = \sigma_\pi^2 \boldsymbol{I}_p$.

2.3.1 CALCULATIONAL FORMULAE

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

$$\begin{aligned} \boldsymbol{\delta}^{HB}(\boldsymbol{x}) &= E^{\pi(\boldsymbol{\theta}|\boldsymbol{x})}[\boldsymbol{\theta}] \\ &= E^{\pi_2(\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi|\boldsymbol{x})}[\boldsymbol{\delta}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi)] \end{aligned} \tag{2.19}$$

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

$$\boldsymbol{\delta}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi) = \boldsymbol{x} - \boldsymbol{\Sigma}\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{2.20}$$

and

$$\pi_2(\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi|\boldsymbol{x}) \propto (\det \boldsymbol{W})^{1/2} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{W}(\boldsymbol{x}-\boldsymbol{\mu})\}\pi_2(\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi). \tag{2.21}$$

Note that $\boldsymbol{\delta}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi)$ is the conditional mean of $\boldsymbol{\theta}$ given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\pi$. This decomposition can be calculationally advantageous when $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\pi$ have low dimensional distributions; in that case, the calculation of (2.19) requires only low dimensional integration. Also, when $\boldsymbol{\mu}$ 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, $\boldsymbol{\mu} = \beta\mathbf{1}$ and β has a normal distribution. Then (2.19) becomes

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

with

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

and

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

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

$$\delta^{HB}(\boldsymbol{x}) = \boldsymbol{x} - E^{\pi_2^2(\sigma_\pi^2|\boldsymbol{x})}[\frac{\sigma^2}{\sigma^2+\sigma_\pi^2}](\boldsymbol{x} - \overline{x}\mathbf{1}), \tag{2.24}$$

where

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

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

Note that δ^{HB} is defined even for $p = 3$, as long as $A < \infty$; when $A = \infty$, so $m = 0$, δ^{HB} does not exist for $p = 3$ (see Lemma 2.2). Thus, when $A < \infty$, δ^{HB} defines an exchangeable shrinkage estimator when $p = 3$, while (1.2) requires $p \geq 4$. Furthermore, δ^{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 θ_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

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

Thus, significant practical gains when $p = 3$ will only be available if subjective information about β 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 $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, $\pi_1(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma_\pi^2)$ is $\mathcal{N}_p(\boldsymbol{y\beta}, \sigma_\pi^2\boldsymbol{I}_p)$, and $\boldsymbol{\beta}$ is $\mathcal{N}_\ell(\boldsymbol{\beta}^0, \boldsymbol{A})$, it is shown in Berger (1985, pp. 190–192) that

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

where

$$\delta(\boldsymbol{x}|\sigma_\pi^2) = \boldsymbol{x} - \boldsymbol{\Sigma}\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{y}\hat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{y}\boldsymbol{U}\boldsymbol{A}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0), \tag{2.27}$$
$$\boldsymbol{W} = (\boldsymbol{\Sigma} + \sigma_\pi^2 \boldsymbol{I}_p)^{-1}, \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y})^{-1}\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{x},$$
$$\boldsymbol{U} = (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y} + \boldsymbol{A}^{-1})^{-1},$$

$$\pi_2^2(\sigma_\pi^2|\boldsymbol{x}) \propto m(\boldsymbol{x}|\sigma_\pi^2)\pi_2^2(\sigma_\pi^2), \tag{2.28}$$

$$m(\boldsymbol{x}|\sigma_\pi^2) = \frac{\exp\{-\frac{1}{2}[(\boldsymbol{x} - \boldsymbol{y}\hat{\boldsymbol{\beta}})^t\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{y}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^t(\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y})\boldsymbol{U}\boldsymbol{A}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)]\}}{[\det \boldsymbol{W}]^{-1/2}[\det \boldsymbol{U}]^{1/2}}. \tag{2.29}$$

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

Example 3 (Case 2, continued). Consider the situation above, except that now $\boldsymbol{\beta} \sim \mathcal{T}_\ell(\alpha, \boldsymbol{\beta}^0, \boldsymbol{A})$. As in Generalization II of Appendix I, we can use the representation of the $\mathcal{T}_\ell(\alpha, \boldsymbol{\beta}^0, \boldsymbol{A})$ distribution as a Gamma $(\frac{2}{\alpha}, \frac{\alpha}{2})$ (denoted π_3) mixture of normals, to derive analogous expressions for $\boldsymbol{\delta}^{HB}$. Indeed, one need only replace $\boldsymbol{A}^{-1}$ by $\lambda\boldsymbol{A}^{-1}$ in (2.27) and (2.28) (call the resulting expressions $\boldsymbol{\delta}(\boldsymbol{x}|\lambda, \sigma_\pi^2)$ and $m(\boldsymbol{x}|\lambda, \sigma_\pi^2)$, respectively), and define

$$\pi_2^2(\lambda, \sigma_\pi^2|\boldsymbol{x}) = m(\boldsymbol{x}|\lambda, \sigma_\pi^2)\lambda^{m/2}\pi_3(\lambda)\pi_2^2(\sigma_\pi^2)$$

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

$$\boldsymbol{\delta}^{HB}(\boldsymbol{x}) = E^{\pi_2^2(\lambda, \sigma_\pi^2|\boldsymbol{x})}[\boldsymbol{\delta}(\boldsymbol{x}|\lambda, \sigma_\pi^2)]. \tag{2.30}$$

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

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

$$\boldsymbol{\delta}^{HB}(\boldsymbol{x}) = \boldsymbol{x} - h_{(p-\ell-2)}(||\boldsymbol{x}||_*^2)\boldsymbol{\Sigma}\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{P}\boldsymbol{x}), \tag{2.31}$$

where

$$\boldsymbol{P} = \boldsymbol{y}^t[\boldsymbol{y}\boldsymbol{C}^{-1}\boldsymbol{y}^t]^{-1}\boldsymbol{y}\boldsymbol{C}^{-1}, \tag{2.32}$$

$$||\boldsymbol{x}||_*^2 = \boldsymbol{x}\boldsymbol{C}^{-1}(\boldsymbol{I}_p - \boldsymbol{P})\boldsymbol{x}, \tag{2.33}$$

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 $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, is

$$\boldsymbol{\delta}^{HB}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\Sigma}\nabla \log m(\boldsymbol{x}), \tag{2.34}$$

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

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

so that

$$m(\boldsymbol{x}) = \int m(\boldsymbol{x}|\boldsymbol{\mu})\pi_2^1(\boldsymbol{\mu})d\boldsymbol{\mu}.$$

Then

$$\begin{aligned}
\boldsymbol{\delta}^{HB}(\boldsymbol{x}) &= \boldsymbol{x} + \boldsymbol{\Sigma}\frac{\nabla m(\boldsymbol{x})}{m(\boldsymbol{x})} \\
&= \boldsymbol{x} + \boldsymbol{\Sigma}\frac{\int [\nabla m(\boldsymbol{x}|\boldsymbol{\mu})]\pi_2^1(\boldsymbol{\mu})d\boldsymbol{\mu}}{m(\boldsymbol{x})} \\
&= \int \boldsymbol{\delta}(\boldsymbol{x}|\boldsymbol{\mu})\pi_2^1(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu},
\end{aligned} \tag{2.35}$$

where

$$\boldsymbol{\delta}(\boldsymbol{x}|\boldsymbol{\mu}) = \boldsymbol{x} + \boldsymbol{\Sigma}\nabla \log m(\boldsymbol{x}|\boldsymbol{\mu}), \tag{2.36}$$

$$\pi_2^1(\boldsymbol{\mu}|\boldsymbol{x}) = \frac{m(\boldsymbol{x}|\boldsymbol{\mu})\pi_2^1(\boldsymbol{\mu})}{m(\boldsymbol{x})}. \tag{2.37}$$

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

3. ESTIMATED ACCURACY AND LOSS

What error measures are to be associated with the hierarchical Bayes estimator $\boldsymbol{\delta}^{HB}$? 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 $\boldsymbol{\delta}^{HB}$ is given by (2.19) and the posterior covariance matrix is

$$\boldsymbol{V}^{HB}(\boldsymbol{x}) = E^{\pi_2(\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi|\boldsymbol{x})}[\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{\Sigma} + (\boldsymbol{\delta}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi) - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))(\boldsymbol{\delta}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi) - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))^t] \quad (3.1)$$

where $\boldsymbol{\delta}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi)$ is given by (2.20) and $\pi_2(\boldsymbol{\mu},\boldsymbol{\Sigma}_\pi|\boldsymbol{x})$ 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

$$\begin{aligned}\rho^{HB}(\boldsymbol{x}) &= E^{\pi(\boldsymbol{\theta}|\boldsymbol{x})}[(\boldsymbol{\theta} - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))^t\boldsymbol{Q}(\boldsymbol{\theta} - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))] \\ &= \text{tr}\ (\boldsymbol{V}^{HB}(\boldsymbol{x})\boldsymbol{Q}). \end{aligned} \quad (3.2)$$

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

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

$$\begin{aligned}\boldsymbol{V}^{HB}(\boldsymbol{x}) = E^{\pi_2^2(\sigma_\pi^2|\boldsymbol{x})}[&\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{y}\boldsymbol{U}\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{\Sigma} \\ &+ (\boldsymbol{\delta}(\boldsymbol{x}|\sigma_\pi^2) - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))(\boldsymbol{\delta}(\boldsymbol{x}|\sigma_\pi^2) - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))^t], \end{aligned} \quad (3.3)$$

where $\boldsymbol{\delta}(\boldsymbol{x}|\sigma_\pi^2)$, $\boldsymbol{W}$, $\boldsymbol{U}$, and $\pi_2^2(\sigma_\pi^2|\boldsymbol{x})$ are given by (2.27) through (2.29).

For Case 2, the same formula holds, but with $\lambda\boldsymbol{A}^{-1}$ replacing $\boldsymbol{A}^{-1}$ in $\boldsymbol{U}$ (and elsewhere), $\boldsymbol{\delta}(\boldsymbol{x}|\lambda,\sigma_\pi^2)$ and $\pi_2^2(\sigma_\pi^2,\lambda|\boldsymbol{x})$ replacing $\boldsymbol{\delta}(\boldsymbol{x}|\sigma_\pi^2)$ and $\pi_2^2(\sigma_\pi^2|\boldsymbol{x})$, and $\boldsymbol{\delta}^{HB}(\boldsymbol{x})$ 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

$$\begin{aligned}\boldsymbol{V}^{HB}(\boldsymbol{x}) = \boldsymbol{\Sigma} &- h_{(p-\ell-2)}(||\boldsymbol{x}||_*^2)\boldsymbol{\Sigma}\boldsymbol{C}^{-1}\boldsymbol{\Sigma} \\ &+ g_{(p-\ell-2)}(||\boldsymbol{x}||_*^2)\boldsymbol{\Sigma}\boldsymbol{C}^{-1}(\boldsymbol{I}_p - \boldsymbol{P})\boldsymbol{x}\boldsymbol{x}^t(\boldsymbol{I} - \boldsymbol{P})^t\boldsymbol{C}^{-1}\boldsymbol{\Sigma}; \end{aligned} \quad (3.5)$$

here $||\boldsymbol{x}||_*^2$ and $\boldsymbol{P}$ are defined in (2.33) and (2.32), while h_m and g_m are defined in Appendix III.

□

5

3.1.2 Theoretical formulae

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

$$\boldsymbol{V}^{HB}(\boldsymbol{x}) = \boldsymbol{\Sigma} + \boldsymbol{\Sigma}\frac{\boldsymbol{H}_m(\boldsymbol{x})}{m(\boldsymbol{x})}\boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\nabla \log m(\boldsymbol{x}))(\nabla \log m(\boldsymbol{x}))^t \boldsymbol{\Sigma}. \tag{3.6}$$

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

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

$$\rho^{HB}(\boldsymbol{x}) = \ tr(\boldsymbol{Q}\boldsymbol{\Sigma}) + \frac{1}{m(\boldsymbol{x})}\ tr(\boldsymbol{H}_m(\boldsymbol{x})\tilde{\boldsymbol{Q}}) - (\nabla \log m(\boldsymbol{x}))^t \tilde{\boldsymbol{Q}}(\nabla \log m(\boldsymbol{x})), \tag{3.7}$$

where

$$\tilde{\boldsymbol{Q}} = \boldsymbol{\Sigma}\boldsymbol{Q}\boldsymbol{\Sigma}. \tag{3.8}$$

3.2 UNBIASED ESTIMATORS OF ACCURACY

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

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

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

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

here $\mathcal{D}$ 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 $\boldsymbol{\delta}$*, defined as

$$\boldsymbol{V}(\boldsymbol{\theta}, \boldsymbol{\delta}) = E_{\boldsymbol{\theta}}[(\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{X}))(\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{X}))^t]. \tag{3.10}$$

While dominance of one estimator over another according to this criterion is rare, an unbiased estimator of $\boldsymbol{V}(\boldsymbol{\theta}, \boldsymbol{\delta}^{HB})$ can be used as a frequentist version of $\boldsymbol{V}^{HB}(\boldsymbol{x})$; i.e., it can be used as an estimated "accuracy matrix" and to calculate the unbiased estimate of risk.

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

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

for all i. Then

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

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

$$\hat{\boldsymbol{V}}_{\delta^{HB}}(\boldsymbol{x}) = \boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}\frac{\boldsymbol{H}_m(\boldsymbol{x})}{m(\boldsymbol{x})}\boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\nabla \log m(\boldsymbol{x}))(\nabla \log m(\boldsymbol{x}))^t \boldsymbol{\Sigma}. \tag{3.12}$$

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 $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{HB})$ is given by*

$$\hat{R}_{\delta^{HB}}(\boldsymbol{x}) = tr(\boldsymbol{Q\Sigma}) + \frac{2}{m(\boldsymbol{x})} tr(\boldsymbol{H}_m(\boldsymbol{x})\tilde{\boldsymbol{Q}}) - (\nabla \log m(\boldsymbol{x}))^t \tilde{\boldsymbol{Q}}(\nabla \log m(\boldsymbol{x})), \tag{3.13}$$

where $\tilde{\boldsymbol{Q}} = \boldsymbol{\Sigma Q \Sigma}$.

Proof. Follows immediately from Proposition 3.3, since

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^{HB}) = \text{tr}(\boldsymbol{Q V}(\boldsymbol{\theta}, \boldsymbol{\delta}^{HB})). \quad \square$$

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

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

and

$$\hat{R}_{\delta^{HB}}(\boldsymbol{x}) = 2\rho^{HB}(\boldsymbol{x}) - \text{tr}(\boldsymbol{Q\Sigma}) + (\boldsymbol{x} - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))\boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))^t. \tag{3.15}$$

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 δ^{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 δ_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 ρ^{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 δ^{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 ρ^{HB} and $\hat{R}_{\delta^{HB}}$) will be used in exactly the same way: to convey the possible error in δ^{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 δ^{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 $\boldsymbol{V}^{HB}$ and $\hat{\boldsymbol{V}}_{\delta^{HB}}$ will allow for easier comparison. We also restrict attention to the $\ell = 0$ case, with $\boldsymbol{C}$ 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 $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{A})$ prior, $\boldsymbol{\mu}$ and $\boldsymbol{A}$ being subjectively specified location and "scale" factors for $\boldsymbol{\theta}$.

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

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

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

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

so that

$$\begin{aligned}
\boldsymbol{\delta}^{HB}(\boldsymbol{x}) &= \boldsymbol{x} - h(||\boldsymbol{x}||^2)\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \\
\boldsymbol{V}^{HB}(\boldsymbol{x}) &= \boldsymbol{\Sigma} - h(||\boldsymbol{x}||^2)\boldsymbol{B}\boldsymbol{B}^t + g(||\boldsymbol{x}||^2)\boldsymbol{B}\boldsymbol{z}\boldsymbol{z}^t\boldsymbol{B}^t, \\
\hat{\boldsymbol{V}}_{\delta^{HB}}(\boldsymbol{x}) &= \boldsymbol{\Sigma} - 2h(||\boldsymbol{x}||^2)\boldsymbol{B}\boldsymbol{B}^t - (h^2(||\boldsymbol{x}||^2) + 2g(||\boldsymbol{x}||^2))\boldsymbol{B}\boldsymbol{z}\boldsymbol{z}^t\boldsymbol{B}^t, \\
\rho^{HB}(\boldsymbol{x}) &= \operatorname{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}) - h(||\boldsymbol{x}||^2)\operatorname{tr}(\boldsymbol{Q}\boldsymbol{B}\boldsymbol{B}^t) + g(||\boldsymbol{x}||^2)\boldsymbol{z}^t\boldsymbol{B}^t\boldsymbol{Q}\boldsymbol{B}\boldsymbol{z}, \\
\hat{R}_{\delta^{HB}}(\boldsymbol{x}) &= \operatorname{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}) - 2h(||\boldsymbol{x}||^2)\operatorname{tr}(\boldsymbol{Q}\boldsymbol{B}\boldsymbol{B}^t) + (h^2(||\boldsymbol{x}||^2) + 2g(||\boldsymbol{x}||^2))\boldsymbol{z}^t\boldsymbol{B}^t\boldsymbol{Q}\boldsymbol{B}\boldsymbol{z}.
\end{aligned}$$

3.3.1 Comparison of variances

I. Small values of $||\boldsymbol{x}||^2$

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 $||\boldsymbol{x}||^2$ is small,

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

$$\hat{\boldsymbol{V}}_{\delta^{HB}}(\boldsymbol{x}) \cong \boldsymbol{\Sigma} - 2\boldsymbol{B}\boldsymbol{B}^t = \boldsymbol{\Sigma} - 2\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\boldsymbol{\Sigma}.$$

This exposes a potential problem with $\hat{V}_{\delta^{HB}}$, since $\hat{V}_{\delta^{HB}}$ will have negative eigenvalues unless $\boldsymbol{\Sigma} \leq \boldsymbol{A}$. 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 $||\boldsymbol{x}||^2$,

$$\boldsymbol{\delta}^{HB}(\boldsymbol{x}) \cong \boldsymbol{x} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}),$$

which happens to be the posterior mean w.r.t. a conjugate $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{A})$ prior. For this conjugate prior, $\boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}\boldsymbol{\Sigma}$ 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 $||\boldsymbol{x}||^2$

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

$$\boldsymbol{V}^{HB}(\boldsymbol{x}) \cong \boldsymbol{\Sigma} - \frac{(p-2)}{||\boldsymbol{x}||^2}\boldsymbol{B}\boldsymbol{B}^t + \frac{2(p-2)}{||\boldsymbol{x}||^4}\boldsymbol{B}\boldsymbol{z}\boldsymbol{z}^t\boldsymbol{B}^t,$$
$$\hat{\boldsymbol{V}}_{\delta^{HB}}(\boldsymbol{x}) \cong \boldsymbol{\Sigma} - \frac{2(p-2)}{||\boldsymbol{x}||^2}\boldsymbol{B}\boldsymbol{B}^t + \frac{(p^2-4)}{||\boldsymbol{x}||^4}\boldsymbol{B}\boldsymbol{z}\boldsymbol{z}^t\boldsymbol{B}^t.$$

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

Note next that

$$\hat{\boldsymbol{V}}_{\delta^{HB}}(\boldsymbol{x}) - \boldsymbol{V}^{HB}(\boldsymbol{x}) \cong \frac{(p-2)}{||\boldsymbol{x}||^2}\left[-\boldsymbol{B}\boldsymbol{B}^t + p\boldsymbol{B}\left(\frac{\boldsymbol{z}}{|\boldsymbol{z}|}\right)\left(\frac{\boldsymbol{z}}{|\boldsymbol{z}|}\right)^t\boldsymbol{B}^t\right].$$

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

III. Moderate $||\boldsymbol{x}||^2$ and large p

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

$$||\boldsymbol{x}||^2 = (\boldsymbol{x} - \boldsymbol{\mu})^t(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \cong p.$$

Indeed, as $p \to \infty$, $||\boldsymbol{x}||^2/p$ would then converge to 1.

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

$$V^{HB}(\boldsymbol{x}) \cong \boldsymbol{\Sigma} - \boldsymbol{B}\boldsymbol{B}^t + (2 - \frac{4}{\pi})\boldsymbol{B}\tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}^t\boldsymbol{B}^t,$$
$$\hat{V}_{\delta HB}(\boldsymbol{x}) \cong \boldsymbol{\Sigma} - 2\boldsymbol{B}\boldsymbol{B}^t + p\boldsymbol{B}\tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}^t\boldsymbol{B}^t,$$

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

Interestingly, this exhibits features of both the "small $||\boldsymbol{x}||^2$" and "large $||\boldsymbol{x}||^2$" cases simultaneously. To see this, let

$$\boldsymbol{w}_{(1)} = \frac{\boldsymbol{B}\tilde{\boldsymbol{z}}}{|\boldsymbol{B}\tilde{\boldsymbol{z}}|} = \frac{\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}(\boldsymbol{x} - \boldsymbol{\mu})}{|\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}(\boldsymbol{x} - \boldsymbol{\mu})|}$$

and $\{\boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}, \ldots, \boldsymbol{w}_{(p)}\}$ be an orthonormal basis. Then the "variances" of the "contrasts" $\boldsymbol{w}_{(i)}(\boldsymbol{\theta} - \boldsymbol{\delta}^{HB}(\boldsymbol{x}))$ are, for $i \geq 2$,

$$\boldsymbol{w}_{(i)}^t V^{HB}(\boldsymbol{x})\boldsymbol{w}_{(i)} = \boldsymbol{w}_{(i)}^t(\boldsymbol{\Sigma} - \boldsymbol{B}\boldsymbol{B}^t)\boldsymbol{w}_{(i)},$$
$$\boldsymbol{w}_{(i)}^t \hat{V}_{\delta HB}(\boldsymbol{x})\boldsymbol{w}_{(i)} = \boldsymbol{w}_{(i)}^t(\boldsymbol{\Sigma} - 2\boldsymbol{B}\boldsymbol{B}^t)\boldsymbol{w}_{(i)},$$

and, for $i = 1$,

$$\boldsymbol{w}_{(1)}^t V^{HB}(\boldsymbol{x})\boldsymbol{w}_{(1)} = \boldsymbol{w}_{(1)}^t(\boldsymbol{\Sigma} - \boldsymbol{B}\boldsymbol{B}^t)\boldsymbol{w}_{(1)} + (2 - \frac{4}{\pi})\tilde{\boldsymbol{z}}^t\boldsymbol{B}^t\boldsymbol{B}\tilde{\boldsymbol{z}},$$
$$\boldsymbol{w}_{(1)}^t \hat{V}_{\delta HB}(\boldsymbol{x})\boldsymbol{w}_{(1)} = \boldsymbol{w}_{(1)}^t(\boldsymbol{\Sigma} - 2\boldsymbol{B}\boldsymbol{B}^t)\boldsymbol{w}_{(1)} + p\tilde{\boldsymbol{z}}^t\boldsymbol{B}^t\boldsymbol{B}\tilde{\boldsymbol{z}}.$$

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

IV. A NUMERICAL EXAMPLE

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

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

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

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

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

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

σ_i^2 and A_i being the diagonal elements of $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$, 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 $V_k^{HB}(\boldsymbol{x})$ and $\hat{V}_k(\boldsymbol{x})$ for any value of $|\boldsymbol{z}|^2$. For instance, if $|\boldsymbol{z}|^2 = 6$ (recall that $|\boldsymbol{z}|^2 = p$ 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 $V_j^{HB}(\boldsymbol{x})$ for $j \neq i$ are the "conservative"

$$V_j^{HB}(\boldsymbol{x}) = \sigma_j^2 - (.587)\frac{\sigma_j^4}{(\sigma_j^2 + A_j)}$$

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

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

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

$$V_i^{HB}(\boldsymbol{x}) = \sigma_i^2 + (.149)\frac{\sigma_i^4}{(\sigma_i^2 + A_i)},$$

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

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

$$V_6^{HB}(\boldsymbol{x}) = 10 + (.149)\frac{100}{16.5} = 10.90,$$

$$\hat{V}_6(\boldsymbol{x}) = 10 + (2.36)\frac{100}{16.5} = 24.30.$$

To emphasize: in this example if one observes $\boldsymbol{z} = (0,0,0,0,0,\sqrt{6})^t$ (i.e., $\boldsymbol{x} = (0,0,0,0,0,9.95)^t$), then the estimated variance of θ_6 obtained from $\boldsymbol{V}^{HB}$ is 10.90, while that obtained from $\hat{\boldsymbol{V}}_{\delta^{HB}}$ is 24.30. Note that, here,

$$\begin{aligned}\delta_6^{HB}(\boldsymbol{x}) &= x_6 - h(6)\sigma_6^2(\sigma_6^2 + A_6)^{-1}x_6 \\ &= 9.95 - (.587)\frac{10}{16.5}(9.95) = 6.41,\end{aligned}$$

so that δ_6^{HB} shifts $x_6 = 9.95$ about one sample standard deviation ($\sqrt{10}$).

Of some interest is the observation that

$$\begin{aligned}&\sup_{\boldsymbol{z}} \sup_{\boldsymbol{w}:\ |\boldsymbol{w}|=1} \boldsymbol{w}^t(\hat{\boldsymbol{V}}_{\delta^{HB}} - \boldsymbol{V}^{HB})\boldsymbol{w} \\ &\quad = \sup_{|\boldsymbol{z}|} \boldsymbol{e}_6^t(\hat{\boldsymbol{V}}_{\delta^{HB}} - \boldsymbol{V}^{HB})\boldsymbol{e}_6 \\ &\quad = \frac{\sigma_6^4}{(\sigma_6^2 + A_6)} \sup_v[-h(v) + (h^2(v) + g(v))v] \\ &\quad = (6.0606) \sup_v \left[\frac{20}{v} - \frac{9v(10+3v)}{4(4e^{3v/4-4-3v})}\right] \\ &= 13.44\end{aligned}$$

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

3.3.2 Comparison of risks

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

III. Moderate $||\boldsymbol{x}||^2$ and large p

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

$$\begin{aligned}\rho^{HB}(\boldsymbol{x}) &\cong \text{tr } \boldsymbol{\Sigma} - \text{tr } \boldsymbol{B}\boldsymbol{B}^t + (2 - \frac{4}{\pi})\tilde{\boldsymbol{z}}^t\boldsymbol{B}^t\boldsymbol{B}\tilde{\boldsymbol{z}}, \\ \hat{R}_{\delta^{HB}}(\boldsymbol{x}) &\cong \text{tr } \boldsymbol{\Sigma} - 2\text{tr } \boldsymbol{B}\boldsymbol{B}^t + p\tilde{\boldsymbol{z}}^t\boldsymbol{B}^t\boldsymbol{B}\tilde{\boldsymbol{z}}.\end{aligned}$$

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

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

and

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

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

IV. The numerical example

For $\boldsymbol{z} = |\boldsymbol{z}|\boldsymbol{e}_i$,

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

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

For $\boldsymbol{z} = |\boldsymbol{z}|\boldsymbol{e}_6$, these become

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

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

At $\boldsymbol{z} = \sqrt{6}\boldsymbol{e}_6$, $\rho^{HB} = 14.23$ and $\hat{R}_{\delta^{HB}} = 26.83$. This is quite a discrepancy, $\hat{R}_{\delta^{HB}}$ estimating the risk as being almost twice ρ^{HB}.

At the other extreme, for $\boldsymbol{z} = |\boldsymbol{z}|\boldsymbol{e}_1$,

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

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

which at $\boldsymbol{z} = \sqrt{6}\boldsymbol{e}_1$ 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 ρ^{HB}. (Again, we have chosen $|\boldsymbol{z}|^2 = 6$ to make the comparison because it is what we "expect" to observe.)

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

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

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

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

Graphs of $\rho^{HB}(\boldsymbol{x})$ and $\hat{R}_{\delta^{HB}}(\boldsymbol{x})$ for the three cases $\boldsymbol{z} = |\boldsymbol{z}|\boldsymbol{e}_6$, $\boldsymbol{z} = |\boldsymbol{z}|\boldsymbol{e}_1$, and $\boldsymbol{z} = |\boldsymbol{z}|\ (1,1,1,1,1,1)^t/\sqrt{6}$, are given as functions of $|\boldsymbol{z}|$ in Figure 2. They are labelled ρ_1, ρ_2, ρ_3 and $\hat{R}_1, \hat{R}_2, \hat{R}_3$, respectively. Note that the $\hat{R}_i \to -0.7$ as $|\boldsymbol{z}| \to 0$, and are always substantially smaller than the corresponding ρ_i for small $|\boldsymbol{z}|$.

3.3.3 Discussion

The differences between $\hat{V}_{\delta^{HB}}$ or $\hat{R}_{\delta^{BH}}$ and V^{HB} or ρ^{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 $\boldsymbol{x} = (0,0,0,0,0,10)^t$ is about 14. The large $\hat{R}_{\delta^{HB}}(\boldsymbol{x}) \cong 27$ for this $\boldsymbol{x}$ 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 $\boldsymbol{x}$ 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 $(x_1,\ldots,x_5)^t$ to the prior beliefs about $(\theta_1,\ldots,\theta_5)^t$ overcomes the "bad" fit of x_6 to the prior belief about θ_6 (recall that $\mu_6 = 0$ and $\sqrt{A_6} = \sqrt{6.5} = 2.55$), so that the Bayesian estimator will substantially shrink towards $\boldsymbol{\mu} = \mathbf{0}$. 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 θ_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 $\rho^{HB}(\boldsymbol{x})$ would cause us to investigate the robustness of $\boldsymbol{\delta}^{HB}$ more carefully.

Of course, we are not considering the report of $R(\boldsymbol{\theta},\boldsymbol{\delta}^{HB})$, but instead the report of $\hat{R}_{\delta^{HB}}(\boldsymbol{x})$ (or $\hat{V}_{\delta^{HB}}(\boldsymbol{x})$). And we have identified a seemingly systematic problem with the latter: when $||\boldsymbol{x}||$ is small, $\hat{R}_{\delta^{HB}}$ or $\hat{V}_{\delta^{HB}}$ seem themselves to be too small (even sometimes negative) while if $||\boldsymbol{x}||^2$ 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 ρ^{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

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

Let $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_M$ be values of $\boldsymbol{\theta}$ minimizing and maximizing $R(\boldsymbol{\theta},\boldsymbol{\delta}^{HB})$. (In the numerical example, $\boldsymbol{\theta}_m = \mathbf{0}$ and $\boldsymbol{\theta}_M \cong (0,0,0,0,0,12)^t$.) For $\boldsymbol{x}$ in the immediate vicinity of $\boldsymbol{\theta}_m$, it must be the case that $\hat{R}_{\delta^{HB}}(\boldsymbol{x})$ is generally less than $R(\boldsymbol{\theta}_m,\boldsymbol{\delta}^{HB})$ or (3.16) will not hold when $\hat{R}_{\delta^{HB}}(\boldsymbol{x})$ is averaged over all $\boldsymbol{x}$. Similarly, for $\boldsymbol{x}$ in the immediate vicinity of $\boldsymbol{\theta}_M$, $\hat{R}_{\delta^{HB}}(\boldsymbol{x})$ must typically exceed $R(\boldsymbol{\theta}_M,\boldsymbol{\delta}^{HB})$ 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 $\boldsymbol{V}^{HB}(\boldsymbol{x})$ (or $\rho^{HB}(\boldsymbol{x})$) is very hard to justify, and is the most serious potential failing of $\hat{\boldsymbol{V}}_{\boldsymbol{\delta}^{HB}}$ and $\hat{R}_{\boldsymbol{\delta}^{HB}}$.

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

4. MINIMAXITY OF $\boldsymbol{\delta}^{HB}$

4.1 ANALYTIC SUFFICIENT CONDITIONS

To show that $\boldsymbol{\delta}^{HB} = \boldsymbol{x} + \boldsymbol{\Sigma}\nabla \log m(\boldsymbol{x})$ is minimax, it is sufficient to show that (see (3.13))

$$\hat{R}_{\boldsymbol{\delta}^{HB}}(\boldsymbol{x}) \leq \operatorname{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}) \qquad \text{for all } \boldsymbol{x}, \tag{4.1}$$

since then $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{HB}) = E_{\boldsymbol{\theta}}\hat{R}_{\boldsymbol{\delta}^{HB}}(\boldsymbol{X}) \leq \operatorname{tr}(\boldsymbol{Q}\boldsymbol{\Sigma})$, the minimax risk for the problem. It is straightforward to show that (4.1) can be rewritten as

$$\tilde{\nabla}(\tilde{\boldsymbol{Q}}\nabla\sqrt{m(\boldsymbol{x})}) \leq 0 \qquad \text{for all } \boldsymbol{x}, \tag{4.2}$$

where $\tilde{\boldsymbol{Q}} = \boldsymbol{\Sigma}\boldsymbol{Q}\boldsymbol{\Sigma}$, and

$$\tilde{\nabla}(\boldsymbol{v}(\boldsymbol{x})) \equiv \sum_{i=1}^{p} \frac{\partial}{\partial x_i} v_i(\boldsymbol{x}).$$

When $\tilde{\boldsymbol{Q}} = \boldsymbol{I}_p$, (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 $\boldsymbol{\delta}^{HB}$. 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, $\boldsymbol{\delta}^{HB}$ is minimax if*

$$\tilde{\nabla}(\tilde{\boldsymbol{Q}}\nabla m(\boldsymbol{x}|\boldsymbol{\mu})) \leq 0 \qquad \textit{for all } \boldsymbol{x} \textit{ and } \boldsymbol{\mu}. \tag{4.3}$$

(See Section 2.3.2 for definition of $m(\boldsymbol{x}|\boldsymbol{\mu})$.)

Proof. Clearly,

$$\tilde{\nabla}(\tilde{\boldsymbol{Q}}\nabla m(\boldsymbol{x})) = \int_{\mathbb{R}^p} \tilde{\nabla}(\tilde{\boldsymbol{Q}}\nabla m(\boldsymbol{x}|\boldsymbol{\mu}))\pi_2^1(\boldsymbol{\mu})d\boldsymbol{\mu},$$

so $\tilde{\nabla}(\tilde{\boldsymbol{Q}}\nabla m(\boldsymbol{x})) \leq 0$. 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 $\boldsymbol{\mu}$. Furthermore, if (4.3) is satisfied, then $\boldsymbol{\delta}^{HB}$ is minimax *regardless of the distribution,* π_2^1, *chosen for* $\boldsymbol{\mu}$ (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 ($\boldsymbol{\mu}$) can be utilized while maintaining complete frequentist justification (minimaxity). In the next section we will discuss conditions on $\pi_2^2(\boldsymbol{\Sigma}_\pi|\boldsymbol{\mu})$ under which (4.3) holds.

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

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

$$\pi_2(\boldsymbol{\mu}, \xi) = \pi_2^1(\boldsymbol{\mu})\pi_2^2(\xi|\boldsymbol{\mu}).$$

We will choose

$$\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{\Sigma}^{-1}. \tag{4.4}$$

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

Theorem 4.2 *If, for all* $\boldsymbol{\mu}$, $\pi_2^2(\xi|\boldsymbol{\mu})$ *is non-decreasing on* $(0,\infty)$, *then* (4.3) *is satisfied.*

Proof. Since (see Section 2.3.1)

$$\boldsymbol{W} = (\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_\pi)^{-1} = \xi^{-1}\boldsymbol{C}^{-1} \quad \text{and} \quad \tilde{\boldsymbol{Q}} = \boldsymbol{\Sigma}\boldsymbol{Q}\boldsymbol{\Sigma} = \boldsymbol{C},$$

calculation yields

$$m(\boldsymbol{x}|\boldsymbol{\mu}) = K\int \xi^{-p/2}\exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2\xi}\right\}\pi_2^2(\xi|\boldsymbol{\mu})d\xi$$

and

$$\tilde{\nabla}(\tilde{\boldsymbol{Q}}\nabla m(\boldsymbol{x}|\boldsymbol{\mu})) = K\int_0^\infty \left\{-\frac{p}{\xi} + \frac{(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\xi^2}\right\}$$
$$\times \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2\xi}\right\}\xi^{-p/2}\pi_2^2(\xi|\boldsymbol{\mu})d\xi.$$

Defining $a = (\boldsymbol{x}-\boldsymbol{\mu})^t C^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2$, condition (4.3) is equivalent to

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

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

$$\psi(a) = -2e^{-a/t_0} t_o^{-p/2} \pi_2^2(t_0|\boldsymbol{\mu}) - 2\int_{t_0}^{\infty} \xi^{-p/2} e^{-a/\xi} \pi_2'(\xi|\boldsymbol{\mu}) d\xi, \tag{4.5}$$

where π_2' denotes the (almost everywhere existing) derivative of $\pi_2^2(\xi|\boldsymbol{\mu})$. (Note that the monotonicity condition on π_2^2 ensures that this integration by parts is valid.) But, since $\pi_2^2(\xi|\boldsymbol{\mu})$ is nondecreasing, $\pi_2'(\xi|\boldsymbol{\mu}) \geq 0$, and the right hand side of (4.5) is clearly negative, completing the proof. □

A natural choice for π_2^2 is $\pi_2^2(\xi|\boldsymbol{\mu}) \equiv 1$ (on a subset $[t_0, \infty)$ of (2.8)). This clearly is nondecreasing, and so (4.3) is satisfied and the resulting $\boldsymbol{\delta}^{HB}$ 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 ξ that was considered was (with $\boldsymbol{\mu}$ being given)

$$\pi_2^2(\xi|\boldsymbol{\mu}) \propto \xi^{-(n+1-p/2)} \qquad \text{on } (1, \infty),$$

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 $\pi_2^2(\xi|\boldsymbol{\mu})$. (Note that, for fixed $\boldsymbol{\mu}$, 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* $\pi_2^2(\xi|\boldsymbol{\mu})$. Unfortunately, (4.3) cannot be satisfied for such π_2^2, as the following lemma shows.

LEMMA 4.3. *If there exists t_1 such that, on (t_1, ∞), $\pi_2^2(\xi|\boldsymbol{\mu})$ is continuous, nonincreasing and nonconstant, then* (4.3) *cannot hold for all $\boldsymbol{x}$.*

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 π_2^2, there exists an interval (b, c), $b > t_0$, such that $\pi_2'(\xi|\boldsymbol{\mu}) < -\varepsilon < 0$ on (b, c). Clearly

$$\begin{aligned}\int_{t_0}^{\infty} \xi^{-p/2} e^{-a/\xi} \pi_2'(\xi|\boldsymbol{\mu}) d\xi &< (-\varepsilon) \int_b^c \xi^{-p/2} e^{-a/\xi} d\xi \\ &\leq (-\varepsilon) c^{-p/2} e^{-a/b} (c - b).\end{aligned}$$

Thus (see (4.5))

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

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 $\boldsymbol{x}$ and $\boldsymbol{\mu}$. For instance, certain π_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 $\boldsymbol{\Sigma}_\pi = \sigma_\pi^2 \boldsymbol{\Sigma}$; then simply set $\boldsymbol{C} = \boldsymbol{\Sigma}$ in Example 4, so that $\sigma_\pi^2 = (\xi - 1)$.

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

Because of the special choice of $\boldsymbol{Q}$ and the special nature of $\pi_2^2(\xi|\boldsymbol{\mu})$ required for the analytic minimaxity proof in Section 4.2, an alternative general method for verifying minimaxity of $\boldsymbol{\delta}^{HB}$ 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 $\hat{R}_{\delta^{HB}}$, so that the numerical problem can be rewritten as showing that

$$\Delta(\boldsymbol{x}) = 2[\text{tr}\,(\boldsymbol{Q}\boldsymbol{\Sigma}) - \rho^{HB}(\boldsymbol{x})] - [\boldsymbol{x} - \boldsymbol{\delta}^{HB}(\boldsymbol{x})]^t \boldsymbol{Q}[\boldsymbol{x} - \boldsymbol{\delta}^{HB}(\boldsymbol{x})] \geq 0. \tag{4.6}$$

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

Numerically minimizing $\Delta(\boldsymbol{x})$ is not necessarily trivial. First of all, calculation of $\boldsymbol{\delta}^{HB}$ and ρ^{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 $\boldsymbol{\delta}^{HB}$ *is* minimax, $\Delta(\boldsymbol{x})$ will converge to its minimum (of zero) as $|\boldsymbol{x}| \to \infty$, so that one has to truncate the minimization algorithm when $\Delta(\boldsymbol{x})$ gets within ε of 0 and $|\boldsymbol{x}|$ is large. (Strictly speaking, one has then only shown that $\boldsymbol{\delta}^{HB}$ is probably ε-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 ε 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 $\pi_2^2(\sigma_\pi^2) = 1$. For sum of squares error loss ($\boldsymbol{Q} = \boldsymbol{I}_7$

in (1.1)), the results in Section 4.2 (and Lemma 2.2) show that $\boldsymbol{\delta}^{HB}$ is minimax. Here, however, the *current* IQ, θ_7, might be of substantially more importance than the previous IQs, so that $\boldsymbol{Q} =$ diag.$\{1,1,1,1,1,1,q\}$ (with $q > 1$) might be deemed to be more reasonable. We will investigate the minimaxity of $\boldsymbol{\delta}^{HB}$ for such $\boldsymbol{Q}$, using the numerical method.

For this example, algebra yields

$$\begin{aligned}(100)^{-2}\Delta(\boldsymbol{x}) &= 2(E_1 - 225E_5)(6+q) + (E_1^2 - 2E_3)[s^2 + (q-1)(x_7 - \overline{x})^2] \\ &+ (E_2^2 - 2E_4)(\overline{x} - 100)^2(6+q) + 2(E_1E_2 - 2E_5)(x_7 - \overline{x})(\overline{x} - 100)(q-1), \end{aligned} \tag{4.7}$$

where $s^2 = ||\boldsymbol{x} - \overline{x}\mathbf{1}||^2$ and the E_i are the expectations with respect to $\pi_2^2(\sigma_\pi^2|\boldsymbol{x})$ of, respectively, $(100+\sigma_\pi^2)^{-1}$, $(1875+\sigma_\pi^2)^{-1}$, $(100+\sigma_\pi^2)^{-2}$, $(1875+\sigma_\pi^2)^{-2}$, and $(1875+\sigma_\pi^2)^{-1}$ $(100+\sigma_\pi^2)^{-1}$. From (4.7) (and (2.23)) it is not hard to show that $\Delta(\boldsymbol{x})$ actually depends only on the three quantities x_7, $\overline{x}^* = \sum_{i=1}^{6} x_i/6$, and $s^{*2} = \sum_{i=1}^{6}(x_i - \overline{x}^*)^2$, 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 $\Delta(\boldsymbol{x})/100$, as a function of q. (The accuracy of the minima is about 0.05.) For $q \le 1.7$, the minimum is zero, indicating that $\boldsymbol{\delta}^{HB}$ is minimax for such q. For $q > 1.7$, however, $\boldsymbol{\delta}^{HB}$ 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 occassionally 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 $\Delta(\boldsymbol{x})$) on the degree of non-minimaxity (since $R(\boldsymbol{\theta},\boldsymbol{\delta}) -$ tr $(\boldsymbol{Q}\not{\boldsymbol{\Sigma}}) \le -\inf\Delta(\boldsymbol{x})$).

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

$$R(\boldsymbol{\theta},\boldsymbol{\delta}^{HB}) = \text{ tr } (\boldsymbol{Q}\not{\boldsymbol{\Sigma}}) - E_\theta\Delta(\boldsymbol{X});$$

the presence of the additional expectation over $\boldsymbol{X}$, in calculation of $R(\boldsymbol{\theta},\boldsymbol{\delta}^{HB})$, so complicates the numerical problem as to make it unmanageable on a routine basis for complicated $\boldsymbol{\delta}^{HB}$. 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

Angers, J. F. (1987). Development of robust Bayes estimators for a multivariate normal mean, Ph.D. Thesis, Purdue University.

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

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

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8**, 716–761.

Berger, J. (1982a). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. Berger, eds.). Academic Press, New York.

Berger, J. (1982b). Selecting a minimax estimator of a multivariate normal mean. *Ann. Statist.* **10**, 81–92.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer–Verlag, New York.

Berger, J. and Chen, S. Y. (1987). Minimaxity of empirical Bayes estimators derived from subjective hyperpriors. In *Advances in Multivariate Statistical Analysis* (A. K. Gupta, ed.). 1–12, D. Reidel Publishing Company.

Berger, J. and Dey, D. K. (1985). Truncation of shrinkage estimators in the nonsymmetric case. In *Multivariate Estimation* **VI** (P. R. Krishnaiah, ed.). North–Holland, Amsterdam.

Berger, J. and Srinivasan, C. (1978). Generalized Bayes estimators in multivariate problems. *Ann. Statist.* **6**, 783–801.

Bock, M. E. (1988). Shrinkage estimators: Pseudo–Bayes rules for normal mean vectors. In *Statistical Decision Theory and Related Topics* **IV** (S. S. Gupta and J. Berger, eds.). Springer–Verlag, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison–Wesley, Reading.

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

Brown, L. (1986). *Foundations of Statistical Exponential Families.* IMS Monograph Series, Hayward, California.

Brown, L. (1987). An ancillarity paradox which appears in multiple linear regression. Technical

Report, Department of Mathematics, Cornell University, Ithaca.

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

Casella, G. (1985). Condition number and minimax ridge regression estimation. *J. Amer. Statist. Assoc.* **80**, 753–758.

Chen, S. Y. (1983). Restricted risk Bayes estimation. Ph.D. Thesis, Purdue University.

Chen, S. Y. (1988). Restricted risk Bayes estimation for the mean of the multivariate normal distribution. *J. Multivariate Analysis* **24**, 207–217.

Cooley, E. A. and Lin, P. E. (1983). Bayes minimax estimators of a multivariate normal mean, with application to generalized ridge regression. *Commun. Statist. Theor. Meth.* **12**, 2861–2869.

DasGupta, A. and Rubin, H. (1988). Bayesian estimation subject to minimaxity of a multivariate normal mean in the case of a common unknown variance. In *Statistical Decision Theory and Related Topics* **IV** (S. S. Gupta and J. Berger, eds.). Springer–Verlag, New York.

Deely, J. and Lindley, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* **76**, 833–841.

Dickey, J. M. (1968). Three multidimensional integral identities with Bayesian applications. *Ann. Math. Statist.* **39**, 1615–1627.

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. Fienberg and A. Zellner, eds.). North–Holland, Amsterdam.

DuMouchel, W. M. and Harris, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.* **78**, 293–315.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators – Part II: The empirical Bayes case. *J. Amer. Statist. Assoc.* **67**, 130–139.

Faith, R. E. (1978). Minimax Bayes and point estimators of a multivariate normal mean. *J. Multivariate Analysis* **8**, 372–379.

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

Haff, L. R. (1988). The variational form of certain Bayes estimators. Technical Report, Department of Mathematics, University of California, San Diego.

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

George, E. I. (1986b). Combining minimax shrinkage estimators. *J. Amer. Statist. Assoc.* **81**, 437–445.

George, E. I. (1986c). A formal Bayes multiple shrinkage estimator. *Communications in Statistics* A, 2099–2114.

Johnstone, I. (1988). On inadmissibility of some unbiased estimates of loss. In *Statistical Decision Theory and Related Topics* **IV** (S. S. Gupta and J. Berger, eds.). Springer–Verlag, New York.

Judge, G. and Bock, M. E. (1978). *The Statistical Implications of Pre–test and Stein–rule estimators in Econometrics.* North–Holland, Amsterdam.

Li, T. F. (1982). A note on James–Stein and Bayes empirical Bayes estimators. *Commun. Statist.* **A11**, 1029–1043.

Lindley, D. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. B* **34**, 1–41.

Lu, K. L. and Berger, J. (1988a). Estimation of normal means: frequentist estimation of loss. To appear in *Ann. Statist.*.

Lu, K. L. and Berger, J. (1988b). Estimated confidence procedures for multivariate normal means. *J. Statist. Planning and Inference.*

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78**, 47–65.

Smith, A. F. M. (1973). A general Bayesian linear model. *J. Roy. Statist. Soc. B* **35**, 67–75.

Spruill, M. (1986). Some approximate restricted Bayes estimators of a normal mean. *Statistics and Decisions* **4**, 337–351.

Stein, C. (1973). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. Asymp. Stat.*, 345–381.

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

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

Strawderman, W. (1973). Proper Bayes minimax estimators of the multivariate normal mean for the case of common unknown variances. *Ann. Statist.* **1**, 1189–1194.

Zheng, Z. (1982). A class of generalized Bayes minimax estimators. In *Statistical Decision Theory and Related Topics* **III** (S. S. Gupta and J. Berger, eds.). Academic Press, New York.

APPENDIX I

Generalizations of Examples 3 and 4

I. Frequently, we consider the case $\boldsymbol{\Sigma}_\pi = \sigma_\pi^2 \boldsymbol{I}_p$ in the examples. An apparently more general case of considerable interest is that in which correlation among the θ_i is allowed, i.e. in which

$$\boldsymbol{\Sigma}_\pi = \sigma_\pi^2 \boldsymbol{I}_p + \rho \mathbf{1}\mathbf{1}^t.$$

It is of interest to note that this case can be reduced to the $\boldsymbol{\Sigma}_\pi = \sigma_\pi^2 \boldsymbol{I}_p$ case by defining

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

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

$$\begin{aligned}\pi(\boldsymbol{\theta}) &= \int \pi_1(\boldsymbol{\theta}|\boldsymbol{\mu}, \rho, \sigma_\pi^2)\pi_2(\boldsymbol{\mu}, \rho, \sigma_\pi^2) d\boldsymbol{\mu} d\rho d\sigma_\pi^2 \\ &= \int \pi_1^*(\boldsymbol{\theta}|\boldsymbol{\mu}^*, \sigma_\pi^2)\pi_2^*(\boldsymbol{\mu}^*, \sigma_\pi^2) d\boldsymbol{\mu}^* d\sigma_\pi^2;\end{aligned}$$

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

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

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

$$\pi_1(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi) = \int \pi_1^*(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi, \lambda) d\tilde{\pi}_3(\lambda),$$

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

$$\pi(\boldsymbol{\theta}) = \int \pi_1^*(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi)\pi_2^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi) d\boldsymbol{\mu} d\boldsymbol{\Sigma}_\pi,$$

where

$$\pi_2^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi) = \int \pi_2(\boldsymbol{\mu}, \frac{1}{\rho}\boldsymbol{\Sigma}_\pi)\rho\tilde{\pi}_3(\rho) d\rho.$$

APPENDIX II

Proof of Lemma 2.2.

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

$$m(\boldsymbol{x}) = \int_0^\infty m(\boldsymbol{x}|\sigma_\pi^2)\pi_2^2(\sigma_\pi^2)d\sigma_\pi^2,$$

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

$$\int_0^\infty (\det \boldsymbol{W})^{1/2}[\det (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y} + \boldsymbol{A}^{-1})]^{-1/2}\pi_2^2(\sigma_\pi^2)d\sigma_\pi^2 < \infty. \tag{A1}$$

For $0 < \sigma_\pi^2 < K$,

$$(\det \boldsymbol{W}) \leq \det (\boldsymbol{\Sigma}^{-1}), \tag{A2}$$

and

$$\det (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y} + \boldsymbol{A}^{-1}) \geq \det (\boldsymbol{y}^t(\boldsymbol{\Sigma} + K\boldsymbol{I}_p)^{-1}\boldsymbol{y}). \tag{A3}$$

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

$$\int_0^K (\det \boldsymbol{W})^{1/2}[\det (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y} + \boldsymbol{A}^{-1})]^{-1/2}\pi_2^2(\sigma_\pi^2)d\sigma_\pi^2 < \infty.$$

For $\sigma_\pi^2 > K$,

$$\det (\boldsymbol{W}) \leq \det (\sigma_\pi^2\boldsymbol{I}_p)^{-1} = \sigma_\pi^{-2p}, \tag{A4}$$

and

$$\begin{aligned}\det (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y} + \boldsymbol{A}^{-1}) &\geq \det (K'\sigma_\pi^{-2}\boldsymbol{y}^t\boldsymbol{y} + \boldsymbol{A}^{-1}) \\ &= \det (K'\sigma_\pi^{-2}\boldsymbol{y}^t\boldsymbol{y}) \det (\boldsymbol{I}_p + \frac{\sigma_\pi^2}{K'}\boldsymbol{A}^{-1}(\boldsymbol{y}^t\boldsymbol{y})^{-1}) \\ &= \det (K'\sigma_\pi^{-2}\boldsymbol{y}^t\boldsymbol{y})\prod_{i=1}^m(1 + \frac{\sigma_\pi^2}{K'}\rho_i) \\ &\geq K^*(\sigma_\pi^2)^{-(\ell-m)}(\prod_{i=1}^m \rho_i);\end{aligned} \tag{A5}$$

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

$$\int_K^\infty (\det \boldsymbol{W})^{1/2}[\det (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y} + \boldsymbol{A}^{-1})]^{-1/2}\pi_2^2(\sigma_\pi^2)d\sigma_\pi^2 < \infty.$$

Case 2. Using the representation discussed in Section 2.3.1, Example 3 — Case 2, one has

$$m(\boldsymbol{x}) = \int_0^\infty \int_0^\infty m(\boldsymbol{x}|\lambda, \sigma_\pi^2)\pi_2^2(\sigma_\pi^2)\lambda^{m/2}\pi_3(\lambda)d\lambda d\sigma_\pi^2,$$

where $m(\boldsymbol{x}|\lambda, \sigma_\pi^2)$ is given by (2.29) with $\boldsymbol{A}^{-1}$ replaced by $\lambda\boldsymbol{A}^{-1}$, and π_3 is a Gamma $(\frac{2}{\alpha}, \frac{\alpha}{2})$ density. Again, the exponential part of $m(\boldsymbol{x}|\lambda, \sigma_\pi^2)$ is bounded by 1, so it suffices to show that

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

For $K < \sigma_\pi^2 < \infty$, 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

$$\int_K^\infty \int_0^\infty (\det \boldsymbol{W})^{1/2}(\det (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y} + \lambda\boldsymbol{A}^{-1})]^{-1/2}\pi_2^2(\sigma_\pi^2)\lambda^{m/2}\pi_3(\lambda)d\lambda d\sigma_\pi^2 < \infty.$$

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

$$\det (\boldsymbol{y}^t\boldsymbol{W}\boldsymbol{y} + \lambda A^{-1}) \geq \det (\boldsymbol{y}^t(\boldsymbol{\Sigma} + K\boldsymbol{I}_p)^{-1}\boldsymbol{y})\lambda^m \prod_{i=1}^m \rho_i^*,$$

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

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

completing the proof.

□

APPENDIX III

Define

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

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

where

$$H_m(v) = \begin{cases} v^{m/2}\left[\frac{m}{2}!\{e^v - \sum_{i=1}^{(m/2-1)} \frac{v^i}{i!}\}\right]^{-1} & \text{if } m \text{ is even} \\ v^{m/2}\left[\Gamma(\frac{m}{2}+1)\{e^v[2\Phi(\sqrt{2v}) - 1] - \sum_{i=0}^{(m-3)/2} \frac{v^{i+1/2}}{\Gamma(i+3/2)}\}\right]^{-1} & \text{if } m \text{ is odd;} \end{cases}$$

where Φ 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* $||x||^2/p \to 1$ *as* $p \to \infty$. *Then* $h(||x||^2) \to 1$ *and* $pg(||x||^2) \to (2 - \frac{4}{\pi})$.

Proof. The result that $h(||x||^2) \to 1$ follows easily from Lemma 2.1.1 (vi) of Berger (1980). To show that $pg(||x||^2) \to (2-\frac{4}{\pi})$, note first that it is easy to show that any $||x||^2$ such that $||x||^2/p \to 1$ 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)

$$\begin{aligned} vg(v) &= [t_n(v) - r_n^2(v)]/v \\ &= 2\frac{r_n(v)}{v} + (2n - r_n(v))(\frac{r_n(v)}{v} - 1), \end{aligned}$$

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 $(2n)^{-1}r_n(2n) \to 1$ as $n \to \infty$, so that we need only show that (see Berger (1980))

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

as $n \to \infty$. Stirling's formula gives

$$\begin{aligned} \frac{n^i n!}{(n+i)!} &= \frac{n^i e^{-(n+1)}(n+1)^{(n+\frac{1}{2})}\sqrt{2\pi}(1+O(\frac{1}{n}))}{e^{-(n+i+1)}(n+i+1)^{(n+i+\frac{1}{2})}\sqrt{2\pi}(1+O(\frac{1}{n}))} \\ &= (1 - \frac{i+1}{(n+i+1)})^i(1 - \frac{i}{(n+i+1)})^{n+\frac{1}{2}}e^i(1+O(\frac{1}{n})). \end{aligned} \tag{A5}$$

(Within this proof, $O(\cdot)$ and $o(\cdot)$ are to be understood to be uniform in $0 \le i < \infty$.) 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 $i \le n^\alpha$ where $\alpha < 1$,

$$\frac{n^i n!}{(n+i)!} = \left(1 - \frac{i}{(n+i+1)}\right)^{n+i+1} e^i(1+o(1)). \tag{A6}$$

Next note that, for $i \le n^\alpha$ 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 $\alpha > \frac{1}{2}$,

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

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

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

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. □

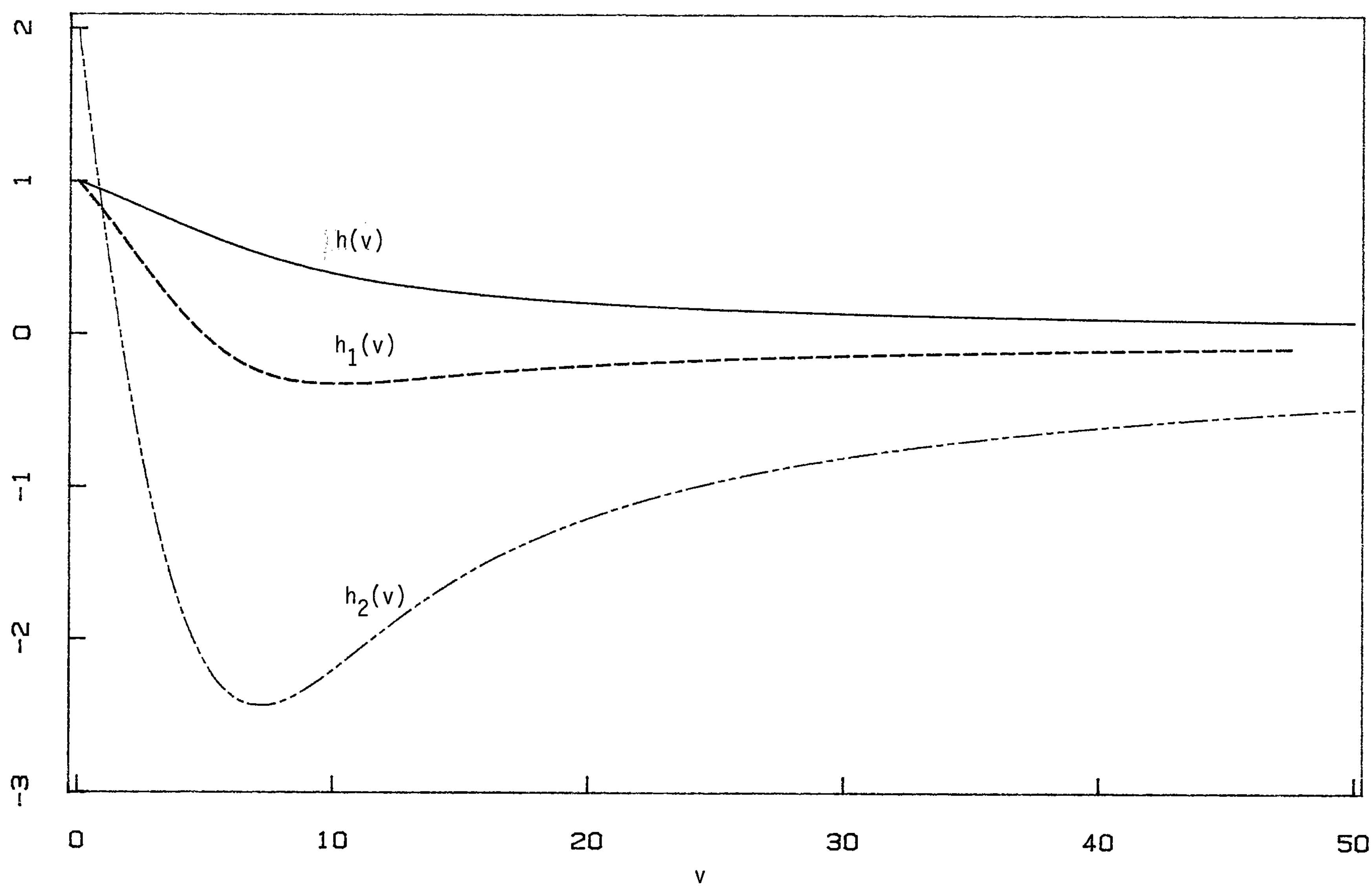

Figure 1. Graphs of $h(v)$, $h_1(v)$, and $h_2(v)$.

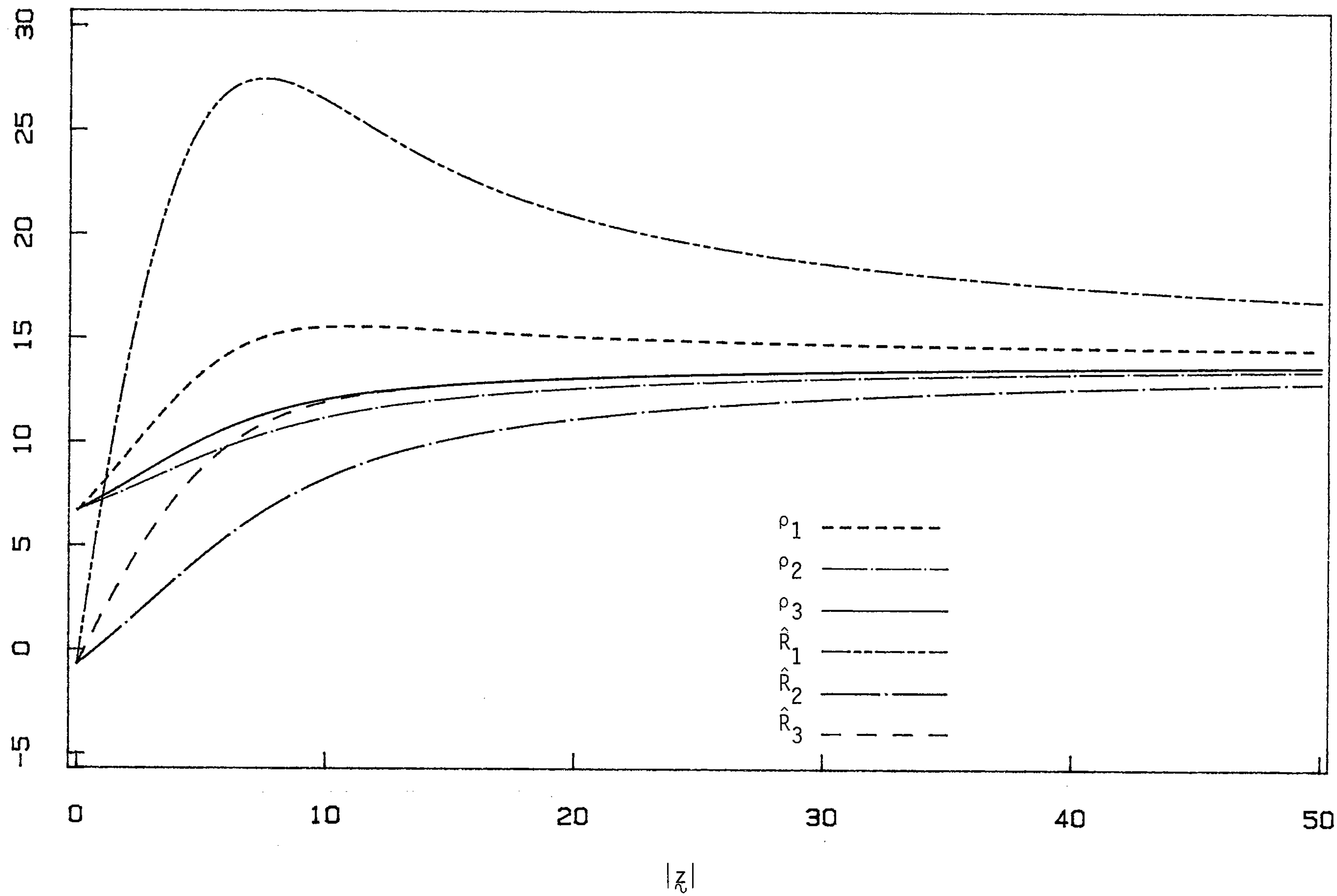

Figure 2. Graphs along various rays of the posterior expected losses, $\rho_i\,(|z|)$, and unbiased estimators of risk, $\hat{R}_i\,(|z|)$; here i=1, 2, and 3 denote the $|z|\,(0,0,0,0,0,1)^t$, $|z|\,(1,0,0,0,0,0)^t$, and $|z|\,(1,1,1,1,1,1)^t/\sqrt{6}$ rays, respectively.

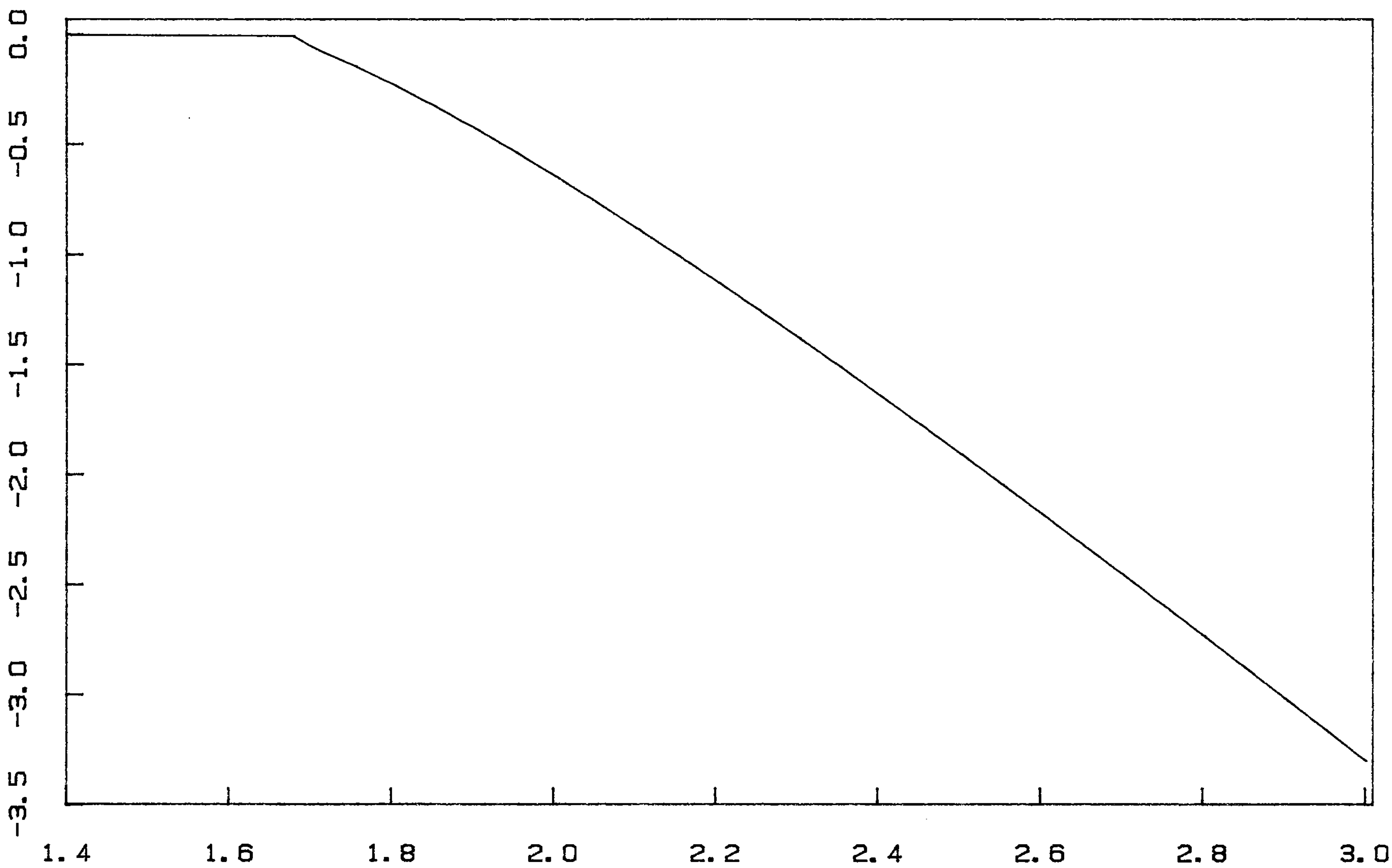

Figure 3. The value of $\inf_x \Delta(x)/100$, as a function of the loss q.