Journal of Statistical Planning and Inference 8 (1983) 143-160
North-Holland

COMBINING COORDINATES IN SIMULTANEOUS ESTIMATION OF NORMAL MEANS*

James O. BERGER

Purdue University, West Lafayette, IN 47907, USA

Dipak K. DEY

University of Kentucky, Lexington, KY 40506, and Texas Tech University, Lubbock, TX 79409, USA

Received 16 February 1982; revised manuscript received 6 October 1982
Recommended by J. Sacks

Abstract: The problem of combining coordinates in Stein-type estimators, when simultaneously estimating normal means, is considered. The question of deciding whether to use all coordinates in one combined shrinkage estimator or to separate into groups and use separate shrinkage estimators on each group is considered. A Bayesian viewpoint is (of necessity) taken, and it is shown that the 'combined' estimator is, somewhat surprisingly, often superior.

AMS 1970 Subject Classification: Primary 62F15; Secondary 62F10, 62F23.

Key words and phrases: Shrinkage estimator; Robust generalized Bayes estimator; Minimax estimator; Bayes risk; Normal prior; Flat prior.

1. Introduction

Let $X=(X_1,\dots,X_k)^t$ have a k-variate normal distribution with mean vector $\theta=(\theta_1,\dots,\theta_k)^t$ and known positive definite covariance matrix Σ. It is desired to estimate θ, using an estimator $\delta(X)=(\delta_1(X),\dots,\delta_k(X))^t$, under a quadratic loss

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

Q being a known positive definite matrix. As usual we will evaluate an estimator δ in terms of its risk function (i.e. expected loss)

$$R(\theta,\delta)=E_\theta L(\theta,\delta(X)).$$

Stein (1955) showed that (for $Q=\Sigma=I_k$) the usual estimator $\delta^0(X)=X$ is inadmissible for $k\geq 3$. Estimators improving upon δ^0 for the above general case have

* Research supported by the National Science Foundation, Grant MCS 78-02300 and MCS 81-01670A1.

been found by a variety of authors. (See Berger (1980, 1982a) for references.) In actually selecting an alternative to δ^0, however, it was shown in Berger (1980) that prior information concerning θ must be taken into account (to choose the region of the parameter space in which significant improvement over δ^0 is to be obtained). In Berger (1980) a robust generalized Bayes estimator was developed using, as inputs a 'prior mean' μ and a 'prior covariance matrix' A. (Alternatively, μ and A can be thought of as specifying an ellipse in the parameter space in which significant improvement over δ^0 is desired.) The estimator is given by

$$\delta^{\mathrm{RB}}(X)=\left(I_k-\frac{r((X-\mu)^{\mathrm{t}}(\Sigma+A)^{-1}(X-\mu))}{(X-\mu)^{\mathrm{t}}(\Sigma+A)^{-1}(X-\mu)}\,\Sigma(\Sigma+A)^{-1}\right)(X-\mu)+\mu, \tag{1.1}$$

where r is a certain increasing function which can be reasonably approximated by

$$r(z)=\min\{k-2,z\}.$$

This estimator was shown to be a very attractive alternative to δ^0, but unfortunately is not necessarily uniformly better than δ^0 (and hence not minimax). For those desiring uniform dominance of δ^0, a minimax Bayes estimator δ^{MB} (to be described later) was developed in Berger (1982a).

For extensive discussion of the above estimators, see Berger (1980) and Berger (1982a). A few brief comments concerning the motivation for their use seems in order, however. First, the estimators are concerned with the situation in which both μ and A must be determined subjectively. In many 'empirical Bayes' types of problems, where the prior information about θ includes beliefs in certain relationships such as exchangeability among the θ_i, the observation X can be used to estimate μ and A or certain facets of them. In many (undoubtedly the majority) of simultaneous estimation problems, however, no such relationships exist, and it then becomes necessary to employ subjective prior inputs. (Even from a non-Bayesian viewpoint, Stein-type estimators must be told the region to shrink towards, which should logically be the region in which θ is expected to lie.)

A complete Bayesian analysis could, of course, be performed, but typical (say, the conjugate prior) Bayes estimators are very nonrobust with respect to the specification of certain features of the prior distribution. For instance, it is very difficult to decide whether a normal or Cauchy functional form for the prior distribution is appropriate, yet this can have a profound effect on the analysis. What are sought, therefore, are robust Bayes estimators which make valuable and significant use of easy to specify parts of the prior distribution, yet are not dependent or sensitive to other features of the prior distribution (such as the exact functional form). In Berger (1982a) it is shown that in doing Stein estimation in typical nonsymmetric situations it is *necessary* to at least specify μ and A, which can be thought of either as specifying an ellipse in which θ is thought to lie or as a prior mean and covariance matrix. (Actually, thinking of μ as simply the 'best guess' for θ, and A as reflecting the 'accuracies' of this guess is better, since one rarely even knows whether or not one's true prior opinions are reflected by a prior distribution which has moments.) Coincidentally (and luckily) μ and A are precisely the features of one's prior beliefs which are relatively easy to specify or elicit. Thus from either a non-Bayesian or a robust Bayesian viewpoint, the problem can be stated as that of developing a reasonable Stein-type estimator incorporating μ and A which is satisfactory to both the non-Bayesian (i.e. has satisfactory $R(\theta,\delta)$) and to the robust Bayesian (i.e. has satisfactory Bayes risk

$$r(\pi,\delta)=E^{\pi}R(\theta,\delta)$$

for all reasonable π with the specified μ and A.) The estimators δ^{RB} and δ^{MB} achieve both of these goals to varying degrees, δ^{MB} being somewhat more attractive to the non-Bayesian (being minimax) and δ^{RB} being somewhat more attractive to the Bayesian (having better Bayes risk).

A major problem (some would say *the* major problem) remaining in Stein estimation is that of deciding how coordinates should be grouped together for use in Stein estimators. To achieve the Stein effect (see Berger (1982b) for justification of the Stein effect in robust Bayesian terms), coordinates must be grouped together, yet it is well known that the Stein effect is most significant when "similar" coordinates are grouped together. To our knowledge, the only previous work on this grouping problem was that of Efron and Morris (1973b). They considered use of the James–Stein estimator in a two group situation, i.e. where the θ_i could be divided into two homogeneous groups which were suspected of having different prior variances. This is definitely an empirical Bayes type of situation (the homogeneity of each group allows estimation of the prior mean and variance) and Efron and Morris developed very reasonable estimators for this situation which acted like the 'combined' James–Stein estimator when the prior variance estimates were close, yet acted like separate James–Stein estimators on each group when the prior variance estimates were substantially different. The situation considered here is substantially different than the Efron and Morris situation, since no symmetries or empirical Bayes relationships in the problem are assumed to exist.

The following notation will be used throughout the paper. Suppose that θ is divided into s groups of sizes $k_1,\dots,k_s$ ($k_i\geq 3$, $i=1,\dots,s$), where if necessary the coordinates are relabeled so that the l-th group is given by (defining $k_0=0$)

$$\theta_{(l)}=(\theta_{k_0+\dots+k_{l-1}+1},\dots,\theta_{k_0+\dots+k_l})^{\mathrm{t}}.$$

Similarly define $X_{(l)}$, and let $\delta_{(l)}^{\mathrm{RB}}$ and $\delta_{(l)}^{\mathrm{MB}}$ be the appropriate estimator for estimating $\theta_{(l)}$ based solely on $X_{(l)}$. The estimator δ^{RB} or δ^{MB} will be called the 'combined estimator', while $\delta^{\mathrm{RBS}}=(\delta_{(1)}^{\mathrm{RB}},\dots,\delta_{(s)}^{\mathrm{RB}})^{\mathrm{t}}$ or $\delta^{\mathrm{MBS}}=(\delta_{(1)}^{\mathrm{MB}},\dots,\delta_{(s)}^{\mathrm{MB}})^{\mathrm{t}}$ will be called the 'separate estimator'.

The first difficulty encountered in pursuing the grouping problem is that of measuring the performance of estimators. We will do so in terms of Bayes risk $r(\pi,\delta)$. A Bayesian would have few qualms with this. A non-Bayesian, however, should also accept this as a measure, since δ^{RB} and δ^{MB} (particularly the latter) are designed to have good frequentist risk $R(\theta,\delta)$. Hence the main question of interest is how *much* improvement over δ^0 is being obtained, and the natural way to measure this is to average $R(\theta,\delta)$ with respect to a prior measure π reflecting the likelihood of the various θ. (All estimators considered will have $R(\theta,\delta)$ which cross, and hence are noncomparable as functions.) The obvious problem is how to choose π.

Note, first of all, that δ^{RB} (but not δ^{MB}) was developed as a generalized Bayes rule with respect to a prior π_0 which incorporates μ and A. The prior π_0 was prejudiced in favor of 'combining', however, being of a form which imposed considerable dependence on the coordinates of θ. The use of such π would artificially prejudice the issue in favor of combining, and if indeed the groups of coordinates are felt to be 'unrelated' (which we will assume is the case and is reflected in the problem by Q, Σ and A being block diagonal with l-th blocks $Q_{(l)}$, $\Sigma_{(l)}$ and $A_{(l)}$ (all $k_l\times k_l$ matrices) respectively), then any π used should reflect this independence (i.e. be a product of priors on each group). Since μ and A (and the group independence) are known facts of the prior distribution we will be concerned with evaluation for a wide range of functional forms for π which incorporate μ, A, and the desired group independence. It is crucial to realize that μ, A, and the group independence are the only prior features we assume known.

In Section 2 the π chosen will be $\mathrm{N}(\mu, A)$. This allows explicit calculation of the Bayes risks of the combined and separate estimators (for both δ^{RB} and δ^{MB}), from which it follows (Theorems 2.1.1 and 2.2.1) that the combined estimators have smaller Bayes risks than the separate estimators. (For technical reasons, it is necessary to restrict consideration to the situation of diagonal Q, Σ, and A for δ^{MB}.) These formulas for the Bayes risks are of interest in themselves.

In Section 3, various flat tailed π are considered for the separation problem involving δ^{RB}. (The calculations for δ^{MB} quickly became too messy to be of much use.) First, a type of flat tailed prior (see (3.1.1)) is considered, which leads to an easy expression (see (3.1.10)) for the difference of Bayes risks between the combined and separate estimators, an expression involving constants (see Table 1) easily calculated numerically. A wide variety of cases is considered, and the combined estimator is always seen to be superior. Next, under some technical conditions on π, the case where $k\to\infty$ (and $k_l/k\to\tau_l$ for $l=1,\ldots,s$) is considered. This allows development of asymptotic expressions for the difference in Bayes risks of the combined and separate estimators, which should be useful indicators of behavior for large dimensions. The results are not quite as clearcut as the previous results, however, in that it is shown that separation can be better if the fourth moment of the prior is large enough (compared to the square of the second moment). Such priors are somewhat unusual, however, so the combined estimator is (asymptotically) superior for typical priors.

The above mentioned results came as something of a surprise to us, in that it seemed natural to imagine that when π was a product of the individual group priors and the groups were quite dissimilar then the separate estimators would perform better than the combined estimator. That the opposite was generally true seems to indicate a hitherto unsuspected power of the Stein effect (see also Berger (1982b)).

One limitation of this work should be mentioned. It is natural to also investigate the grouping question when worried about misspecification of μ and A. If, for example, one is quite certain about the specification of these quantities for one group, but not for a different group, then separation will probably be better. This is an interesting problem currently under investigation. The work in this paper should thus be considered a necessary first step along the path to a final solution.

In the remainder of the paper it will be assumed, without loss of generality, that $\mu=0$. This can be accomplished by a simple translation of the problem. (Only translation invariant π will be considered.)

2. Separation under normal priors

2.1. The robust generalized Bayes estimator

For normal prior distributions, linear transformations of the problem do not affect Bayes risk. It is easy to check that a linear transformation can be made which preserves the block diagonal structure of Σ and A and for which the transformed loss is sum of squares error loss (i.e. $Q=I_k$). It will, therefore, be assumed in this section that $Q=I_k$ and π is $\mathrm{N}(0, A)$.

The calculation of $r(\pi, \delta^{\mathrm{RB}})$ is made difficult by the presence of the function r in (1.1). We will, therefore, replace r by $(k-2)$, and consider the estimator

$$\delta^{\mathrm{RB}*}(X)=\left(I_k-\frac{(k-2)}{X^{\mathrm{t}}(\Sigma+A)^{-1}X}\,\Sigma(\Sigma+A)^{-1}\right)X. \tag{2.1.1}$$

(Recall we set $\mu=0$, without loss of generality.) δ^{RB} and $\delta^{\mathrm{RB}*}$ are very similar. (Indeed they have virtually identical risks for moderately large k.) Numerical studies in Dey (1980) (which for the sake of brevity will not be reported) indicate that the separation results for the two estimators are identical.

As in Berger (1980) a calculation using integration by parts gives that

$$\begin{aligned} R(\theta,\delta^{\mathrm{RB}*})=\operatorname{tr}\Sigma+E_\theta\Bigg[&-\frac{2(k-2)}{\|X\|^2}\Bigg\{\operatorname{tr}\Sigma^2(\Sigma+A)^{-1} \\ &-\frac{2X^{\mathrm{t}}(\Sigma+A)^{-1}\Sigma^2(\Sigma+A)^{-1}X}{\|X\|^2}\Bigg\} \\ &+\frac{(k-2)^2X^{\mathrm{t}}(\Sigma+A)^{-1}\Sigma^2(\Sigma+A)^{-1}X}{\|X\|^4}\Bigg] \end{aligned} \tag{2.1.2}$$

where $\|X\|^2=X^{\mathrm{t}}(\Sigma+A)^{-1}X$ and tr stands for the trace of a matrix.

The following lemma then gives the desired Bayes risk.

Lemma 2.1.1. *If $Q=I_k$ and π is* N(0, A), *then*

$$r(\pi,\delta^{\mathrm{RB}*})=\operatorname{tr}\Sigma-\frac{(k-2)}{k}\operatorname{tr}\Sigma^2(\Sigma+A)^{-1}. \tag{2.1.3}$$

Proof. It is clear that marginally X is N(0, $\Sigma+A$), and hence $X^t(\Sigma+A)^{-1}X$ has a chi-square distribution with k-degrees of freedom. Thus it can be easily shown that

$$E^X\left[\frac{1}{\|X\|^2}\right]=\frac{1}{k-2}. \tag{2.1.4}$$

Now let O be a $k\times k$ orthogonal matrix chosen so that

$$\Lambda=O^t(\Sigma+A)^{-1/2}\Sigma^2(\Sigma+A)^{-1/2}O$$

is diagonal, with diagonal elements $\lambda_1,\lambda_2,\dots,\lambda_k$. Let a random variable Y be defined as $Y=O(\Sigma+A)^{-1/2}X$. Then we have

$$\|X\|^2=X^t(\Sigma+A)^{-1}X=Y^tY \tag{2.1.5}$$

and

$$X^t(\Sigma+A)^{-1}\Sigma^2(\Sigma+A)^{-1}X=Y^t\Lambda Y=\sum_{i=1}^{n}\lambda_iY_i^2. \tag{2.1.6}$$

Therefore,

$$E^X\left[\frac{X^t(\Sigma+A)^{-1}\Sigma^2(\Sigma+A)^{-1}X}{\|X\|^4}\right]=\frac{1}{k(k-2)}\operatorname{tr}\Sigma^2(\Sigma+A)^{-1}, \tag{2.1.7}$$

using Lemma 1 of the appendix, since Y_i^2, $i=1,\dots,k$, are independent chi-square with 1 degree of freedom. Now (2.1.2) follows immediately using (2.1.4) and (2.1.7), which completes the proof. □

For completeness, we note that the separate estimator $\delta^{\mathrm{RB}*\mathrm{S}}=(\delta_{(1)}^{\mathrm{RB}*},\dots,\delta_{(s)}^{\mathrm{RB}*})^t$ is defined by

$$\delta_{(l)}^{\mathrm{RB}*}=\left(I_{k_l}-\frac{(k_l-2)}{X_{(l)}^t(\Sigma_{(l)}+A_{(l)})^{-1}X_{(l)}}\Sigma_{(l)}(\Sigma_{(l)}+A_{(l)})^{-1}\right)X_{(l)}. \tag{2.1.8}$$

Applying Lemma 2.1.1 to each group separately and summing shows that $\delta^{\mathrm{RB}*\mathrm{S}}$ has Bayes risk

$$\begin{aligned}r(\pi,\delta^{\mathrm{RB}*\mathrm{S}})&=\sum_{l=1}^{s}\left\{\operatorname{tr}\Sigma_{(l)}-\frac{(k_l-2)}{k_l}\operatorname{tr}\Sigma_{(l)}^2(\Sigma_{(l)}+A_{(l)})^{-1}\right\}\\&=\operatorname{tr}\Sigma-\sum_{l=1}^{s}\left(1-\frac{2}{k_l}\right)\operatorname{tr}\Sigma_{(l)}^2(\Sigma_{(l)}+A_{(l)})^{-1}.\end{aligned} \tag{2.1.9}$$

The major result is given in the following theorem, namely that the combined estimator is always better than the separate estimator in this situation.

Theorem 2.1.1. *Suppose $Q=I_k$ and π is* N(0, A). *Then $\delta^{\mathrm{RB}*}$ is better than $\delta^{\mathrm{RB}*\mathrm{S}}$ in terms of Bayes risk.*

Proof. Comparing (2.1.3) and (2.1.9), noting that $k_l \le k$, the conclusion follows. □

2.2. The minimax Bayes estimator

In this section, it will be assumed that π is N(0, A); that Q, Σ, and A are diagonal with diagonal elements q_i, d_i and a_i, respectively, and that the coordinates are indexed so that $q_1^* \ge q_2^* \ge \cdots \ge q_k^*$, where $q_i^* = q_i d_i^2/(d_i + a_i)$. In Berger (1982a) the following estimator was shown to be minimax (i.e. uniformly better than δ^0), and yet allowed incorporation of μ and A: $\delta^{\mathrm{MB}*} = (\delta_1^{\mathrm{MB}*}, \ldots, \delta_k^{\mathrm{MB}*})$, where

$$\delta_i^{\mathrm{MB}*}(X) = q_i^{*-1} \sum_{j=i}^{k} (q_j^* - q_{j+1}^*) \left[\left(1 - \frac{(j-2)^+}{\|X^j - \mu^j\|^2} \frac{d_i}{(d_i + a_i)} \right)(X_i - \mu_i) + \mu_i \right], \tag{2.2.1}$$

where

$$\|X^j - \mu^j\|^2 = \sum_{l=1}^{j} (X_l - \mu_l)^2/(d_l + a_l),$$

$q_{k+1}^* \equiv 0$, and $(j-2)^+$ denotes the positive part of $(j-2)$.

A complication arises in this situation when attempting to divide θ into groups for separate estimation. The complication is that it is important to retain the ordering of the q_i^* in each group. (The q_i^* reflect the 'importance' of the coordinates to improvement in simultaneous estimation. See Berger (1982a) for further discussion.) We will therefore only consider groups formed from the given ordering; i.e. the first group will be $\theta_{(1)} = (\theta_1, \ldots, \theta_{k_1})^{\mathrm{t}}$, the second group will be $\theta_{(2)} = (\theta_{k_1+1}, \ldots, \theta_{k_1+k_2})^{\mathrm{t}}$, etc., where the θ_i are as above (with corresponding q_i^* that are decreasing). Such grouping in terms of decreasing q_i^* is natural, in any case, since 'similar' coordinates should have similar q_i^*. The following theorem shows that the combined estimator $\delta^{\mathrm{MB}*}$ has smaller Bayes risk than the separate estimator $\delta^{\mathrm{MB}*\mathrm{S}} = (\delta_{(1)}^{\mathrm{MB}*}, \ldots, \delta_{(s)}^{\mathrm{MB}*})^{\mathrm{t}}$, where $\delta_{(l)}^{\mathrm{MB}*}$ is given componentwise as

$$\delta^{\mathrm{MB}*\mathrm{S}} = q_i^{*-1} \sum_{j=T_{l-1}+1}^{T_l} (q_j^* - q_{j+1}^*) \times \left[\left(1 - \frac{(j-2-T_{l-1})^+}{\|X^j - \mu^j\|_l^2} \frac{d_i}{d_i + a_i} \right)(X_i - \mu_i) + \mu_i \right],$$

where $T_l = k_1 + \cdots + k_l (l = 1, \ldots, s)$, $T_0 \equiv 0$,

$$\|X^j - \mu^j\|_l^2 = \sum_{i=T_{l-1}+1}^{j} (X_i - \mu_i)^2/(d_i + a_i),$$

and (by an abuse of notation) $q^*_{T_l+1}$ in the above expression is understood to be zero for each l.

Theorem 2.2.1. *In the above situation, where π is* N(0, A), $\delta^{\mathrm{MB}*}$ *has smaller Bayes risk than* $\delta^{\mathrm{MB}*\mathrm{S}}$.

Proof. For simplicity, we will drop the * from q_i^*. Equation (A1) of Berger (1982a) shows that

$$\begin{aligned}
r(\pi,\delta^{\mathrm{MB}*}) &= \operatorname{tr} Q\Sigma - \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{(j-2)^+}{j}\,\frac{(q_j-q_{j+1})}{q_i}\,[2q_i-(q_j+q_{j+1})] \\
&\geq \operatorname{tr} Q\Sigma - \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{(q_j-q_{j+1})}{q_i}\,[2q_i-(q_j+q_{j+1})] \\
&= \operatorname{tr} Q\Sigma - \sum_{i=1}^{k}\sum_{j=1}^{k} 2(q_j-q_{j+1}) + \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{(q_j^2-q_{j+1}^2)}{q_i} \\
&= \operatorname{tr} Q\Sigma - \sum_{i=1}^{k} 2q_i + \sum_{i=1}^{k} \frac{q_i^2}{q_i} \\
&= \operatorname{tr} Q\Sigma - \sum_{i=1}^{k} q_i.
\end{aligned} \tag{2.2.2}$$

This inequality clearly also holds for the $\delta_{(l)}^{\mathrm{MB}*\mathrm{S}}$, implying that

$$\begin{aligned}
r(\pi,\delta^{\mathrm{MB}*\mathrm{S}}) &= \sum_{l=1}^{s} r(\pi,\delta_{(l)}^{\mathrm{MB}*\mathrm{S}}) \\
&\geq r(\pi,\delta_{(1)}^{\mathrm{MB}*\mathrm{S}}) + \sum_{l=2}^{s} \left\{ \operatorname{tr} Q_{(l)}\Sigma_{(l)} - \sum_{i=T_{l-1}+1}^{T_l} q_i \right\}.
\end{aligned} \tag{2.2.3}$$

Lemma 2 of Berger (1982a) states that

$$r(\pi,\delta_{(1)}^{\mathrm{MB}*}) = \operatorname{tr} Q_{(1)}\Sigma_{(1)} - \sum_{i=3}^{T_1} q_i - 2\sum_{i=3}^{T_1} \frac{q_i}{i}\left[1-\frac{q_i}{(i-1)}\sum_{j=1}^{(i-1)}\frac{1}{q_j}\right].$$

Using this in (2.2.3) together with the fact that

$$1-\frac{q_i}{(i-1)}\sum_{j=1}^{(i-1)}\frac{1}{q_j} \geq 0, \quad i=1,\ldots,k$$

(since the q_i^* are nonincreasing), we obtain

$$\begin{aligned}
r(\pi^{\mathrm{MB}*\mathrm{S}}) &\geq \operatorname{tr} Q\Sigma - \sum_{i=3}^{k} q_i - 2\sum_{i=3}^{T_1}\frac{q_i}{i}\left[1-\frac{q_i}{(i-1)}\sum_{j=1}^{(i-1)}\frac{1}{q_j}\right] \\
&\geq \operatorname{tr} Q\Sigma - \sum_{i=3}^{k} q_i - 2\sum_{i=3}^{k}\frac{q_i}{i}\left[1-\frac{q_i}{(i-1)}\sum_{j=1}^{(i-1)}\frac{1}{q_j}\right] \\
&= r(\pi,\delta^{\mathrm{MB}*}) \quad \text{(Lemma 2 of Berger (1982a)).}
\end{aligned}$$

This establishes the theorem. □

3. Separation under a flat prior

The results of the previous section are somewhat surprising. Even for a normal prior which reflects independence of the various groups of coordinates, the combined estimators seem better than the separate estimators. To alleviate concerns that this result may be due to the sharp tails of the normal prior, we consider in this section priors which incorporate μ and A but have flat tails. We restrict consideration to $\delta^{\mathrm{RB}*}$ and $\delta^{\mathrm{RB}*\mathrm{S}}$, and again assume (w.l.o.g.) that $\mu=0$.

3.1. Numerical results for a certain flat-tailed prior

We will assume that Σ and A are diagonal with diagonal elements d_i and a_i, respectively. Also assume that given λ_i, the θ_i's are independent $\mathrm{N}(0, b(\lambda_i))$, $i=1,\dots,k$, where the λ_i's are independently distributed with density $f(\lambda)=n\lambda^{n-1}$, for $n>0$ and $0<\lambda<1$, and $b(\lambda_i)=(c_i/\lambda_i)-d_i$, $i=1,\dots,k$, where c_i is defined here as $c_i=d_i+a_i$. Thus the generalized prior density for θ_i is

$$g_n(\theta_i)=\frac{n}{\sqrt{2\pi}}\int_0^1 (b(\lambda_i))^{-1/2}\exp\left\{-\frac{\theta_i^2}{2b(\lambda_i)}\right\}\lambda_i^{n-1}\,\mathrm{d}\lambda_i. \tag{3.1.1}$$

It can be shown asymptotically (for large θ_i) that $g_n(\theta_i)$ behaves like $C_1(\theta_i)^{-2n}$, for some constant C_1. Thus g_n is a prior density with a tail considerably flatter than that of a normal density. (This particular density is chosen for its comparative ease in calculation.) Clearly, given λ_i, the X_i's are independent $\mathrm{N}(0, c_i/\lambda_i)$ for all $i=1,\dots,k$. Thus from (2.1.2) the Bayes risk of $\delta^{\mathrm{RB}*}$ with respect to the above prior is given as

$$\begin{aligned} r(g_n,\delta^{\mathrm{RB}*}) &= \operatorname{tr}\Sigma-2(k-2)\sum_{i=1}^k\frac{d_i^2}{c_i}E^{\lambda}E^{X\mid\lambda}\left[\frac{1}{\sum_{i=1}^k X_i^2/c_i}\right] \\ &\quad +(k^2-4)E^{\lambda}E^{X\mid\lambda}\left[\frac{\sum_{i=1}^k d_i^2X_i^2/c_i^2}{(\sum_{i=1}^k X_i^2/c_i)^2}\right] \end{aligned} \tag{3.1.2}$$

where $\lambda=(\lambda_1,\dots,\lambda_k)$ and $E^{X\mid\lambda}(\cdot)$ stands for the expectation under the conditional distribution of X given λ.

By choosing different values of n, several flat priors can be generated. We take $n=2$ for simplicity. Note that

$$E^{\lambda}E^{X\mid\lambda}\left[\frac{1}{\sum_{i=1}^k X_i^2/c_i}\right]=E^X\left[\frac{1}{\sum_{i=1}^k X_i^2/c_i}\right] \tag{3.1.3}$$

and

$$E^{\lambda}E^{X\mid\lambda}\left[\frac{\sum_{i=1}^k d_i^2X_i^2/c_i^2}{(\sum_{i=1}^k X_i^2/c_i)^2}\right]=\sum_{i=1}^k\frac{d_i^2}{c_i}E^X\left[\frac{X_i^2/c_i}{(\sum_{i=1}^k X_i^2/c_i)^2}\right]. \tag{3.1.4}$$

Since the λ_i's are independent and identically distributed, the unconditional distributions of X_i^2/c_i are independent of i. Therefore

$$kE^X\left[\frac{X_i^2/c_i}{(\sum_{i=1}^k X_i^2/c_i)^2}\right]=E^X\left[\frac{1}{\sum_{i=1}^k X_i^2/c_i}\right]. \tag{3.1.5}$$

Using (3.1.3), (3.1.4) and (3.1.5) in (3.1.2) we have

$$r(g_1,\delta^{\mathrm{RB}*})=\operatorname{tr}\Sigma-\frac{(k-2)^2}{k}VS, \tag{3.1.6}$$

where $V=\sum_{i=1}^k d_i^2/c_i$ and $S=E^X[1/\sum_{i=1}^k X_i^2/c_i]$. S can be computed by a numerical integration technique using Lemma 2 in the appendix. Table 1 gives the values of S for various k. (Using the conditional representation in (3.1.3) it is easy to see that S does not depend on the c_i.)

Table 1
Values of S for different values of k

k	S	k	S
3	0.38339	13	0.05117
4	0.23961	14	0.04670
5	0.17750	15	0.04295
6	0.13959	20	0.03055
7	0.11382	25	0.02366
8	0.09540	30	0.01928
9	0.08173	35	0.01646
10	0.07129	40	0.01406
11	0.06310	45	0.01236
12	0.05654	50	0.01020

In calculating the Bayes risk of the separate estimator, define S_l and V_l for the l-th group as

$$S_l=E^{X_{(l)}}\left[\frac{1}{\sum_{i=T_{l-1}+1}^{T_l} X_i^2/c_i}\right], \tag{3.1.7}$$

and

$$V_l=\sum_{i=T_{l-1}+1}^{T_l} d_i^2/c_i,\quad l=1,2,\ldots,s, \tag{3.1.8}$$

where T_l is defined as in Section 2. Note that S_l can be found from Table 1. Now using (3.1.7), (3.1.8) and (3.1.2), the Bayes risk of $\delta_{(l)}^{\mathrm{RB}*}$ is given by

$$r(g_1,\delta_{(l)}^{\mathrm{RB}*})=\operatorname{tr}\Sigma_{(l)}-\frac{(k_l-2)^2}{k_l}V_lS_l,\quad l=1,\ldots,s. \tag{3.1.9}$$

Then we can conclude that the difference of the Bayes risks of the combined and separate estimators is given by

$$\begin{aligned} \Delta &= r(g_1, \delta^{\mathrm{RB}*}) - r(g_1, \delta^{\mathrm{RB}*\mathrm{S}}) \\ &= r(g_1, \delta^{\mathrm{RB}*}) - \sum_{l=1}^{s} r(g_1, \delta^{\mathrm{RB}*}_{(l)}) \\ &= \sum_{l=1}^{s} \frac{(k_l-2)^2}{k_l} V_l S_l - \frac{(k-2)^2}{k} VS. \end{aligned} \tag{3.1.10}$$

Using Table 1, Δ can be computed for different c_i, k, and k_l. A representative sample of such calculations is given in Table 2 for $s=2$. All calculated values of Δ are negative, which indicates that the combined estimator is better than the separate estimator with respect to the Bayes risk, under the given flat prior.

Remark 3.1. If $s=2$ and $k_1=k_2$, we have $S_1=S_2$. Then (3.1.10) reduces to

$$\Delta = -V\left\{\frac{(k-2)^2}{k} S - \frac{(k_1-2)^2}{k_1} S_1\right\}.$$

Numerical calculations indicate that the expression within the brackets is always positive and it is clearly constant when k_1 and k are fixed. Thus for fixed k_1 and k, Δ is proportional to V.

Table 2
Difference of Bayes risks of combined and separate estimators; $k=10$

k_1	(d_i), $i=1,2,\ldots,10$	(c_i), $i=1,2,\ldots,10$	Δ
5	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	(10, 10, 10, 10, 10, 10, 10, 10, 10, 10)	−0.1256
5	(0.0001, 0.01, 0.1, 0.1, 1, 2, 2, 2, 2, 4)	(10, 10, 10, 10, 10, 12, 12, 12, 100, 100)	−0.1766
5	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)	(10, 10, 10, 10, 10, 10, 10, 10, 10, 10)	−0.3139
5	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)	(1, 1, 1, 1, 1, 1000, 1000, 1000, 1000, 1000)	−0.5558
5	(0.1, 0.1, 0.1, 0.1, 0.1, 1, 1, 1, 1, 1)	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	−0.5592
5	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)	(1, 1, 1, 1, 1, 200, 200, 200, 200, 200)	−0.5647
5	(1, 1, 1, 1, 1, 10, 10, 10, 10, 10)	(1, 1, 1, 1, 1, 1000, 1000, 1000, 1000, 1000)	−0.6090
5	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	(2, 2, 2, 2, 2, 2, 2, 2, 2, 2)	−0.6278
5	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)	(1, 1, 1, 1, 1, 10, 10, 10, 10, 10)	−0.7751
5	(1, 1, 1, 1, 1, 10, 10, 10, 10, 10)	(1, 1, 1, 1, 1, 200, 200, 200, 200, 200)	−0.8305
5	(1, 1, 1, 1, 1, 10, 10, 10, 10, 10)	(1, 1, 1, 1, 1, 100, 100, 100, 100, 100)	−1.1073
5	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	−1.2556
5	(1, 1, 1, 1, 1, 10, 10, 10, 10, 10)	(1, 1, 1, 1, 1, 10, 10, 10, 10, 10)	−6.0902
5	(1, 1, 1, 1, 1, 10, 10, 10, 10, 10)	(10, 10, 10, 10, 10, 10, 10, 10, 10, 10)	−6.3406
4	(0.0001, 0.0001, 0.0001, 0.01, 0.01, 10, 10, 10, 10, 10)	(1, 1, 1, 2, 2, 10, 10, 100, 100, 100)	−11.6132
5	(0.0001, 0.0001, 0.0001, 0.01, 0.01, 10, 10, 10, 10, 10)	(1, 1, 1, 2, 2, 10, 10, 100, 100, 100)	−25.4680
6	(0.0001, 0.0001, 0.0001, 0.01, 0.01, 10, 10, 10, 10, 10)	(1, 1, 1, 2, 2, 10, 10, 100, 100, 100)	−62.2560
7	(0.0001, 0.0001, 0.0001, 0.01, 0.01, 10, 10, 10, 10, 10)	(1, 1, 1, 2, 2, 10, 10, 100, 100, 100)	−158.4928

3.2. Asymptotic (as $k \to \infty$) results for separation

Under the assumption that Σ is a diagonal matrix and the prior has uniformly bounded tenth moments, we will approximate the Bayes risk of the combined estimator $\delta^{\mathrm{RB}*}$ and the separate estimator $\delta^{\mathrm{RB}*\mathrm{S}}$ for large k. Assume that

$$\Sigma = \operatorname{diag}(\underbrace{d_1, \ldots, d_1}_{k_1}, \ldots, \underbrace{d_s, \ldots, d_s}_{k_s})$$

and that the prior π on θ is such that $\theta_1, \ldots, \theta_k$ are independent with $E(\theta_i) = 0$, $E(\theta_i^2) = a_l$ and $E(\theta_i^2 - a_l)^2 = v_l$ when $T_{l-1} + 1 \le i \le T_l$; $l = 1, \ldots, s$ (the T_l being defined as before). Suppose also that there exists a $T < \infty$ such that, for all i, $E(\theta_i^{10}) < T$ and $a_i/d_i < T$. Finally, assume that $\tau_l = \lim_{k\to\infty}(k_l/k)$ exists for all l, and that $0 < \tau_l < 1$.

Theorem 3.2.1. *If Σ and π are as above, then Δ, defined as the difference of the Bayes risks of $\delta^{\mathrm{RB}*}$ and $\delta^{\mathrm{RB}*\mathrm{S}}$, is given by*

$$\Delta = \sum_{l=1}^{s} \frac{d_l^2}{d_l + a_l} \left\{ 2(\tau_l - 1) + \frac{v_l - 2a_l^2}{(d_l + a_l)^2} - \tau_l \sum_{j=1}^{s} \frac{(v_j - 2a_j^2)\tau_j}{(d_j + a_j)^2} \right\} + \mathrm{o}(1) \tag{3.2.1}$$

where $\mathrm{o}(1)$ *converges to zero as $k \to \infty$.*

Proof. Given in the appendix.

Remark 3.2. For normal priors, an easy calculation shows that $v_l = 2a_l^2$. Hence $\Delta < 0$ asymptotically, agreeing with the more explicit results of Section 2.

Remark 3.3. If $v_l = v$, $a_l = a$, and $d_l = d$ $(l = 1, \ldots, s)$, then

$$\Delta = \frac{d^2}{d+a}(s-1)\left\{\frac{v - 2a^2}{(d+a)^2} - 2\right\}.$$

Thus separation is asymptotically *better*, even in this symmetric situation, if

$$\frac{v - 2a^2}{(d+a)^2} > 2. \tag{3.2.2}$$

This inequality can be satisfied for very flat tailed priors. For example, it can be shown to hold for the truncated t priors

$$\pi(\theta_i) = c\left(1 + \frac{\theta_i^2}{\beta}\right)^{-(\alpha+1)/2} I_{\{|\theta_i| \le M\}},$$

providing β and M are large enough and $4 < \alpha < 7$. (The truncation at M is to ensure that the moment assumptions of Theorem 3.2.1 are satisfied.)

This is somewhat discouraging, in that it shows that separation can be better if the fourth moment of the prior is large enough. Trying a variety of possible forms for π, however, will convince the reader that it is quite rare for (3.2.2) to be satisfied (providing the appropriate numbers of moments exist). Since it is rare to have accurate enough prior knowledge to be able to specify a fourth moment, use of the combined estimator is again indicated therefore.

Remark 3.4. It is natural to question the assumption that the prior even has finite moments. Priors with very flat tails are not at all unreasonable. But if the prior does not have (say) finite variances, then it is clear that the estimators considered here are all inadequate, since $(k-2)/(X-\mu)^{t}(\Sigma+A)^{-1}(X-\mu)$ becomes infinitely small (with probability one) as $k\to\infty$. To deal with this problem, Stein (1981) proposed (for the symmetric case) truncating excessively large values of the X_i. Determination of the optimal truncation point is an interesting problem discussed in Dey (1980). Numerical results indicate, however, that for properly truncated versions, the combined estimator is still better than the separate estimator (assuming the prior tail for each coordinate is felt to be similar).

4. Appendix

Lemma 1. *If $\{X_i\}$, $i=1,2,\ldots,k$, is a sequence of independent and identically distributed chi-square random variables with one degree of freedom, then*

$$E\left[\frac{\sum_{i=1}^{k} l_i X_i}{(\sum_{i=1}^{k} X_i)^2}\right]=\frac{1}{k(k-2)}\sum_{i=1}^{k} l_i, \tag{A1}$$

where the l_i are scalars.

Proof. We have

$$E\left[\frac{l_i X_i}{(\sum_{i=1}^{k} X_i)^2}\right]=l_i E\left[\frac{(1/k)\sum_{i=1}^{k} X_i}{(\sum_{i=1}^{k} X_i)^2}\right]=\frac{l_i}{k}E\left[\frac{1}{\sum_{i=1}^{k} X_i}\right]=\frac{l_i}{k(k-2)}.$$

Now summing over i, the proof is complete. □

Lemma 2.

$$S=E^{X}\left[\frac{1}{\sum_{i=1}^{k} X_i^2/c_i}\right]=\int_0^\infty\left\{(1+2t)^{1/2}(1-3t)+6t^2\sinh^{-1}\frac{1}{(2t)^{-1/2}}\right\}^{k}\mathrm{d}t. \tag{A2}$$

Proof. Suppose $Z_i=X_i^2$, $i=1,\ldots,k$. Then $Z_i\mid\lambda_i$ is $(c_i/\lambda_i)\chi_1^2$ where χ_1^2 represents a chi-square random variable with one degree of freedom. Thus given λ_i, Z_i/c_i is $(1/\lambda_i)\chi_1^2$. Hence given λ_i, the Laplace transform of Z_i/c_i is

$$\phi(t\mid\lambda_i)=E^{Z_i\mid\lambda_i}[\mathrm{e}^{-tZ_i/c_i}]=\left(1+\frac{2t}{\lambda_i}\right)^{-1/2}.$$

It follows that the unconditional Laplace transform of Z_i/c_i is

$$\phi(t)=\int_0^1 2\left(1+\frac{2t}{\lambda_i}\right)^{-1/2}\lambda_i\,\mathrm{d}\lambda_i=4t\int_0^{\theta^*}\tanh\theta\sinh^2\theta\,4tc\sinh\theta\cosh\theta\,\mathrm{d}\theta,$$

the last step following by the change of variables $\lambda=2t\sinh^2\theta$ and defining $\theta^*=\sinh^{-1}(2t)^{-1/2}$. Integrating by parts gives

$$\begin{aligned}\phi(t)&=16t^2\int_0^{\theta^*}\sinh^4\theta\,\mathrm{d}\theta=16t^2\left\{\tfrac{1}{4}\sinh^3\theta^*\cosh\theta^*-\tfrac{3}{4}\int_0^{\theta^*}\sinh^2\theta\,\mathrm{d}\theta\right\}\\&=16t^2\{\tfrac{1}{4}(2t)^{-2}(1+2t)^{1/2}-\tfrac{3}{4}[\tfrac{1}{4}\sinh 2\theta-\tfrac{1}{2}\theta]_0^{\theta^*}\}\\&=(1+2t)^{1/2}-3t(1+2t)^{1/2}+6t^2\sinh^{-1}(2t)^{-1/2}.\end{aligned}$$

Now using the independence of the Z_i/c_i for $i=1,\dots,k$, the Laplace transform of $\sum_{i=1}^k Z_i/c_i$ (or $\sum_{i=1}^k X_i^2/c_i$) is given as

$$E\left[\exp\left(-t\sum_{i=1}^k X_i^2/c_i\right)\right]=(1+t)^{1/2}(1-3t)+6t^2\sinh^{-1}(2\mathrm{t})^{-1/2}.$$

Finally by Fubini's theorem,

$$\begin{aligned}E^X\left[\frac{1}{\sum_{i=1}^k X_i^2/c_i}\right]&=\int_0^\infty E\left[\exp\left(-t\sum_{i=1}^k X_i^2/c_i\right)\right]\mathrm{d}t\\&=\int_0^\infty\{(1+2t)^{1/2}(1-3t)+6t^2\sinh^{-1}(2t)^{-1/2}\}^k\,\mathrm{d}t,\end{aligned}$$

which completes the proof of the lemma. □

The following lemmas are neded in the proof of Theorem 3.2.1. All relevant notation and conditions are given in Subsection 3.2.

Lemma 3. *For* $T_{l-1}+1\le i\le T_l$,

$$E^X\left[\frac{X_i^2}{d_l+a_l}-1\right]=2+\frac{v_l-2a_l^2}{(d_l+a_l)^2}\,. \tag{A3}$$

where E^X *stands for expectation under the marginal distribution of* X.

Proof. We have $X_i\,|\,\theta_i$ is $\mathrm{N}(\theta_i,d_l)$, $T_{l-1}\le i\le T_l$. Therefore

$$E^X(X_i^2)=E^{\theta_i}E^{X_i|\theta_i}(X_i^2)=E^{\theta_i}(d+\theta_i^2)=d_l+a_l.$$

We know that $E^{X_i|\theta_i}(X_i-\theta_i)^4=3d_l^2$. Therefore,

$$E^{X_i|\theta_i}(X_i^4)=3d_l^2+6\theta_i^2d_l+\theta_i^4.$$

Thus,

$$E^X(X_i^4)=3d_l^2+6d_la_l+v_l+a_l^2$$

(since $E(\theta_i^4) = v_l + a_l^2$). Therefore

$$E^X\left[\frac{X_i^2}{d_l + a_l} - 1\right]^2 = E^X\left[\frac{X_i^4}{(d_l + a_l)^2} + 1 - \frac{2X_i^2}{d_l + a_l}\right] = 2 + \frac{v_l - 2a_l^2}{(d_l + a_l)^2}.$$

This completes the proof of the lemma. □

Lemma 4. *Under the assumptions given at the beginning of Section* 3.2,

$$E^X\left[\frac{k_l}{\|X_{(l)}\|^2}\right] = 1 + \frac{2}{k_l} + \frac{v_l - 2a_l^2}{k_l(d_l + a_l)^2} + o\left(\frac{1}{k_l}\right). \tag{A4}$$

Proof. Define $Y_i^2 = X_i^2/(d_l + a_l)$ and

$$\varepsilon = E^{Y_{(l)}}\left\{\left[\sum_{i=T_{l-1}+1}^{T_l}(Y_i^2 - 1)\right]^3 \Big/ \left[k_l \sum_{i=T_{l-1}+1}^{T_l} Y_i^2\right]\right\}. \tag{A5}$$

An easy calculation then gives that

$$E^X\left[\frac{k_l}{\|X_{(l)}\|^2}\right] = 1 - E^{Y_{(l)}}\left[\frac{1}{k_l}\sum_{i=T_{l-1}+1}^{T_l}(Y_i^2 - 1)\right] + E^{Y_{(l)}}\left[\frac{1}{k_l}\sum_{i=T_{l-1}+1}^{T_l}(Y_i^2 - 1)\right]^2 - \frac{\varepsilon}{k_l}.$$

Observing that the Y_i^2 are independent with $E^Y(Y_i^2) = 1$ and using (A3), it follows that

$$E^X\left[\frac{k_l}{\|X_{(l)}\|^2}\right] = 1 + \frac{1}{k_l}\left\{2 + \frac{v_l - 2a_l^2}{(d_l + a_l)^2}\right\} - \frac{\varepsilon}{k_l}.$$

Hence it is only necessary to prove that $\varepsilon = o(1)$.

Expanding the numerator in (A5) gives

$$\begin{aligned}\varepsilon &= E^{Y_{(l)}}\left\{\frac{\sum_i (Y_i^2 - 1)^3}{k_l \sum_i Y_i^2}\right\} + 3E^{Y_{(l)}}\left\{\frac{\sum\sum_{i\neq j}(Y_i^2 - 1)^2(Y_j^2 - 1)}{k_l \sum_i Y_i^2}\right\} \\ &\quad + 6E^{Y_{(l)}}\left\{\frac{\sum\sum\sum_{i\neq j\neq m}(Y_i^2 - 1)(Y_j^2 - 1)(Y_m^2 - 1)}{k_l \sum_i Y_i^2}\right\} \\ &= I_1 + I_2 + I_3 \quad \text{(say).}\end{aligned} \tag{A6}$$

(In the above expression and in the following, it is to be understood that summations or products are only over the integers $T_{l-1}+1, \ldots, T_l$.)

To deal with I_1, define for some some $a > 0$

$$\Omega = \left\{(Y_{T_{l-1}+1}, \ldots, Y_{T_l})^t : \sum_{i\neq j} Y_i^2 \geq a^2 \text{ for } j = T_{l-1}+1, \ldots, T_l\right\},$$

and write (letting $f(Y_i)$ denote the marginal density of Y_i)

$$I_1 = \int_{\Omega} \frac{\sum_i (Y_i^2-1)^3}{k_l \sum_i Y_i^2} \prod_i [f(Y_i)\,\mathrm{d}Y_i] + \int_{\Omega^c} \frac{\sum_i (Y_i^2-1)^3}{k_l \sum_i Y_i^2} \prod_i [f(Y_i)\,\mathrm{d}Y_i]$$
$$= I_1^1 + I_1^2 \quad \text{(say)}. \tag{A7}$$

Now on Ω,

$$\frac{|\sum_i (Y_i^2-1)^3|}{k_l \sum_i Y_i^2} \le \sum_i \frac{(Y_i^2-1)^2 |Y_i^2-1|}{k_l(a^2+Y_i^2)}$$
$$\le \sum_i \frac{(Y_i^2-1)^2}{k_l}\left(1+\frac{1}{a^2}\right) \equiv g_{k_l}(Y_{(l)}) \quad \text{(say)}.$$

From the assumption of uniformly bounded 10th moments, it can be assumed without loss of generality that $E(Y_i^2-1)^2 \le T$ for all i. An easy argument, using the strong law of large numbers and the extended Lebesgue dominated convergence theorem (see Rao (1973) for a statement of this theorem due to V. Johns and J. Pratt), then shows that $|I_1^1| = \mathrm{o}(1)$.

To deal with I_1^2 from (A7), define

$$\Omega_j = \left\{ Y_{(l)} : \sum_{i \ne j} Y_i^2 < a^2 \right\},$$

observe that $\Omega^c = \bigcup_{j=T_{l-1}+1}^{T_l} \Omega_j$, and note from this and the independence of the Y_i that

$$|I_1^2| \le \sum_i \int_{\Omega^c} \frac{|Y_i^2-1|^3}{k_l \sum_{j\ne i} Y_j^2} \prod_j [f(Y_j)\,\mathrm{d}Y_j]$$
$$\le \sum_i \left\{ (E|Y_i^2-1|^3) \sum_m \int_{\Omega_m} \frac{1}{k_l \sum_{j \ne i,m} Y_j^2} \prod_{j\ne i} [f(Y_j)\,\mathrm{d}Y_j] \right\}$$
$$\le \sum_i \sum_m [E|Y_i^2-1|^3] \int_{\{\sum_{j\ne i,m} Y_j^2 < a^2\}} \frac{1}{k_l \sum_{j\ne i,m} Y_j^2} \prod_{j \ne i,m} [f(Y_j)\,\mathrm{d}Y_j].$$

Using the fact that $a_l/d_l \le T$, it is clear that

$$f(Y_i) = E^{\pi(\theta_i)}\left[\left\{ \frac{(d_l+a_l)}{2\pi d_l} \right\}^{1/2} \exp\left\{ -\frac{(d_l+a_l)}{2d_l} \left(Y_i - \frac{\theta_i}{(d_l+a_l)^{1/2}} \right)^2 \right\} \right]$$
$$\le \left\{ \frac{(d_l+a_l)}{2\pi d_l} \right\}^{1/2} \le \left(\frac{1+T}{2\pi} \right)^{1/2}.$$

Together with the uniform bound T on the moments $E|Y_i^2-1|^3$, it follows that

$$|I_1^2| \le \sum_i \sum_m T\left(\frac{1+T}{2\pi}\right)^{(k_l-2)/2} \int_{\{\sum_{j\ne i,m} Y_j^2 < a^2\}} \frac{1}{k_l \sum_{j\ne i,m} Y_j^2} \prod_{j\ne i,m} \mathrm{d}Y_j$$
$$= k_l^2 T \left(\frac{1+T}{2\pi}\right)^{(k_l-2)/2} \frac{a^{k_l-4}\pi^{(k_l-2)/2}}{k_l \Gamma(k_l/2)}$$
$$= \mathrm{o}(1).$$

To deal with I_2 from (A6), observe first that

$$\frac{1}{\sum_i Y_i^2} = \frac{1}{\sum_{i\neq j} Y_i^2} - \frac{Y_j^2}{(\sum_{i\neq j} Y_i^2)(\sum_i Y_i^2)}.$$

It follows that

$$I_2 = 3 \sum_{i\neq j}\sum \left[E^{Y_{(l)}}\left\{\frac{(Y_i^2-1)^2(Y_j^2-1)}{k_l \sum_{i\neq j} Y_i^2}\right\} - E^{Y_{(l)}}\left\{\frac{(Y_i^2-1)^2(Y_j^2-1)Y_j^2}{k_l(\sum_{i\neq j} Y_i^2)(\sum_i Y_i^2)}\right\}\right]$$

$$= -3 \sum_{i\neq j}\sum E^{Y_{(l)}}\left\{\frac{(Y_i^2-1)^2(Y_j^2-1)Y_j^2}{k_l(\sum_{i\neq j} Y_i^2)(\sum_i Y_i^2)}\right\} \quad (\text{since } E(Y_j^2)=1).$$

The argument that $|I_2| = \mathrm{o}(1)$ proceeds as did that for I_1.

The argument for I_3 follows similarly after applying the identity

$$\frac{1}{\sum_i Y_i^2} = \frac{1}{\sum_{i\neq j} Y_i^2} - \frac{Y_j^2}{(\sum_{i\neq j} Y_i^2)^2} + \frac{Y_j^4}{(\sum_{i\neq j} Y_i^2)^2(\sum_i Y_i^2)}.$$

It is for this term that uniformly bounded tenth moments are required. □

Lemma 5.

$$E^X\left[\frac{k}{\|X\|^2}\right] = 1 + \frac{2}{k} + \frac{1}{k^2}\sum_{l=1}^{s}\frac{(v_l - 2a_l^2)}{(d_l+a_l)^2} + \mathrm{o}\left(\frac{1}{k}\right). \tag{A8}$$

Proof. Similar to the proof of Lemma 4. □

Proof of Theorem 3.2.1. Using (2.1.2) and Lemma 5 we have that the Bayes risk of $\delta^{\mathrm{RB}*}$ is

$$r(\pi, \delta^{\mathrm{RB}*}) = \sum_{l=1}^{s} k_l d_l - \frac{(k-2)^2}{k^2}\left\{\sum_{l=1}^{s}\frac{d_l^2}{d_l+a_l}k_l\right\}$$
$$\times\left\{1 + \frac{2}{k} + \frac{1}{k^2}\sum_{l=1}^{s}\frac{k_l(v_l-2a_l^2)}{(d_l+a_l)^2} + \mathrm{o}\left(\frac{1}{k}\right)\right\}. \tag{A9}$$

Similarly by Lemma 4, we have

$$r(\pi, \delta_{(l)}^{\mathrm{RB}*}) = k_l d_l - \frac{(k_l-2)^2 d_l^2}{k_l(d_l+a_l)}\left\{1 + \frac{2}{k_l} + \frac{v_l - 2a_l^2}{k_l(d_l+a_l)^2} + \mathrm{o}\left(\frac{1}{k_l}\right)\right\}.$$

Thus the Bayes risk of the separate estimator is

$$r(\pi, \delta^{\mathrm{RB}*\mathrm{S}}) = \sum_{l=1}^{s} k_l d_l - \sum_{l=1}^{s}\frac{(k_l-2)^2}{k_l}\,\frac{d_l^2}{d_l+a_l}$$
$$\times\left\{1 + \frac{2}{k_l} + \frac{v_l - 2a_l^2}{k_l(d_l+a_l)^2} + \mathrm{o}\left(\frac{1}{k_l}\right)\right\}. \tag{A10}$$

From (A9) and (A10), it follows that the difference of Bayes risks is

$$\begin{aligned}\Delta = & -4\sum_{l=1}^{s}\frac{d_l}{d_l+a_l}+4\sum_{l=1}^{s}\frac{d_l^2\tau_l}{d_l+a_l}+\sum_{l=1}^{s}\frac{d_l^2}{d_l+a_l}\left\{2+\frac{\upsilon_l-2a_l^2}{(d_l+a_l)^2}\right\}\\ & -\sum_{l=1}^{s}\frac{2d_l^2\tau_l}{d_l+a_l}-\left\{\sum_{l=1}^{s}\frac{d_l^2\tau_l}{d_l+a_l}\sum_{l=1}^{s}\frac{(\upsilon_l-2a_l^2)\tau_l}{(d_l+a_l)^2}\right\}+\mathrm{o}(1)\\ = & \sum_{l=1}^{s}\frac{d_l^2}{d_l+a_l}\left\{2(\tau_l-1)+\frac{\upsilon_l-2a_l^2}{(d_l+a_l)^2}-\tau_l\sum_{j=1}^{s}\frac{(\upsilon_j-2a_j^2)\tau_j}{(d_j+a_j)^2}\right\}+\mathrm{o}(1),\end{aligned}$$

which completes the proof of the theorem. □

References

Berger, J. (1975). Minimax estimation of location vectors of a wide class of densities. *Ann. Statist.* 3, 1318–1328.

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

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

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

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

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

Berger, J. (1982b). Bayesian robustness and the Stein effect. *J. Amer. Statist. Assoc.* 77, 358–368.

Bhattacharya, P.K. (1966). Estimating the mean of a multivariate normal population with general loss function. *Ann. Math. Statist.* 37, 1819–1927.

Bock, M.E. (1975). Minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* 3, 209–218.

Chung, K.L. (1968). *A Course in Probability Theory*. Harcourt, Brace & World, New York.

Dey, D.K. (1980). Ph.D. Thesis, Purdue University.

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

Efron, B. and C. Morris (1973b). Combining possibly related estimation problems. *J. Roy. Statist. Soc. Ser. B* 35, 379–421.

Hudson, M. (1974). Empirical Bayes Estimation. Technical Report #58, Stanford University.

James, W. and C. Stein (1960). Estimation with quadratic loss. *Proc. Fourth Berkeley Symposium Math. Stat. Prob.* 1, 361–379. University of California Press.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. Second Edition. John Wiley and Sons, New York.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Stat. Prob.* 1, 197–206. University of California Press.

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

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