Bernoulli **21**(1), 2015, 574–603
DOI: 10.3150/13-BEJ580

Improved minimax estimation of a multivariate normal mean under heteroscedasticity

ZHIQIANG TAN

[1]*Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA. E-mail: ztan@stat.rutgers.edu*

Consider the problem of estimating a multivariate normal mean with a known variance matrix, which is not necessarily proportional to the identity matrix. The coordinates are shrunk directly in proportion to their variances in Efron and Morris' (*J. Amer. Statist. Assoc.* **68** (1973) 117–130) empirical Bayes approach, whereas inversely in proportion to their variances in Berger's (*Ann. Statist.* **4** (1976) 223–226) minimax estimators. We propose a new minimax estimator, by approximately minimizing the Bayes risk with a normal prior among a class of minimax estimators where the shrinkage direction is open to specification and the shrinkage magnitude is determined to achieve minimaxity. The proposed estimator has an interesting simple form such that one group of coordinates are shrunk in the direction of Berger's estimator and the remaining coordinates are shrunk in the direction of the Bayes rule. Moreover, the proposed estimator is scale adaptive: it can achieve close to the minimum Bayes risk simultaneously over a scale class of normal priors (including the specified prior) and achieve close to the minimax linear risk over a corresponding scale class of hyper-rectangles. For various scenarios in our numerical study, the proposed estimators with extreme priors yield more substantial risk reduction than existing minimax estimators.

Keywords: Bayes risk; empirical Bayes; minimax estimation; multivariate normal mean; shrinkage estimation; unequal variances

1. Introduction

A fundamental statistical problem is shrinkage estimation of a multivariate normal mean. See, for example, the February 2012 issue of *Statistical Science* for a broad range of theory, methods, and applications. Let $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ be multivariate normal with *unknown* mean vector $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and *known* variance matrix Σ. Consider the problem of estimating θ by an estimator $\delta = \delta(X)$ under the loss $L(\delta, \theta) = (\delta - \theta)^{\mathrm{T}} Q(\delta - \theta)$, where Q is a *known* positive definite, symmetric matrix. The risk of δ is $R(\delta, \theta) = E_\theta\{L(\delta, \theta)\}$. The general problem can be transformed into a canonical form such that Σ is diagonal and $Q = I$, the identity matrix (e.g., Lehmann and Casella [21], Problem 5.5.11). For simplicity, assume except in Section 3.2 that Σ is $D = \mathrm{diag}(d_1, \ldots, d_p)$ and $L(\delta, \theta) = \|\delta - \theta\|^2$, where $\|x\|^2 = x^{\mathrm{T}} x$ for a column vector x. The letter D is substituted for Σ to emphasize that it is diagonal.

For this problem, we aim to develop shrinkage estimators that are both minimax and capable of effective risk reduction over the usual estimator $\delta_0 = X$ even in the heteroscedastic case (i.e., $d_1, \ldots, d_p$ are not equal). An estimator of θ is minimax if and only if, *regardless of*

$\theta \in \mathbb{R}^p$, its risk is always no greater than $\sum_{j=1}^p d_j$, the risk of δ_0. For $p \geq 3$, minimax estimators different from and hence dominating δ_0 are first discovered in the homoscedastic case where $D = \sigma^2 I$ (i.e., $d_1 = \cdots = d_p = \sigma^2$). James and Stein [19] showed that $\delta_c^{\mathrm{JS}} = (1 - c\sigma^2/\|X\|^2)X$ is minimax provided $0 \leq c \leq 2(p-2)$. Stein [26] suggested the positive-part estimator $\delta_c^{\mathrm{JS}+} = (1 - c\sigma^2/\|X\|^2)_+ X$, which dominates δ_c^{JS}. Throughout, $a_+ = \max(0, a)$. Shrinkage estimation has since been developed into a general methodology with various approaches, including empirical Bayes (Efron and Morris [17]; Morris [22]) and hierarchical Bayes (Strawderman [28]; Berger and Robert [7]). While these approaches are prescriptive for constructing shrinkage estimators, minimaxity is not automatically achieved but needs to be checked separately.

For the heteroscedastic case, there remain challenging issues on how much observations with different variances should be shrunk relatively to each other (e.g., Casella [15], Morris [22]). For the empirical Bayes approach (Efron and Morris [17]), the coordinates of X are shrunk directly in proportion to their variances. But the existing estimators are, in general, non-minimax (i.e., may have a greater risk than the usual estimator δ_0). On the other hand, Berger [3] proposed minimax estimators, including admissible minimax estimators, such that the coordinates of X are shrunk inversely in proportion to their variances. But the risk reduction achieved over δ_0 is insubstantial unless all the observations have similar variances.

To address the foregoing issues, we develop novel minimax estimators for multivariate normal means under heteroscedasticity. There are two central ideas in our approach. The first is to develop a class of minimax estimators by generalizing a geometric argument essentially in Stein [25] (see also Brandwein and Strawderman [11]). For the homoscedastic case, the argument shows that δ_c^{JS} can be derived as an approximation to the best linear estimator of the form $(1 - \lambda)X$, where λ is a scalar. In fact, the optimal choice of λ in minimizing the risk is $p\sigma^2/E_\theta(\|X\|^2)$. Replacing $E_\theta(\|X\|^2)$ by $\|X\|^2$ leads to δ_c^{JS} with $c = p$. This derivation is highly informative, even though it does not yield the optimal value $c = p - 2$.

Our class of minimax estimators are of the linear form $(I - \lambda A)X$, where A is a nonnegative definite, diagonal matrix indicating the direction of shrinkage and λ is a scalar indicating the magnitude of shrinkage. The matrix A is open to specification, depending on the variance matrix D but *not* on the data X. For a fixed A, the scalar λ is determined to achieve minimaxity, depending on both D and X. Berger's [3] minimax estimator corresponds to the special choice $A = D^{-1}$, thereby leading to the unusual pattern of shrinkage discussed above.

The second idea of our approach is to choose A by approximately minimizing the Bayes risk with a normal prior in our class of minimax estimators. The Bayes risk is used to measure average risk reduction for θ in an elliptical region as in Berger [4,5]. It turns out that the solution of A obtained by our approximation strategy has an interesting simple form. In fact, the coordinates of X are automatically segmented into two groups, based on their Bayes "importance" (Berger [5]), which is of the same order as the coordinate variances when the specified prior is homoscedastic. The coordinates of high Bayes "importance" are shrunk inversely in proportion to their variances, whereas the remaining coordinates are shrunk in the direction of the Bayes rule. This shrinkage pattern may appear paradoxical: it may be expected that the coordinates of high Bayes "importance" are to be shrunk in the direction of the Bayes rule. But that scheme is inherently aimed at reducing the Bayes risk under the specified prior and, in general, fails to achieve minimaxity (i.e., it may lead to even a greater risk than the usual estimator δ_0).

In addition to simplicity and minimaxity, we further show that the proposed estimator is scale adaptive in reducing the Bayes risk: it achieves close to the minimum Bayes risk, with the difference no greater than the sum of the 4 highest Bayes "importance" of the coordinates of X, simultaneously over a scale class of normal priors (including the specified prior). To our knowledge, the proposed estimator seems to be the first one with such a property in the general heteroscedastic case. Previously, in the homoscedastic case, $\delta^{\mathrm{JS}}_{p-2}$ is known to achieve the minimum Bayes risk up to the sum of 2 (equal-valued) Bayes "importance" of the coordinates over the scale class of homoscedastic normal priors (Efron and Morris [17]).

The rest of this article is organized as follows. Section 2 gives a review of existing estimators. Section 3 develops the new approach and studies risk properties of the proposed estimator. Section 4 presents a simulation study. Section 5 provides concluding remarks. All proofs are collected in the Appendix.

2. Existing estimators

We describe a number of existing shrinkage estimators. See Lehmann and Casella [21] for a textbook account and Strawderman [29] and Morris and Lysy [23] for recent reviews. Throughout, $\mathrm{tr}(\cdot)$ denotes the trace and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. Then $\mathrm{tr}(D) = \sum_{j=1}^{p} d_j$ and $\lambda_{\max}(D) = \max(d_1, \ldots, d_p)$.

For a Bayes approach, assume the prior distribution: $\theta \sim \mathrm{N}(0, \gamma I)$, where γ is the prior variance. The Bayes rule is given componentwise by $\delta^{\mathrm{Bayes}}_j = \{1 - d_j/(d_j + \gamma)\}X_j$. Then the greater d_j is, the more X_j is shrunk whether γ is fixed or estimated from the data. For the empirical Bayes approach of Efron and Morris [17], γ is estimated by the maximum likelihood estimator $\hat{\gamma}$ such that

$$\hat{\gamma} = \sum_{j=1}^{p} \frac{X_j^2 - d_j}{(d_j + \hat{\gamma})^2} \Big/ \sum_{j=1}^{p} \frac{1}{(d_j + \hat{\gamma})^2}. \tag{1}$$

Morris [22] suggested the modified estimator

$$\delta^{\mathrm{EB}}_j = \left(1 - \frac{p-2}{p} \frac{d_j}{d_j + \hat{\gamma}_+}\right) X_j. \tag{2}$$

In our implementation, the right-hand side of (1) is computed to update $\hat{\gamma}$ from the initial guess, $p^{-1}\{\sum_{j=1}^{p}(X_j^2 - d_j)\}_+$, for up to 100 iterations until the successive absolute difference in $\hat{\gamma}$ is $\leq 10^{-4}$, or $\hat{\gamma}$ is set to ∞ so that $\delta^{\mathrm{EB}} = X$ otherwise.

Alternatively, Xie *et al.* [31] proposed empirical Bayes-type estimators based on minimizing Stein's [27] unbiased risk estimate (SURE) under heteroscedasticity. Their basic estimator is defined componentwise by

$$\delta^{\mathrm{XKB}}_j = \left(1 - \frac{d_j}{d_j + \tilde{\gamma}}\right) X_j, \tag{3}$$

where $\tilde{\gamma}$ is obtained by minimizing the SURE of δ^{Bayes}, that is, $\text{SURE}(\gamma) = X^{\text{T}}D\{D + \gamma I\}^{-1}X + 2\gamma\,\text{tr}\{D(D+\gamma I)^{-1}\} - \text{tr}(D)$. In general, the two types of empirical Bayes estimators, δ^{EB} and δ^{XKB}, are non-minimax, as shown in Section 4.

For a direct extension of δ_c^{JS}, consider the estimator $\delta_c^{\text{S}} = (1 - c/\|X\|^2)X$ and, more generally, $\delta_r^{\text{S}} = \{1 - r(\|X\|^2)/\|X\|^2\}X$, where c is a scalar constant and $r(\cdot)$ a scalar function. See Lehmann and Casella [21], Theorem 5.7, although there are some typos. Both δ_c^{S} and δ_r^{S} are spherically symmetric. The estimator δ_c^{S} is minimax provided

$$0 \le c \le 2\big\{\text{tr}(D) - 2\lambda_{\max}(D)\big\}, \tag{4}$$

and δ_r^{S} is minimax provided $0 \le r(\cdot) \le 2\{\text{tr}(D) - 2\lambda_{\max}(D)\}$ and $r(\cdot)$ is nondecreasing. No such $c \ne 0$ exists unless $\text{tr}(D) > 2\lambda_{\max}(D)$, which restricts how much $(d_1, \ldots, d_p)$ can differ from each other. For example, condition (4) fails when $p = 10$ and

$$d_1 = 40, \qquad d_2 = 20, \qquad d_3 = 10, \qquad d_4 = \cdots = d_{10} = 1, \tag{5}$$

because $\text{tr}(D) = 77$ and $\lambda_{\max}(D) = 40$.

Berger [3] proposed estimators of the form $\delta_c^{\text{B}} = \{I - cD^{-1}/(X^{\text{T}}D^{-2}X)\}X$ and $\delta_r^{\text{B}} = \{I - r(X^{\text{T}}D^{-2}X)/(X^{\text{T}}D^{-2}X)D^{-1}\}X$, where c is a scalar constant and $r(\cdot)$ a scalar function. Then δ_c^{B} is minimax provided $0 \le c \le 2(p-2)$, and δ_r^{B} is minimax provided $0 \le r(\cdot) \le 2(p-2)$ and $r(\cdot)$ is nondecreasing, regardless of differences between $(d_1, \ldots, d_p)$. However, a striking feature of δ_c^{B} and δ_r^{B}, compared with δ^{EB} and δ^{XKB}, is that the smaller d_j is, the more X_j is shrunk. For example (5), under δ_c^{B}, the coordinates (X_1, X_2, X_3) are shrunk only slightly, whereas $(X_4, \ldots, X_{10})$ are shrunk as if they were shrunk as a 7-dimensional vector under δ_c^{JS}. The associated risk reduction is insubstantial, because the risk of estimating $(\theta_4, \ldots, \theta_{10})$ is a small fraction of the overall risk of estimating θ.

Define the positive-part version of δ_c^{B} componentwise as

$$\big(\delta_c^{\text{B}+}\big)_j = \left(1 - \frac{cd_j^{-1}}{X^{\text{T}}D^{-2}X}\right)_+ X_j. \tag{6}$$

The estimator $\delta_c^{\text{B}+}$ dominates δ_c^{B} by Baranchik [1], Section 2.5. Berger [6], Equation (5.32), stated a different positive-part estimator, δ_r^{B} with $r(t) = \min(p-2, t)$, but the jth component may not be of the same sign as X_j.

Given a prior $\theta \sim \text{N}(0, \Gamma)$, Berger [5] suggested an approximation of Berger's [4] robust generalized Bayes estimator as

$$\delta^{\text{RB}} = \left[I - \min\left\{1, \frac{p-2}{X^{\text{T}}(D+\Gamma)^{-1}X}\right\}D(D+\Gamma)^{-1}\right]X. \tag{7}$$

The estimator is expected to provide significant risk reduction over $\delta_0 = X$ if the prior is correct and be robust to misspecification of the prior, but it is, in general, non-minimax. In the case of $\Gamma = 0$, δ^{RB} becomes $\{1 - (p-2)/(X^{\text{T}}D^{-1}X)\}_+X$, in the form of spherically symmetric estimators $\delta_r^{\text{SS}} = \{1 - r(X^{\text{T}}D^{-1}X)/(X^{T}D^{-1}X)\}X$, where $r(\cdot)$ is a scalar function (Bock [10],

Brown [12]). The estimator δ_r^{SS} is minimax provided $0 \le r(\cdot) \le 2\{\mathrm{tr}(D)/\lambda_{\max}(D) - 2\}$ and $r(\cdot)$ is nondecreasing. Moreover, if $\mathrm{tr}(D) \le 2\lambda_{\max}(D)$, then δ_r^{SS} is non-minimax unless $r(\cdot) = 0$.

To overcome the non-minimaxity of δ^{RB}, Berger [5] developed a minimax estimator δ^{MB} by combining δ_r^{B}, δ^{RB}, and a minimax estimator of Bhattacharya [9]. Suppose that $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$ and the indices are sorted such that $d_1^* \ge \cdots \ge d_p^*$, where $d_j^* = d_j^2/(d_j + \gamma_j)$. Define δ^{MB} componentwise as

$$\delta_j^{\mathrm{MB}} = X_j - \left[\frac{1}{d_j^*}\sum_{k=j}^{p}\left(d_k^* - d_{k+1}^*\right)\min\left\{1, \frac{(k-2)_+}{\sum_{\ell=1}^{k} X_\ell^2/(d_\ell + \gamma_\ell)}\right\}\right]\frac{d_j}{d_j + \gamma_j}X_j, \tag{8}$$

where $d_{p+1}^* = 0$. In the case of $\Gamma = 0$, δ^{MB} reduces to the original estimator of Bhattacharya [9]. The factor $(k-2)_+$ is replaced by $2(k-2)_+$ in Berger's [5] original definition of δ^{MB}, corresponding to replacing $p-2$ by $2(p-2)$ in δ^{RB}. In our simulations, the two versions of δ^{MB} somehow yield rather different risk curves, and so do the corresponding versions of other estimators. But there has been limited theory supporting one version over the other. Therefore, we focus on comparisons of only the corresponding versions of δ^{MB} and other estimators.

3. Proposed approach

We develop a useful approach for shrinkage estimation under heteroscedasticity, by making explicit how different coordinates are shrunk differently. The approach not only sheds new light on existing results, but also lead to new minimax estimators.

3.1. A sketch

Assume that $\Sigma = D$ (diagonal) and $Q = I$. Consider estimators of the linear form

$$\delta = (I - \lambda A)X = X - \lambda AX, \tag{9}$$

where A is a nonnegative definite, diagonal matrix indicating the *direction* of shrinkage and λ is a scalar indicating the *magnitude* of shrinkage. Both A and λ are to be determined. A sketch of our approach is as follows.

(i) For a fixed A, the optimal choice of λ in minimizing the risk is

$$\lambda_{\mathrm{opt}} = \frac{\mathrm{tr}(DA)}{E_\theta(X^{\mathrm{T}}A^{\mathrm{T}}AX)}.$$

(ii) For a fixed A and a scalar constant $c \ge 0$, consider the estimator

$$\delta_{A,c} = X - \frac{c}{X^{\mathrm{T}}A^{\mathrm{T}}AX}AX.$$

By Theorem 1, an upper bound on the risk function of $\delta_{A,c}$ is

$$R(\delta_{A,c},\theta) \le \operatorname{tr}(D) + E_\theta\left[\frac{c\{c-2c^*(D,A)\}}{X^{\mathrm{T}}A^{\mathrm{T}}AX}\right], \tag{10}$$

where $c^*(D,A) = \operatorname{tr}(DA) - 2\lambda_{\max}(DA)$. Requiring the second term to be no greater than 0 shows that if $c^*(D,A) \ge 0$, then $\delta_{A,c}$ is minimax provided

$$0 \le c \le 2c^*(D,A). \tag{11}$$

If $c^*(D,A) \ge 0$, then the upper bound (10) has a minimum at $c = c^*(D,A)$.

(iii) By taking $c = c^*(D,A)$ in $\delta_{A,c}$, consider the estimator

$$\delta_A = X - \frac{c^*(D,A)}{X^{\mathrm{T}}A^{\mathrm{T}}AX}AX$$

subject to $c^*(D,A) \ge 0$, so that δ_A is minimax by step (ii). A positive-part estimator dominating δ_A is defined componentwise by

$$\left(\delta_A^+\right)_j = \left\{1 - \frac{c^*(D,A)a_j}{X^{\mathrm{T}}A^{\mathrm{T}}AX}\right\}_+ X_j, \tag{12}$$

where $(a_1,\ldots,a_p)$ are the diagonal elements of A. The upper bound (10) on the risk functions of δ_A and δ_A^+, subject to $c^*(D,A) \ge 0$, gives

$$R(\delta_A,\theta) \le \operatorname{tr}(D) - E_\theta\left\{\frac{c^{*2}(D,A)}{X^{\mathrm{T}}A^{\mathrm{T}}AX}\right\}. \tag{13}$$

We propose to choose A based on some optimality criterion, such as minimizing the Bayes risk with a normal prior centered at 0 (Berger [5]).

Further discussions of steps (i)–(iii) are provided in Sections 3.2–3.3.

3.2. Constructing estimators: Steps (i)–(ii)

We first develop steps (i)–(ii) for the general problem where neither Σ nor Q may be diagonal. The results can be as concisely stated as those just presented for the canonical problem where Σ is diagonal and $Q = I$. Such a unification adds to the attractiveness of the proposed approach.

Consider estimators of the form (9), where A is not necessarily diagonal, but

$$A\Sigma \text{ is nonnegative definite.} \tag{14}$$

Condition (14) is invariant under a linear transformation. To see this, let B be a nonsingular matrix and $\Sigma^* = B\Sigma B^{\mathrm{T}}$ and $A^* = BAB^{-1}$. For the transformed problem of estimating $\theta^* = B\theta$ based on $X^* = BX$ with variance matrix Σ^*, the transformed estimator from (9) is

$\delta^* = X^* - \lambda A^* X^*$. The application of condition (14) to δ^* says that $A^*\Sigma^* = BA\Sigma B^{\mathrm{T}}$ is nonnegative definite and therefore is equivalent to (14) itself. For the canonical problem where $\Sigma = D$ (diagonal), condition (14) only requires that AD is nonnegative definite, allowing A to be nondiagonal. On the other hand, it seems intuitively appropriate to restrict A to be diagonal. Then condition (14) is equivalent to saying that A is nonnegative definite (and diagonal), which is the condition introduced on A in the sketch in Section 3.1.

The risk of an estimator of the form (9) is

$$\begin{aligned} &E_\theta\big\{(X-\theta-\lambda AX)^{\mathrm{T}} Q(X-\theta-\lambda AX)\big\} \\ &\quad = E_\theta\big\{(X-\theta)^{\mathrm{T}} Q(X-\theta)\big\} + \lambda^2 E_\theta\big(X^{\mathrm{T}} A^{\mathrm{T}} QAX\big) - 2\lambda E_\theta\big\{(X-\theta)^{\mathrm{T}} QAX\big\}. \end{aligned}$$

For a fixed A, the optimal λ in minimizing the risk is

$$\lambda_{\mathrm{opt}} = \frac{E_\theta\{(X-\theta)^{\mathrm{T}} QAX\}}{E_\theta(X^{\mathrm{T}} A^{\mathrm{T}} QAX)} = \frac{\mathrm{tr}(\Sigma QA)}{E_\theta(X^{\mathrm{T}} A^{\mathrm{T}} QAX)}.$$

Replacing $E_\theta(X^{\mathrm{T}} A^{\mathrm{T}} QAX)$ by $X^{\mathrm{T}} A^{\mathrm{T}} QAX$ and $\mathrm{tr}(\Sigma QA)$ by a scalar constant $c \geq 0$ leads to the estimator

$$\delta_{A,c} = X - \frac{c}{X^{\mathrm{T}} A^{\mathrm{T}} QAX} AX.$$

For a generalization, replacing c by $r(X^{\mathrm{T}} A^{\mathrm{T}} QAX)$ with a scalar function $r(\cdot) \geq 0$ leads to the estimator

$$\delta_{A,r} = X - \frac{r(X^{\mathrm{T}} A^{\mathrm{T}} QAX)}{X^{\mathrm{T}} A^{\mathrm{T}} QAX} AX.$$

We provide in Theorem 1 an upper bound on the risk function of $\delta_{A,r}$.

Theorem 1. *Assume that $r(\cdot)$ almost differentiable (Stein [27]). If (14) holds and $r(\cdot) \geq 0$ is nondecreasing, then for each θ,*

$$R(\delta_{A,r}, \theta) \leq \mathrm{tr}(\Sigma Q) + E_\theta\left[\frac{r\{r - 2c^*(\Sigma, Q, A)\}}{X^{\mathrm{T}} A^{\mathrm{T}} QAX}\right], \tag{15}$$

where $r = r(X^{\mathrm{T}} A^{\mathrm{T}} QAX)$ and $c^*(\Sigma, Q, A) = \mathrm{tr}(A\Sigma Q) - \lambda_{\max}(A\Sigma Q + \Sigma A^{\mathrm{T}} Q)$. Taking $r(\cdot) \equiv c \geq 0$ in (15) gives an upper bound on $R(\delta_{A,c}, \theta)$.

Requiring the second term in the risk upper bound (15) to be no greater than 0 leads to a sufficient condition for $\delta_{A,r}$ to be minimax.

Corollary 1. *If (14) holds and $c^*(\Sigma, Q, A) \geq 0$, then $\delta_{A,r}$ is minimax provided*

$$0 \leq r(\cdot) \leq 2c^*(\Sigma, Q, A) \quad \text{and} \quad r(\cdot) \text{ is nondecreasing.} \tag{16}$$

Particularly, $\delta_{A,c}$ is minimax provided $0 \leq c \leq 2c^*(\Sigma, Q, A)$.

For the canonical problem, inequality (15) and condition (16) for $\delta_{A,c}$ give respectively (10) and (11). These results generalize the corresponding ones for δ_c^{S} and δ_c^{B} in Section 2, by the specific choices $A = I$ or D^{-1}. The generalization also holds if c is replaced by a scalar function $r(\cdot) > 0$. In fact, condition (16) reduces to Baranchik's [2] condition in the homoscedastic case.

If $c^*(\Sigma, Q, A) \geq 0$, then the risk upper bound (15) has a minimum at $r(\cdot) \equiv c = c^*(\Sigma, Q, A)$. As a result, consider the estimator

$$\delta_A = X - \frac{c^*(\Sigma, Q, A)}{X^{\mathrm{T}} A^{\mathrm{T}} Q A X} AX,$$

which is minimax provided $c^*(\Sigma, Q, A) \geq 0$. If $A = Q^{-1}\Sigma^{-1}$ (Berger [3]), then $c^*(\Sigma, Q, A) = p - 2$ and, by the proof of Theorem 1 in the Appendix, the risk upper bound (15) becomes exact for $\delta_{A,c}$. Therefore, for $A = Q^{-1}\Sigma^{-1}$, the estimator $\delta_A = \delta_{A,p-2}$ is uniformly best in the class $\delta_{A,c}$, in agreement with the result that $\delta_{p-2}^{\mathrm{JS}}$ is uniformly best among δ_c^{JS} in the homoscedastic case.

The estimator δ_A has desirable properties of invariance. First, δ_A is easily shown to be invariant under a multiplicative transformation $A \mapsto aA$ for a scalar $a > 0$. Second, δ_A is invariant under a linear transformation of the inference problem. Similarly as discussed below (14), let B be a nonsingular matrix and $\Sigma^* = B\Sigma B^{\mathrm{T}}$, $Q^* = B^{\mathrm{T}^{-1}} Q B^{-1}$, and $A^* = BAB^{-1}$. For the transformed problem of estimating $\theta^* = B\theta$ based on $X^* = BX$, the transformed estimator from δ_A is $X^* - \{c^*(\Sigma, Q, A)/(X^{*\mathrm{T}} A^{*\mathrm{T}} Q^* A^* X^*)\} A^* X^*$, whereas the application of δ_A is $X^* - \{c^*(\Sigma^*, Q^*, A^*)/(X^{*\mathrm{T}} A^{*\mathrm{T}} Q^* A^* X^*)\} A^* X^*$. The two estimators are identical because $A^*\Sigma^* Q^* = BA\Sigma Q B^{-1}$, $\Sigma^* A^{*\mathrm{T}} Q^* = B\Sigma A^{\mathrm{T}} Q B^{-1}$, and hence $c^*(\Sigma^*, Q^*, A^*) = c^*(\Sigma, Q, A)$.

Finally, we present a positive-part estimator dominating δ_A in the case where both $A\Sigma$ and QA are symmetric, that is,

$$A\Sigma = \Sigma A^{\mathrm{T}} \quad \text{and} \quad QA = A^{\mathrm{T}} Q. \tag{17}$$

Similarly to (14), it is easy to see that this condition is invariant under a linear transformation. Condition (17) is trivially true if Σ, Q, and A are diagonal. In the Appendix, we show that (17) holds if and only if there exists a nonsingular matrix B such that $Q = B^{\mathrm{T}} B$, $\Sigma = B^{-1} D B^{\mathrm{T}^{-1}}$, and $A = B^{-1} A^* B$, where D and A^* are diagonal and the diagonal elements of D or A^* are, respectively, the eigenvalues of ΣQ or A. In the foregoing notation, $\Sigma^* = D$ and $Q^* = I$. For the problem of estimating $\theta^* = B\theta$ based on $X^* = BX$, consider the estimator $\eta = X - \{c^*(D, A^*)/(X^{*\mathrm{T}} A^{*\mathrm{T}} A^* X^*)\} A^* X$ and the positive-part estimator η^+ with the jth component,

$$\left\{1 - \frac{c^*(D, A^*)}{X^{*\mathrm{T}} A^{*\mathrm{T}} A^* X^*} a_j^*\right\}_+ X_j^*,$$

where $(a_1^*, \ldots, a_p^*)$ are the diagonal elements of A^*. The estimator η^+ dominates η by a simple extension of Baranchik [1], Section 2.5. By a transformation back to the original problem, η

yields δ_A, whereas η^+ yields

$$\delta_A^+ = B^{-1}\operatorname{diag}\left[\left\{1 - \frac{c^*(\Sigma, Q, A)}{X^{\mathrm{T}}A^{\mathrm{T}}QAX}a_1^*\right\}_+, \ldots, \left\{1 - \frac{c^*(\Sigma, Q, A)}{X^{\mathrm{T}}A^{\mathrm{T}}QAX}a_p^*\right\}_+\right]BX.$$

Then δ_A^+ dominates δ_A. Therefore, (15) also gives an upper bound on the risk of δ_A^+, with $r(\cdot) \equiv c^*(\Sigma, Q, A)$, even though δ_A^+ is not of the form $\delta_{A,r}$.

In practice, a matrix A satisfying (17) can be specified in two steps. First, find a nonsingular matrix B such that $Q = B^{\mathrm{T}}B$ and $\Sigma = B^{-1}DB^{\mathrm{T}-1}$, where D is diagonal. Second, pick a diagonal matrix A^* and define $A = B^{-1}A^*B$. The first step is always feasible by taking $B = OC$, where C is a nonsingular matrix such that $Q = C^{\mathrm{T}}C$ and O is an orthogonal matrix O such that $O(C\Sigma C^{\mathrm{T}})O^{\mathrm{T}}$ is diagonal. Given (Σ, Q) and D, it can be shown that A and δ_A^+ depend on the choice of A^*, but not on that of B, provided that $a_j^* = a_k^*$ if $d_j = d_k$ for any $j, k = 1, \ldots, p$. In the canonical case where $\Sigma = D$ and $Q = I$, this condition amounts to saying that any coordinates of X with the same variances should be shrunk in the same way.

3.3. Constructing estimators: Step (iii)

Different choices of A lead to different estimators δ_A and δ_A^+. We study how to choose A, depending on (Σ, Q) but *not* on X, to approximately optimize risk reduction while preserving minimaxity for δ_A. The estimator δ_A^+ provides even greater risk reduction than δ_A. We focus on the canonical problem where $\Sigma = D$ (diagonal) and $Q = I$. Further, we restrict A to be diagonal and nonnegative definite.

As discussed in Berger [4], any estimator can have significantly smaller risk than $\delta_0 = X$ only for θ in a specific region. Berger [4,5] considered the situation where significant risk reduction is desired for an elliptical region

$$\left\{\theta\colon\ (\theta - \mu)^{\mathrm{T}}\Gamma^{-1}(\theta - \mu) \le p\right\}, \tag{18}$$

with μ and Γ the prior mean and prior variance matrix. See δ^{RB} and δ^{MB} reviewed in Section 2. To measure average risk reduction for θ in region (18), Berger [5] used the Bayes risk with the normal prior $\theta \sim \mathrm{N}(\mu, \Gamma)$. For simplicity, assume throughout that $\mu = 0$ and $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_p)$ is diagonal.

We adopt Berger's [5] ideas of specifying an elliptical region and using the Bayes risk to quantify average risk reduction in this region. We aim to find A, subject to $c^*(D, A) \ge 0$, minimizing the Bayes risk of δ_A with the prior π_Γ, $\theta \sim \mathrm{N}(0, \Gamma)$,

$$R(\delta_A, \pi_\Gamma) = E^{\pi_\Gamma}E_\theta\left(\|\delta_A - \theta\|^2\right),$$

where E^{π_Γ} denotes the expectation with respect to the prior π_Γ. Given A, the risk $R(\delta_A, \pi_\Gamma)$ can be numerically evaluated. A simple Monte Carlo method is to repeatedly draw $\theta \sim \mathrm{N}(0, \Gamma)$ and $X|\theta \sim \mathrm{N}(\theta, D)$ and then take the average of $\|\delta_A(X) - \theta\|^2$. But it seems difficult to literally implement the foregoing optimization. Alternatively, we develop a simple method for choosing A by two approximations.

First, if $c^*(D, A) \geq 0$, then taking the expectation of both sides of (13) with respect to the prior π_Γ gives an upper bound on the Bayes risk of δ_A:

$$R(\delta_A, \pi_\Gamma) \leq \operatorname{tr}(D) - E^m \left\{ \frac{c^{*2}(D, A)}{X^{\mathrm{T}} A^{\mathrm{T}} A X} \right\}, \tag{19}$$

where E^m denotes the expectation with respect to the marginal distribution of X in the Bayes model, that is, $X \sim \mathrm{N}(0, D + \Gamma)$. An approximation strategy for choosing A is to minimize the upper bound (19) on the Bayes risk or to maximize the second term. The expectation $E^m\{(X^{\mathrm{T}} A^{\mathrm{T}} A X)^{-1}\}$ can be evaluated as a 1-dimensional integral by results on inverse moments of quadratic forms in normal variables (e.g., Jones [20]). But the required optimization problem remains difficult.

Second, approximations can be made to the distribution of the quadratic form $X^{\mathrm{T}} A^{\mathrm{T}} A X$. Suppose that $X^{\mathrm{T}} A^{\mathrm{T}} A X$ is approximated with the same mean by $\{\sum_{j=1}^p (d_j + \gamma_j) a_j^2\} \chi_p^2 / p$, where χ_p^2 is a chi-squared variable with p degrees of freedom. Then $E^m\{(X^{\mathrm{T}} A^{\mathrm{T}} A X)^{-1}\}$ is approximated by $\{p/(p-2)\}\{\sum_{j=1}^p (d_j + \gamma_j) a_j^2\}^{-1}$. We show in the Appendix that this approximation gives a valid lower bound:

$$E^m \left(\frac{1}{X^{\mathrm{T}} A^{\mathrm{T}} A X} \right) \geq \frac{p}{p-2} \cdot \frac{1}{\sum_{j=1}^p (d_j + \gamma_j) a_j^2}. \tag{20}$$

A direct application of Jensen's inequality shows that $E^m\{(X^{\mathrm{T}} A^{\mathrm{T}} A X)^{-1}\} \geq \{\sum_{j=1}^p (d_j + \gamma_j) a_j^2\}^{-1}$. But the lower bound (20) is strictly tighter and becomes exact when $(d_1 + \gamma_1) a_1^2 = \cdots = (d_p + \gamma_p) a_p^2$. No simple bounds such as (20) seem to hold if more complicated approximations (e.g., Satterthwaite [24]) are used.

Combining (19) and (20) shows that if $c^*(D, A) \geq 0$, then

$$R(\delta_A, \pi_\Gamma) \leq \operatorname{tr}(D) - \frac{p}{p-2} \cdot \frac{c^{*2}(D, A)}{\sum_{j=1}^p (d_j + \gamma_j) a_j^2}. \tag{21}$$

Notice that δ_A is invariant under a multiplicative transformation $A \mapsto aA$ for a scalar $a > 0$, and so is the upper bound (21). Our strategy for choosing A is to minimize the upper bound (21) subject to $c^*(D, A) \geq 0$ or, equivalently, to solve the constrained optimization problem:

$$\begin{aligned} \max_A \quad & c^*(D, A) = \sum_{j=1}^p d_j a_j - 2 \max_{j=1,\ldots,p} d_j a_j \\ \text{subject to} \quad & \sum_{j=1}^p (d_j + \gamma_j) a_j^2 = \text{fixed}. \end{aligned} \tag{22}$$

The condition $c^*(D, A) \geq 0$ is dropped, because for $p \geq 3$, the achieved maximum is at least $c^*(D, aD^{-1}) = a(p-2) > 0$ for some scalar $a > 0$. In spite of the approximations used in our approach, Theorem 2 shows that not only the problem (22) admits a non-iterative solution, but

also the solution has a very interesting interpretation. For convenience, assume thereafter that the indices are sorted such that $d_1^2/(d_1+\gamma_1) \geq d_2^2/(d_2+\gamma_2) \geq \cdots \geq d_p^2/(d_p+\gamma_p)$.

Theorem 2. *Assume that* $p \geq 3$, $D = \mathrm{diag}(d_1, \ldots, d_p)$ *with* $d_j > 0$ *and* $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$ *with* $\gamma_j \geq 0$ $(j = 1, \ldots, p)$. *For problem* (22), *assume that* $A = \mathrm{diag}(a_1, \ldots, a_p)$ *with* $a_j \geq 0$ $(j = 1, \ldots, p)$ *and* $\sum_{j=1}^p (d_j + \gamma_j) a_j^2 = \sum_{j=1}^p d_j^2/(d_j + \gamma_j)$, *satisfied by* $a_j = d_j/(d_j + \gamma_j)$. *Then the following results hold.*

(i) *There exists a unique solution,* $A^\dagger = \mathrm{diag}(a_1^\dagger, \ldots, a_p^\dagger)$, *to problem* (22).

(ii) *Let* ν *be the largest index such that* $d_\nu a_\nu^\dagger = \max(d_1 a_1^\dagger, \ldots, d_p a_p^\dagger)$. *Then* $\nu \geq 3$, $d_1 a_1^\dagger = \cdots = d_\nu a_\nu^\dagger > d_j a_j^\dagger$ *for* $j \geq \nu + 1$, *and*

$$a_j^\dagger = K_\nu \left(\sum_{k=1}^{\nu} \frac{d_k + \gamma_k}{d_k^2} \right)^{-1} \frac{\nu - 2}{d_j} \qquad (j = 1, \ldots, \nu),$$

$$a_j^\dagger = K_\nu \frac{d_j}{d_j + \gamma_j} \qquad (j = \nu + 1, \ldots, p),$$

where $K_\nu = \{\sum_{j=1}^p d_j^2/(d_j + \gamma_j)\}^{1/2} M_\nu^{-1/2}$ *and*

$$M_\nu = \frac{(\nu - 2)^2}{\sum_{j=1}^{\nu} (d_j + \gamma_j)/d_j^2} + \sum_{j=\nu+1}^{p} \frac{d_j^2}{d_j + \gamma_j}.$$

The achieved maximum value, $c^*(D, A^\dagger)$, *is* $K_\nu M_\nu$ (> 0).

(iii) *The resulting estimator* $\delta_{A^\dagger}$ *is minimax.*

We emphasize that, although A can be considered a tuning parameter, the solution $A^\dagger$ is *data independent*, so that $\delta_{A^\dagger}$ is automatically minimax. If a data-dependent choice of A were used, minimaxity would not necessarily hold. This result is achieved both because each estimator δ_A with $c^*(D, A) \geq 0$ is minimax and because a global criterion (such as the Bayes risk) is used, instead of a pointwise criterion (such as the frequentist risk at the unknown θ), to select A. By these considerations, our approach differs from the usual exercise of selecting a tuning parameter in a data-dependent manner for a class of candidate estimators.

There is a remarkable property of monotonicity for the sequence $(M_3, M_4, \ldots, M_p)$, which underlies the uniqueness of ν and $A^\dagger$.

Corollary 2. *The sequence* $(M_3, M_4, \ldots, M_p)$ *is nonincreasing: for* $3 \leq k \leq p-1$, $M_k \geq M_{k+1}$, *where the equality holds if and only if*

$$\frac{k-2}{\sum_{j=1}^{k} (d_j + \gamma_j)/d_j^2} = \frac{d_{k+1}^2}{d_{k+1} + \gamma_{k+1}}.$$

The condition $d_\nu a_\nu^\dagger > d_{\nu+1} a_{\nu+1}^\dagger$ is equivalent to saying that the left side is greater than the right-hand side in the above expression for $k = \nu$. Therefore, ν is the smallest index $3 \le k \le p-1$ with this property, and $M_\nu > M_{\nu+1}$.

The estimator $\delta_{A^\dagger}$ is invariant under scale transformations of $A^\dagger$. Therefore, the constant K_ν can be dropped from the expression of $A^\dagger$ in Theorem 1.

Corollary 3. *The solution $A^\dagger = \operatorname{diag}(a_1^\dagger, \ldots, a_p^\dagger)$ can be rescaled such that*

$$a_j^\dagger = \left(\sum_{k=1}^{\nu} \frac{d_k + \gamma_k}{d_k^2} \right)^{-1} \frac{\nu - 2}{d_j} \qquad (j = 1, \ldots, \nu), \tag{23}$$

$$a_j^\dagger = \frac{d_j}{d_j + \gamma_j} \qquad (j = \nu + 1, \ldots, p). \tag{24}$$

Then $c^*(D, A^\dagger) = \sum_{j=1}^p a_j^{\dagger 2}(d_j + \gamma_j) = M_\nu$. Moreover, it holds that

$$a_j^\dagger \le \frac{d_j}{d_j + \gamma_j} \qquad (j = 1, \ldots, \nu). \tag{25}$$

The estimator $\delta_{A^\dagger}$ can be expressed as

$$\delta_{A^\dagger} = X - \frac{\sum_{j=1}^p a_j^{\dagger 2}(d_j + \gamma_j)}{\sum_{j=1}^p a_j^{\dagger 2} X_j^2} A^\dagger X. \tag{26}$$

The foregoing results lead to a simple algorithm for solving problem (22):

(i) Sort the indices such that $d_1^2/(d_1 + \gamma_1) \ge \cdots \ge d_p^2/(d_p + \gamma_p)$.

(ii) Take ν to be the smallest index k (corresponding to the largest M_k) such that $3 \le k \le p-1$ and

$$\frac{k-2}{\sum_{j=1}^k (d_j + \gamma_j)/d_j^2} > \frac{d_{k+1}^2}{d_{k+1} + \gamma_{k+1}},$$

or take $\nu = p$ if there exists no such k.

(iii) Compute $(a_1^\dagger, \ldots, a_p^\dagger)$ by (23)–(24).

This algorithm is guaranteed to find the (unique) solution to problem (22) by a fixed number of numerical operations. No iteration or convergence diagnosis is required. Therefore, the algorithm is exact and non-iterative, in contrast with usual iterative algorithms for nonlinear, constrained optimization.

The estimator $\delta_{A^\dagger}$ has an interesting interpretation. By (23)–(24), there is a dichotomous segmentation in the shrinkage direction of the coordinates of X based on $d_j^* = d_j^2/(d_j + \gamma_j)$. This quantity d_j^* is said to reflect the Bayes "importance" of θ_j, that is, the amount of reduction in

Bayes risk obtainable in estimating θ_j in Berger [5]. The coordinates with high d_j^* are shrunk inversely in proportion to their variances d_j as in Berger's [3] estimator δ_c^{B}, whereas the coordinates with low d_j^* are shrunk in the direction of the Bayes rule. Therefore, $\delta_{A^\dagger}$ mimics the Bayes rule to reduce the Bayes risk, except that $\delta_{A^\dagger}$ mimics δ_c^{B} for some coordinates of highest Bayes "importance" in order to achieve minimaxity. In fact, by inequality (25), the relative shrinkage, $a_j^\dagger/\{d_j/(d_j+\gamma_j)\}$, of each X_j $(j=1,\ldots,\nu)$ in $\delta_{A^\dagger}$ versus the Bayes rule is always no greater than that of X_k $(k=\nu+1,\ldots,p)$.

The expression (26) suggests that there is a close relationship in beyond the shrinkage direction between $\delta_{A^\dagger}$ and the Bayes rule under the Bayes model, $X \sim \mathrm{N}(0, D+\Gamma)$. In this case, $E^m(\sum_{j=1}^p a_j^{\dagger 2} X_j^2) = \sum_{j=1}^p a_j^{\dagger 2}(d_j+\gamma_j)$, and hence $\delta_{A^\dagger}$ behaves similarly to $X - A^\dagger X$. Therefore, *on average* under the Bayes model, the coordinates of X are shrunk in $\delta_{A^\dagger}$ the same as in the Bayes rule, except that some coordinates of highest Bayes "importance" are shrunk no greater than in the Bayes rule. While this discussion seems heuristic, we provide in Section 3.4 a rigorous analysis of the Bayes risk of $\delta_{A^\dagger}$, compared with that of the Bayes rule.

We now examine $\delta_{A^\dagger}$ for two types of priors: $\gamma_1 = \cdots = \gamma_p = \gamma$ and $\gamma_j = \gamma d_j$ $(j=1,\ldots,p)$, referred to as the homoscedastic and heteroscedastic priors. For both types, $(d_1^*,\ldots,d_p^*)$ are of the same order as the variances $(d_1,\ldots,d_p)$. Recall that δ_A is invariant under a multiplicative transformation of A. For both the homoscedastic prior with $\gamma=0$ and the heteroscedastic prior *regardless* of $\gamma \geq 0$, the solution $A^\dagger = \mathrm{diag}(a_1^\dagger,\ldots,a_p^\dagger)$ can be rescaled such that

$$a_j^\dagger = \left(\sum_{k=1}^{\nu} d_k^{-1}\right)^{-1} \frac{\nu-2}{d_j} \qquad (j=1,\ldots,\nu),$$

$$a_j^\dagger = 1 \qquad (j=\nu+1,\ldots,p).$$

Denote by $A_0^\dagger$ this rescaled matrix $A^\dagger$, corresponding to $\Gamma=0$. Then coordinates with high variances are shrunk inversely in proportion to their variances, whereas coordinates with low variances are shrunk symmetrically. For $\Gamma=0$, the proposed method has a purely frequentist interpretation: it seeks to minimize the upper bound (21) on the pointwise risk of δ_A at $\theta=0$.

For the homoscedastic prior with $\gamma\to\infty$, the proposed method is then to minimize the upper bound (21) on the Bayes risk of δ_A with an extremely flat, homoscedastic prior. As $\gamma\to\infty$, the solution $A^\dagger$ can be rescaled such that

$$a_j^\dagger = \left(\sum_{k=1}^{\nu} d_k^{-2}\right)^{-1} \frac{\nu-2}{d_j} \qquad (j=1,\ldots,\nu),$$

$$a_j^\dagger = d_j \qquad (j=\nu+1,\ldots,p).$$

Denote by $A_\infty^\dagger$ this rescaled matrix $A^\dagger$. Then coordinates with low (or high) variances are shrunk directly (or inversely) in proportion to their variances. The direction $A_\infty^\dagger$ can also be obtained

by using a fixed prior in the form $\gamma_j = \gamma d_1 - d_j$ $(j = 1, \ldots, p)$ for arbitrary $\gamma \geq 1$, where $d_1 = \max_{j=1,\ldots,p} d_j$.

Finally, in the homoscedastic case ($d_1 = \cdots = d_p = \sigma^2$), if the prior is also homoscedastic ($\gamma_1 = \cdots = \gamma_p = \gamma$), then $\nu = p$, $a_1^\dagger = \cdots = a_p^\dagger$, and $\delta_{A^\dagger}$ reduces to the James–Stein estimator $\delta_{p-2}^{\mathrm{JS}}$, *regardless* of σ^2 and γ.

3.4. Evaluating estimators

The estimator $\delta_{A^\dagger}$ is constructed by minimizing the upper bound (21) on the Bayes risk subject to minimaxity. In addition to simplicity, interpretability, and minimaxity demonstrated for $\delta_{A^\dagger}$, it remains important to further study risk properties of $\delta_{A^\dagger}$ and show that $\delta_{A^\dagger}$ can provide effective risk reduction over $\delta_0 = X$. Write $\delta_{A^\dagger} = \delta_{A^\dagger(\Gamma)}$ whenever needed to make explicit the dependency of $A^\dagger$ on Γ.

First, we study how close the Bayes risk of $\delta_{A^\dagger(\Gamma)}$ can be to that of the Bayes rule, which is the smallest possible among *all* estimators including non-minimax ones, under the prior π_Γ, $\theta \sim \mathrm{N}(0, \Gamma)$. The Bayes rule $\delta_\Gamma^{\mathrm{Bayes}}$ is given componentwise by $(\delta_\Gamma^{\mathrm{Bayes}})_j = \{1 - d_j/(d_j + \gamma_j)\}X_j$, with the Bayes risk

$$R\big(\delta_\Gamma^{\mathrm{Bayes}}, \pi_\Gamma\big) = \mathrm{tr}(D) - \sum_{j=1}^{p} d_j^*,$$

where $d_j^* = d_j^2/(d_j + \gamma_j)$, indicating the Bayes "importance" of θ_j (Berger [5]). The upper bound (21) on the Bayes risk of $\delta_{A^\dagger(\Gamma)}$ gives

$$R\{\delta_{A^\dagger(\Gamma)}, \pi_\Gamma\} \leq \mathrm{tr}(D) - \frac{p}{p-2} M_\nu = \mathrm{tr}(D) - \frac{p}{p-2}\left\{ \frac{(\nu-2)^2}{\sum_{j=1}^{\nu} d_j^{*-1}} + \sum_{j=\nu+1}^{p} d_j^* \right\}, \tag{27}$$

because $c^*(D, A^\dagger) = \sum_{j=1}^{p} (d_j + \gamma_j) a_j^{\dagger 2} = M_\nu$ and hence $c^{*2}(D, A^\dagger)/\{\sum_{j=1}^{p} (d_j + \gamma_j) a_j^{\dagger 2}\} = M_\nu$ by Corollary 3. It appears that the difference between $R\{\delta_{A^\dagger(\Gamma)}, \pi_\Gamma\}$ and $R(\delta_\Gamma^{\mathrm{Bayes}}, \pi_\Gamma)$ tends to be large if ν is large. But $d_1^* \geq \cdots \geq d_\nu^*$ cannot differ too much from each other because by Corollary 1,

$$k - 2 \leq \sum_{j=1}^{k} \frac{d_{k+1}^*}{d_j^*} \leq k \qquad (k = 3, \ldots, \nu - 1).$$

Then the difference between $R\{\delta_{A^\dagger(\Gamma)}, \pi_\Gamma\}$ and $R(\delta_\Gamma^{\mathrm{Bayes}}, \pi_\Gamma)$ should be limited even if ν is large. A careful analysis using these ideas leads to the following result.

Theorem 3. *Suppose that the prior is* $\theta \sim \mathrm{N}(0, \Gamma)$. *If* $\nu = 3$, *then*

$$R\{\delta_{A^\dagger(\Gamma)}, \pi_\Gamma\} \le \mathrm{tr}(D) - \sum_{j=3}^{p} d_j^* + \left(d_3^* - \frac{2}{p-2} \sum_{j=4}^{p} d_j^* - \frac{p}{p-2} \frac{d_3^*}{3} \right) \tag{28}$$

$$\le \mathrm{tr}(D) - \sum_{j=3}^{p} d_j^* + \frac{2}{3} d_3^*. \tag{29}$$

If $\nu \ge 4$, *then*

$$R\{\delta_{A^\dagger(\Gamma)}, \pi_\Gamma\} \le \mathrm{tr}(D) - \sum_{j=3}^{p} d_j^* + \left(d_3^* + d_4^* - \frac{2}{p-2} \sum_{j=5}^{p} d_j^* - \frac{4p}{p-2} \frac{d_\nu^*}{\nu} \right) \tag{30}$$

$$\le \mathrm{tr}(D) - \sum_{j=3}^{p} d_j^* + \left(d_3^* + d_4^* \right). \tag{31}$$

Throughout, *an empty summation is* 0.

There are interesting implications of Theorem 3. By (29) and (31),

$$R\{\delta_{A^\dagger(\Gamma)}, \pi_\Gamma\} \le R\left(\delta_\Gamma^{\mathrm{Bayes}}, \pi_\Gamma \right) + \left(d_1^* + d_2^* + d_3^* + d_4^* \right). \tag{32}$$

Then $\delta_{A^\dagger(\Gamma)}$ achieves almost the minimum Bayes risk if $d_1^* / \{\mathrm{tr}(D) - \sum_{j=1}^{p} d_j^*\} \approx 0$. In terms of Bayes risk reduction, the bound (32) shows that

$$\mathrm{tr}(D) - R\{\delta_{A^\dagger(\Gamma)}, \pi_\Gamma\} \ge \left(1 - \frac{d_1^* + d_2^* + d_3^* + d_4^*}{\sum_{j=1}^{p} d_j^*} \right) \left\{ \mathrm{tr}(D) - R\left(\delta_\Gamma^{\mathrm{Bayes}}, \pi_\Gamma \right) \right\}.$$

Therefore, $\delta_{A^\dagger(\Gamma)}$ achieves Bayes risk reduction within a negligible factor of that achieved by the Bayes rule if $d_1^* / \sum_{j=1}^{p} d_j^* \approx 0$.

In the homoscedastic case where both $D = \sigma^2 I$ and $\Gamma = \gamma I$, $\delta_{A^\dagger}$ reduces to $\delta_{p-2}^{\mathrm{JS}}$, regardless of $\gamma \ge 0$ (Section 3.3). Then the bounds (28) and (30) become exact and give Efron and Morris's [17] result that $R(\delta_{p-2}^{\mathrm{JS}}, \pi_{\gamma I}) = \mathrm{tr}(D) - (p-2)\{\sigma^4/(\sigma^2 + \gamma)\}$ or equivalently $\mathrm{tr}(D) - R(\delta_{p-2}^{\mathrm{JS}}, \pi_{\gamma I}) = (1 - 2/p)\{\mathrm{tr}(D) - R(\delta_{\gamma I}^{\mathrm{Bayes}}, \pi_{\gamma I})\}$.

It is interesting to compare the Bayes risk bound of $\delta_{A^\dagger(\Gamma)}$ with that of the following simpler version of Berger's [5] estimator δ^{MB}:

$$\delta_j^{\mathrm{MB2}} = X_j - \left\{ \frac{1}{d_j^*} \sum_{k=j}^{p} \left(d_k^* - d_{k+1}^* \right) \frac{(k-2)_+}{\sum_{\ell=1}^{k} X_\ell^2 / (d_\ell + \gamma_\ell)} \right\} \frac{d_j}{d_j + \gamma_j} X_j.$$

By Berger [5], δ^{MB2} is minimax and

$$R\big(\delta^{\mathrm{MB2}}, \pi_\Gamma\big) = \mathrm{tr}(D) - \sum_{j=3}^{p} d_j^* - 2\sum_{j=3}^{p} \frac{d_j^*}{j}\left(1 - \frac{d_j^*}{j-1}\sum_{k=1}^{j-1}\frac{1}{d_k^*}\right) \tag{33}$$

$$\leq \mathrm{tr}(D) - \sum_{j=3}^{p} d_j^*. \tag{34}$$

There seems to be no definite comparison between the bounds (28) and (30) on $R\{\delta_{A^\dagger(\Gamma)}, \pi_\Gamma\}$ and the exact expression (33) for $R(\delta^{\mathrm{MB2}}, \pi_\Gamma)$, although the simple bounds (29) and (31) is slightly higher, by at most $d_3^* + d_4^*$, than the bound (34). Of course, each risk upper bound gives a conservative estimate of the actual performance, and comparison of two upper bounds should be interpreted with caution. In fact, the positive-part estimator $\delta_{A^\dagger}^+$ yields lower risks than those of the non-simplified estimator δ^{MB} in our simulation study (Section 4).

The simplicity of $\delta_{A^\dagger}$ and $\delta_{A^\dagger}^+$ makes it easy to further study them in other ways than using the Bayes (or average) risk. No similar result to the following Theorem 4 has been established for δ^{MB} or δ^{MB2}. Corresponding to the prior $\mathrm{N}(0, \Gamma)$, consider the worst-case (or maximum) risk

$$R(\delta, \mathcal{H}_\Gamma) = \sup_{\theta \in \mathcal{H}_\Gamma} R(\delta, \theta)$$

over the hyper-rectangle $\mathcal{H}_\Gamma = \{\theta\colon\ \theta_j^2 \leq \gamma_j,\ j = 1, \ldots, p\}$ (e.g., Donoho *et al.* [16]). Applying Jensen's inequality to (13) shows that if $c^*(D, A) > 0$, then

$$R(\delta_A, \theta) \leq \mathrm{tr}(D) - \frac{c^{*2}(D, A)}{\sum_{j=1}^{p}(d_j + \theta_j^2)a_j^2},$$

which immediately leads to

$$R(\delta_A, \mathcal{H}_\Gamma) \leq \mathrm{tr}(D) - \frac{c^{*2}(D, A)}{\sum_{j=1}^{p}(d_j + \gamma_j)a_j^2}. \tag{35}$$

By the discussion after (20), a direct application of Jensen's inequality to (19) shows that the Bayes risk $R(\delta_A, \pi_\Gamma)$ is also no greater than the right-hand side of (35), whereas inequality (20) leads to a strictly tighter bound (21). Nevertheless, the upper bound (35) on the worst-case risk of $\delta_{A^\dagger(\Gamma)}$ gives

$$R\{\delta_{A^\dagger(\Gamma)}, \mathcal{H}_\Gamma\} \leq \mathrm{tr}(D) - M_\nu = \mathrm{tr}(D) - \left\{\frac{(\nu-2)^2}{\sum_{j=1}^{\nu} d_j^{*-1}} + \sum_{j=\nu+1}^{p} d_j^*\right\},$$

similarly as how (21) leads to (27) on the Bayes risk of $\delta_{A^\dagger(\Gamma)}$. Therefore, the following result holds by the same proof of Theorem 3.

Theorem 4. *Suppose that* $\mathcal{H}_\Gamma = \{\theta\colon\ \theta_j^2 \le \gamma_j,\ j = 1, \ldots, p\}$. *If* $\nu = 3$, *then*

$$R\{\delta_{A^\dagger(\Gamma)}, \mathcal{H}_\Gamma\} \le \operatorname{tr}(D) - \sum_{j=3}^{p} d_j^* + \frac{2}{3} d_3^*.$$

If $\nu \ge 4$, *then*

$$\begin{aligned} R\{\delta_{A^\dagger(\Gamma)}, \mathcal{H}_\Gamma\} &\le \operatorname{tr}(D) - \sum_{j=3}^{p} d_j^* + \left(d_3^* + d_4^* - 4\frac{d_\nu^*}{\nu} \right) \\ &\le \operatorname{tr}(D) - \sum_{j=3}^{p} d_j^* + \left(d_3^* + d_4^* \right). \end{aligned}$$

There are similar implications of Theorem 4 to those of Theorem 3. By Donoho *et al.* [16], the minimax linear risk over $\mathcal{H}_\Gamma$, $R^L(\mathcal{H}_\Gamma) = \inf_{\delta\,\text{linear}} R(\delta, \mathcal{H}_\Gamma)$, coincides with the minimum Bayes risk $R(\delta_\Gamma^{\text{Bayes}}, \pi_\Gamma)$, and is no greater than 1.25 times the minimax risk over $\mathcal{H}_\Gamma$, $R^N(\mathcal{H}_\Gamma) = \inf_\delta R(\delta, \mathcal{H}_\Gamma)$. These results are originally obtained in the homoscedastic case ($d_1 = \cdots = d_p$), but they remain valid in the heteroscedastic case by the independence of the observations X_j and the separate constraints on θ_j. Therefore, a similar result to (32) holds:

$$\begin{aligned} R\{\delta_{A^\dagger(\Gamma)}, \mathcal{H}_\Gamma\} &\le R^L(\mathcal{H}_\Gamma) + \left(d_1^* + d_2^* + d_3^* + d_4^* \right) \\ &\le 1.25 R^N(\mathcal{H}_\Gamma) + \left(d_1^* + d_2^* + d_3^* + d_4^* \right). \end{aligned}$$

If $d_1^* / \{\operatorname{tr}(D) - \sum_{j=1}^p d_j^*\} \approx 0$, then $\delta_{A^\dagger}$ achieves almost the minimax linear risk (or the minimax risk up to a factor of 1.25) over the hyper-rectangle $\mathcal{H}_\Gamma$, in addition to being globally minimax with θ unrestricted.

The foregoing results might be considered non-adaptive in that $\delta_{A^\dagger(\Gamma)}$ is evaluated with respect to the prior $\mathrm{N}(0, \Gamma)$ or the parameter set $\mathcal{H}_\Gamma$ with the same Γ used to construct $\delta_{A^\dagger(\Gamma)}$. But, by the invariance of δ_A under scale transformations of A, $\delta_{A^\dagger(\Gamma)}$ is identical to the estimator, $\delta_{A^\dagger(\Gamma_\alpha)}$, that would be obtained if Γ is replaced by $\Gamma_\alpha = \alpha(D + \Gamma) - D$ for any scalar α such that the diagonal matrix Γ_α is nonnegative definite. By Theorems 3–4, this observation leads directly to the following adaptive result. In contrast, no adaptive result seems possible for δ^{MB}.

Corollary 4. *Let* $\Gamma_\alpha = \alpha(D + \Gamma) - D$ *and* $\alpha_0 = \max_{j=1,\ldots,p}\{d_j/(d_j + \gamma_j)\}$ (≤ 1). *Then for each* $\alpha \ge \alpha_0$,

$$\begin{aligned} \max\left[R\{\delta_{A^\dagger(\Gamma)}, \pi_{\Gamma_\alpha}\}, R\{\delta_{A^\dagger(\Gamma)}, \mathcal{H}_{\Gamma_\alpha}\} \right] &\le R\left(\delta_{\Gamma_\alpha}^{\text{Bayes}}, \pi_{\Gamma_\alpha}\right) + \alpha^{-1}\left(d_1^* + d_2^* + d_3^* + d_4^* \right) \\ &= R^L(\mathcal{H}_{\Gamma_\alpha}) + \alpha^{-1}\left(d_1^* + d_2^* + d_3^* + d_4^* \right), \end{aligned}$$

where $R(\delta_{\Gamma_\alpha}^{\text{Bayes}}, \pi_{\Gamma_\alpha}) = \operatorname{tr}(D) - \alpha^{-1} \sum_{j=1}^p d_j^*$.

For fixed Γ, $\delta_{A^\dagger(\Gamma)}$ can achieve close to the minimum Bayes risk or the minimax linear risk with respect to each prior in the class $\{\mathrm{N}(0, \Gamma_\alpha)\colon\ \alpha \geq \alpha_0\}$ or each parameter set in the class $\{\mathcal{H}_{\Gamma_\alpha}\colon\ \alpha \geq \alpha_0\}$ under mild conditions. For illustration, consider the case of a heteroscedastic prior with $\Gamma \propto D$. Then $\{\Gamma_\alpha\colon\ \alpha \geq \alpha_0\}$ can be reparameterized as $\{\gamma D\colon\ \gamma \geq 0\}$. By Corollary 4, for each $\gamma \geq 0$,

$$\max\bigl\{R(\delta_{A_0^\dagger}, \pi_{\gamma D}), R(\delta_{A_0^\dagger}, \mathcal{H}_{\gamma D})\bigr\} \leq R\bigl(\delta_{\gamma D}^{\mathrm{Bayes}}, \pi_{\gamma D}\bigr) + \frac{d_1 + d_2 + d_3 + d_4}{1 + \gamma},$$

where $R(\delta_{\gamma D}^{\mathrm{Bayes}}, \pi_{\gamma D}) = \{\gamma/(1+\gamma)\}\,\mathrm{tr}(D)$ and $d_1 \geq d_2 \geq \cdots \geq d_p$. Therefore, if $d_1/\mathrm{tr}(D) \approx 0$, then $\delta_{A_0^\dagger}$ achieves the minimum Bayes risk, within a negligible factor, under the prior $\mathrm{N}(0, \gamma D)$ for each $\gamma > 0$. This can be seen as an extension of the result that in the homoscedastic case, $\delta_{p-2}^{\mathrm{JS}}$ asymptotically achieves the minimum Bayes risk under the prior $\mathrm{N}(0, \gamma I)$ for each $\gamma > 0$ as $p \to \infty$.

Finally, we compare the estimator $\delta_{A^\dagger}$ with a block shrinkage estimator, suggested by the differentiation in the shrinkage of low- and high-variance coordinates by $\delta_{A^\dagger}$. Consider the estimator

$$\delta^{\mathrm{block}} = \begin{Bmatrix} \delta_{\tau-2}^{\mathrm{B}}(X_1, \ldots, X_\tau) \\ \delta_{p-\tau-2}^{\mathrm{B}}(X_{\tau+1}, \ldots, X_p) \end{Bmatrix},$$

where τ is a cutoff index, and $\delta_c^{\mathrm{B}}(Y) = Y$ if Y is of dimension 1 or 2. The index τ can be selected such that the coordinate variances are relatively homogeneous in each block. Alternatively, a specific strategy for selecting τ is to minimize an upper bound on the Bayes risk of δ^{block}, similarly as in the development of $\delta_{A^\dagger}$. Applying (21) with $A = D^{-1}$ to $\delta_{p-2}^{\mathrm{B}}$ in the two blocks shows that $R(\delta^{\mathrm{block}}, \pi_\Gamma) \leq \mathrm{tr}(D) - L_\tau$, where

$$L_k = \frac{k-2}{(1/k)\sum_{j=1}^{k}(d_j + \gamma_j)/d_j^2} + \frac{p-k-2}{(1/(p-k))\sum_{j=k+1}^{p}(d_j + \gamma_j)/d_j^2}.$$

The first (or second) term in L_k is set to 0 if $k \leq 2$ (or $k \geq p-2$). Then τ can be defined as the smallest index such that $L_\tau = \max(L_1, L_2, \ldots, L_p)$. But the upper bound (27) on $R(\delta_{A^\dagger}, \pi_\Gamma)$ is likely to be smaller than the corresponding bound on $R(\delta^{\mathrm{block}}, \pi_\Gamma)$, because $\{k/(k-2)\}M_k \geq L_k$ for each $k \geq 3$ by the Cauchy–Schwarz inequality $\{\sum_{j=k+1}^{p} d_j^2/(d_j + \gamma_j)\}\{\sum_{j=k+1}^{p}(d_j + \gamma_j)/d_j^2\} \geq (p-k)^2$. Therefore, $\delta_{A^\dagger}$ tends to yield greater risk reduction than δ^{block}. This analysis also indicates that $\delta_{A^\dagger}$ can be advantageous over δ^{block} extended to multiple blocks.

The rationale of forming blocks in $\delta_{A^\dagger}$ and δ^{block} differs from that in existing block shrinkage estimators (e.g., Brown and Zhao [13]). As discussed in Cai [14], block shrinkage has been developed mainly in the homoscedastic case as a technique for pooling information: the coordinate means are likely to be similar to each other within a block. Nevertheless, it is possible to both deal with heterogeneity among coordinate variances and exploit homogeneity among coordinate means within individual blocks in our approach using a block-homoscedastic prior (i.e., the prior variances are equal within each block). This topic can be pursued in future work.

4. Simulation study

4.1. Setup

We conduct a simulation study to compare the following 8 estimators,

(i) Non-minimax estimators: δ^{EB} by (2), δ^{XKB} by (3), δ^{RB} by (7) with $\Gamma = 0$;
(ii) Minimax estimators: $\delta^{\mathrm{B}+}_{p-2}$ by (6), δ^{MB} by (8) with $\Gamma = 0$ or γI for some large γ, δ^{+}_{A} by (12) with $A = A^{\dagger}_{0}$ and $A^{\dagger}_{\infty}$.

Recall that $A^{\dagger}_{0}$ corresponds to $\Gamma = 0$ or $\Gamma \propto D$ and $A^{\dagger}_{\infty}$ corresponds to $\Gamma = \gamma I$ with $\gamma \to \infty$. In contrast, letting the diagonal elements of Γ tend to ∞ in any direction in δ^{RB} and δ^{MB} leads to $\delta_0 = X$. Setting Γ to 0 or ∞ is used here to specify the relevant estimators, rather than to restrict the prior on θ.

For completeness, we also study the following estimators: $\delta^{\mathrm{B}+}_{2(p-2)}$ by (6), δ^{RB} with $p-2$ replaced by $2(p-2)$ in (7), δ^{MB} with $(k-2)_+$ replaced by $2(k-2)_+$ in (8), and δ^{+}_{A} with $c^*(D, A)$ replaced by $2c^*(D, A)$ in (12), referred to as the alternative versions of $\delta^{\mathrm{B}+}_{p-2}$, δ^{RB}, δ^{MB}, and δ^{+}_{A} respectively. The usual choices of the factors, $p-2$, $(k-2)_+$, and $c^*(D, A)$, are motivated to minimize the risks of the non-positive-part estimators, but may not be the most desirable for the positive-part estimators. As seen below, the alternative choices $2(p-2)$, $2(k-2)_+$, and $2c^*(D, A)$ can lead to risk curves for the positive-part estimators rather different from those based on the usual choices $(p-2)$, $(k-2)_+$, and $c^*(D, A)$. Therefore, we compare the estimators $\delta^{\mathrm{B}+}_{p-2}$, δ^{RB}, δ^{MB}, and δ^{+}_{A} and, separately, their alternative versions.

Each estimator δ is evaluated by the pointwise risk function $R(\delta, \theta)$ as θ moves in a certain direction or the Bayes risk function $R(\delta, \pi)$ as π varies in a set of priors on θ. Consider the homoscedastic prior $\mathrm{N}(0, \eta^2 I/p)$ or the heteroscedastic prior $\mathrm{N}\{0, \eta^2 D/\operatorname{tr}(D)\}$ for $\eta \geq 0$. As discussed in Section 3.3, the Bayes risk with the first or second prior is meant to measure average risk reduction over the region $\{\theta\colon \|\theta\|^2 \leq \eta^2\}$ or $\{\theta\colon \theta^{\mathrm{T}} D^{-1}\theta \leq p\eta^2/\operatorname{tr}(D)\}$. Corresponding to the two priors, consider the direction along $(\eta/\sqrt{p}, \ldots, \eta/\sqrt{p})$ or $(\eta\sqrt{d_1}, \ldots, \eta\sqrt{d_p})/\sqrt{\operatorname{tr}(D)}$, where η gives the Euclidean distance from 0 to the point indexed by η. The two directions are referred to as the homoscedastic and heteroscedastic directions.

We investigate several configurations for D, including (5) and

$$(d_1, d_2, \ldots, d_{10}) = (40, 20, 10, 5, 5, 5, 1, 1, 1, 1) \quad \text{or} \tag{36}$$

$$= (40, 20, 10, 7, 6, 5, 4, 3, 2, 1) \quad \text{or} \tag{37}$$

$$= 5\%, 15\%, \ldots, 95\% \text{ quantiles of } 8/\chi^2_3 \text{ or } 24/\chi^2_5,$$

where χ^2_k is a chi-squared variable with k degrees of freedom. In the last case, $(d_1, \ldots, d_{10})$ can be considered a typical sample from a scaled inverse chi-squared distribution, which is the conjugate distribution for normal variances. In the case (36), the coordinates may be segmented intuitively into three groups with relatively homogeneous variances. In the case (37), there is no clear intuition about how the coordinates should be segmented into groups.

For fixed D, the pointwise risk $R(\delta, \theta)$ is computed by repeatedly drawing $X \sim \mathrm{N}(\theta, D)$ and then taking the average of $\|\delta - \theta\|^2$. The Bayes risk is computed by repeatedly drawing $\theta \sim$

$\mathrm{N}(0, \Gamma)$ and $X|\theta \sim \mathrm{N}(\theta, D)$ and then taking the average of $\|\delta - \theta\|^2$. Each Monte Carlo sample size is set to 10^5.

4.2. Results

The relative performances of the estimators are found to be consistent across different configurations of D studied. Moreover, the Bayes risk curves under the homoscedastic prior are similar to the pointwise risk curves along the homoscedastic direction. The Bayes risk curves under the heteroscedastic prior are similar to the pointwise risk curves along the heteroscedastic direction. Figure 1 shows the pointwise risks of the estimators with the usual versions of $\delta^{\mathrm{B}+}_{p-2}$, δ^{RB}, δ^{MB}, and δ^{+}_{A} and Figure 2 shows those of the estimators with the alternative versions of $\delta^{\mathrm{B}+}_{p-2}$, δ^{RB}, δ^{MB}, and δ^{+}_{A} for the case (36), with roughly three groups of coordinate variances, which might be considered unfavorable to our approach. For both $A^{\dagger}_{0}$ and $A^{\dagger}_{\infty}$, the cutoff index ν is found to be 3. See the supplementary material (Tan [30]) for the Bayes risk curves of all these estimators for the case (36) and the results for other configurations of D.

A number of observations can be drawn from Figures 1–2. First, δ^{EB}, δ^{XKB}, and δ^{RB} have among the lowest risk curves along the homoscedastic direction. But along the heteroscedastic direction, the risk curves of δ^{EB} and δ^{XKB} rise quickly above the constant risk of X as η increases. Moreover, all the risk curves of δ^{EB}, δ^{XKB}, and δ^{RB} along the θ_1 axis exceed the constant risk of X as $|\theta_1|$ increases. Therefore, δ^{EB}, δ^{XKB}, and δ^{RB} fail to be minimax, as mentioned in Section 2.

Second, $\delta^{\mathrm{B}+}_{p-2}$ or $\delta^{\mathrm{B}+}_{2(p-2)}$ has among the highest risk curve, except where the risk curves of δ^{EB} and δ^{XKB} exceed the constant risk of X along the heteroscedastic direction. The poor performance is expected for $\delta^{\mathrm{B}+}_{p-2}$ or $\delta^{\mathrm{B}+}_{2(p-2)}$, because there are considerable differences between the coordinate variances in (36).

Third, among the minimax estimators, δ^{+}_{A} with $A = A^{\dagger}_{0}$ or $A^{\dagger}_{\infty}$ has the lowest risk curve along various directions, whether the usual versions of $\delta^{\mathrm{B}+}_{p-2}$, δ^{MB}, and δ^{+}_{A} are compared (Figure 1) or the alternative versions are compared (Figure 2).

Fourth, the risk curve of δ^{+}_{A} with $A = A^{\dagger}_{0}$ is similar to that of δ^{+}_{A} with $A = A^{\dagger}_{\infty}$ along the heteroscedastic direction. But the former is noticeably higher than the latter along the homoscedastic direction as η increases, whereas is noticeably lower than the latter along the θ_1 axis as $|\theta_1|$ increases. These results agree with the construction of $A^{\dagger}_{0}$ using a heteroscedastic prior and $A^{\dagger}_{\infty}$ using a flat, homoscedastic prior. Their relative performances depend on the direction in which the risks are evaluated.

Fifth, δ^{MB} with $\Gamma = 0$ has risk curves below that of $\delta^{\mathrm{B}+}_{p-2}$ or $\delta^{\mathrm{B}+}_{2(p-2)}$, but either above or crossing those of δ^{+}_{A} with $A = A^{\dagger}_{0}$ and $A^{\dagger}_{\infty}$. Moreover, δ^{MB} with $\Gamma = (16^2/p)I$ has elevated, almost flat risk curves for η from 0 to 16. This seems to indicate an undesirable consequence of using a non-degenerate prior for δ^{MB} in that the risk tends to increase for θ near 0, and remains high for θ far away from 0.

The foregoing discussion involves the comparison of the risk curves as θ moves away from 0 between δ^{MB} and $\delta^{+}_{A^{\dagger}}$ specified with fixed priors. Alternatively, we compare the pointwise risks at $\theta = (\eta/\sqrt{p}, \ldots, \eta/\sqrt{p})$ or $(\eta\sqrt{d_1}, \ldots, \eta\sqrt{d_p})/\sqrt{\mathrm{tr}(D)}$ and the Bayes risks under the prior $\mathrm{N}(0, \eta^2 I/p)$ or $\mathrm{N}\{0, \eta^2 D/\mathrm{tr}(D)\}$ between δ^{MB} and $\delta^{+}_{A^{\dagger}}$ specified with the prior $\mathrm{N}(0, \eta^2 I/p)$ for

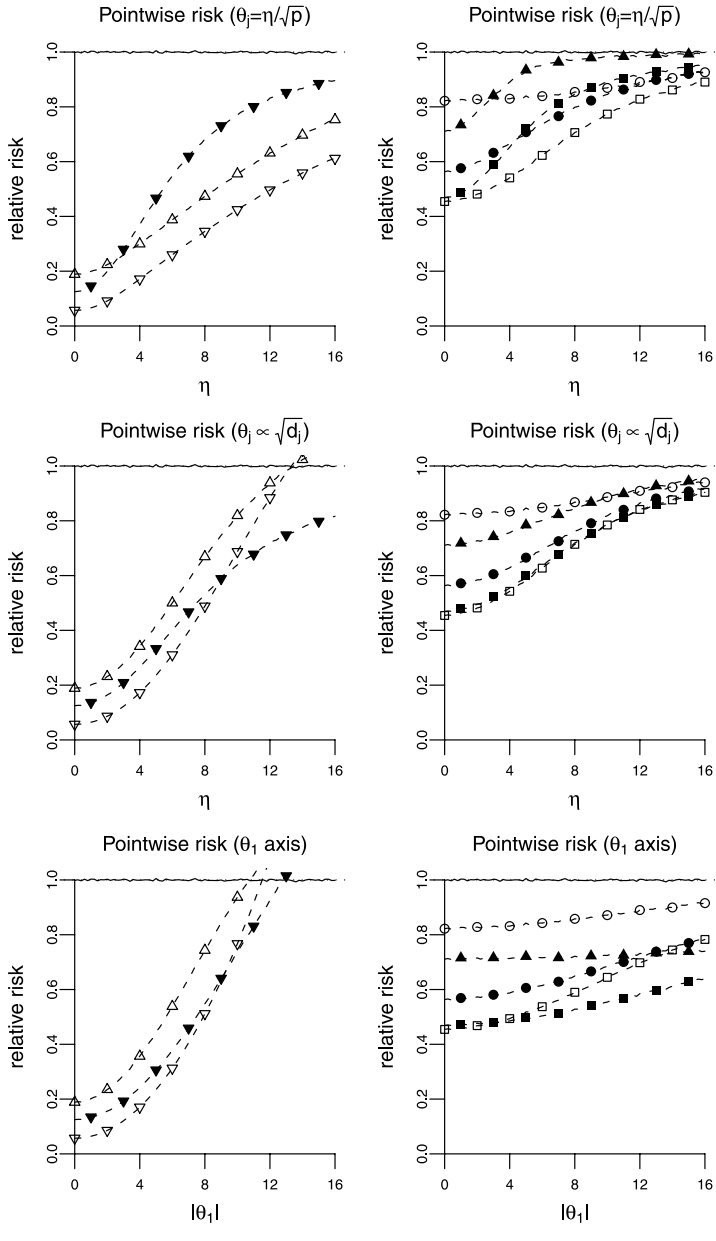

Figure 1. Pointwise risks along the homoscedastic (first row) and heteroscedastic (second row) directions and θ_1 axis (third row) in the case (36). Left: non-minimax estimators δ^{EB} ($\triangledown$), δ^{RB} ($\blacktriangledown$), δ^{XKB} ($\triangle$). Right: minimax estimators $\delta^{\mathrm{B}+}_{p-2}$ ($\blacktriangle$), δ^{MB} with $\Gamma = 0$ ($\bullet$) and $\Gamma = (16^2/p)I$ ($\circ$), δ^{+}_{A} with $A = A_0^{\dagger}$ ($\blacksquare$) and $A_{\infty}^{\dagger}$ ($\square$).

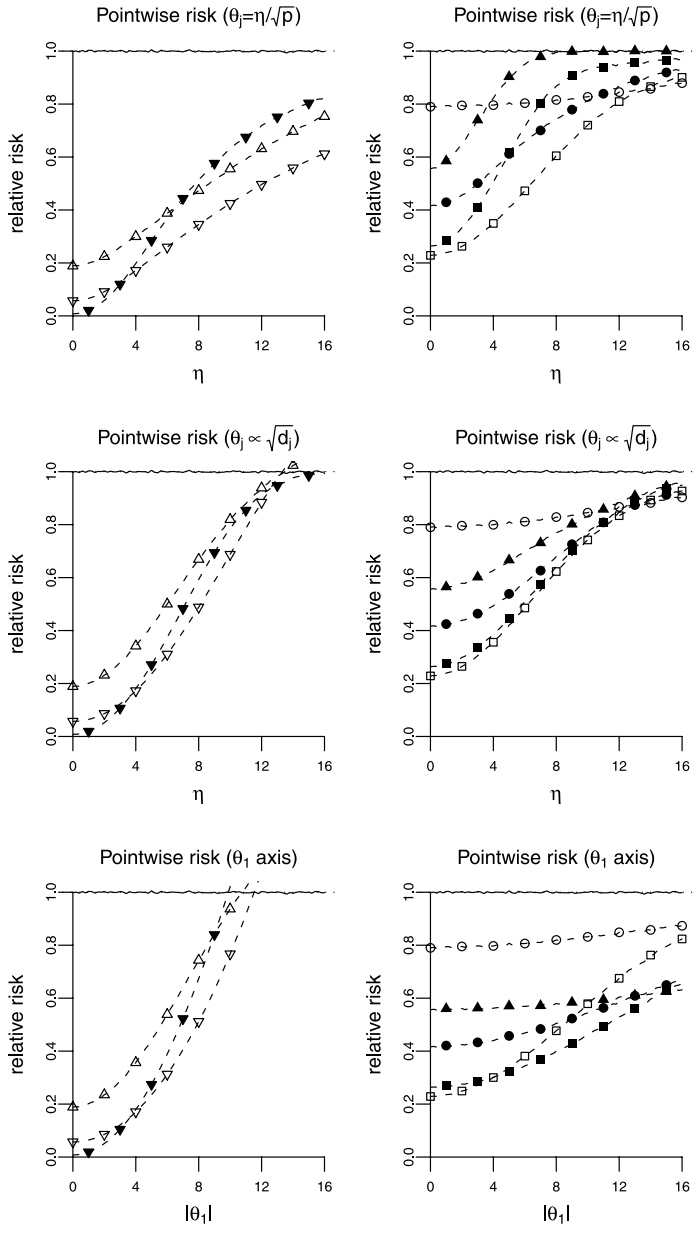

Figure 2. Pointwise risks along the homoscedastic (first row) and heteroscedastic (second row) directions and θ_1 axis (third row) in the case (36), with the same legend as in Figure 1. The alternative versions of δ^{B+}_{p-2}, δ^{RB}, δ^{MB}, and δ^{+}_{A} are used.

a range of η. The homoscedastic prior used in the specification of δ^{MB} and $\delta^{+}_{A^\dagger}$ can be considered correctly specified or misspecified, when the Bayes risks are evaluated under, respectively, the homoscedastic or heteroscedastic prior or when the pointwise risks are evaluated along the homoscedastic or heteroscedastic direction. For each situation, $\delta^{+}_{A^\dagger}$ has lower pointwise or Bayes risks than δ^{MB}. See Figure A2 in the supplementary material (Tan [30]).

5. Conclusion

The estimator $\delta_{A^\dagger}$ and its positive-part version $\delta^{+}_{A^\dagger}$ are not only minimax and but also have desirable properties including simplicity, interpretability, and effectiveness in risk reduction. In fact, $\delta_{A^\dagger}$ is defined by taking $A = A^\dagger$ in a class of minimax estimators δ_A. The simplicity of $\delta_{A^\dagger}$ holds because δ_A is of the linear form $(I - \lambda A)X$, with A and λ indicating the direction and magnitude of shrinkage. The interpretability of $\delta_{A^\dagger}$ holds because the form of $A^\dagger$ indicates that one group of coordinates are shrunk in the direction of Berger's [3] minimax estimator whereas the remaining coordinates are shrunk in the direction of the Bayes rule. The effectiveness of $\delta_{A^\dagger}$ in risk reduction is supported, in theory, by showing that $\delta_{A^\dagger}$ can achieve close to the minimum Bayes risk simultaneously over a scale class of normal priors (Corollary 4). For various scenarios in our numerical study, the estimators $\delta^{+}_{A^\dagger}$ with extreme priors yield more substantial risk reduction than existing minimax estimators.

It is interesting to discuss a special feature of $\delta_{A,r}$ and hence of $\delta_{A,c}$ and δ_A among linear, shrinkage estimators of the form

$$\delta = X - h\big(X^{\mathrm{T}}BX\big)AX, \tag{38}$$

where A and B are nonnegative definite matrices and $h(\cdot)$ is a scalar function. The estimator $\delta_{A,r}$ corresponds to the choice $B \propto A^{\mathrm{T}}QA$, which is motivated by the form of the optimal λ in minimizing the risk of $(I - \lambda A)X$ for fixed A. On the other hand, Berger and Srinivasan [8] showed that under certain regularity conditions on $h(\cdot)$, an estimator (38) can be generalized Bayes or admissible only if $B \propto \Sigma^{-1}A$. This condition is incompatible with $B \propto A^{\mathrm{T}}QA$, unless $A \propto Q^{-1}\Sigma^{-1}$ as in Berger's [3] estimator. Therefore, δ_A including $\delta_{A^\dagger}$ is, in general, not generalized Bayes or admissible. This conclusion, however, does not apply directly to the positive-part estimator δ^{+}_{A}, which is no longer of the linear form $(I - \lambda A)X$.

There are various topics that can be further studied. First, the prior on θ is fixed, independently of data in the current paper. A useful extension is to allow the prior to be estimated within a certain class, for example, homoscedastic priors $N(0, \gamma I)$, from the data, in the spirit of empirical Bayes estimation (e.g., Efron and Morris [17]). Second, the Bayes risk with a normal prior is used to measure average risk reduction in an elliptical region (Section 3.3). It is interesting to study how our approach can be extended when using a non-normal prior on θ, corresponding to a non-elliptical region in which risk reduction is desired.

Appendix

Preparation. The following extends Stein's [27] lemma for computing the expectation of the inner product of $X - \theta$ and a vector of functions of X.

Lemma 1. *Let* $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ *be multivariate normal with mean* θ *and variance matrix* Σ. *Assume that* $g = (g_1, \ldots, g_p)^{\mathrm{T}} : \mathcal{R}^p \to \mathcal{R}^p$ *is almost differentiable Stein* [27] *with* $E_\theta\{|\nabla_j g_i(X)|\} < \infty$ *for* $i, j = 1, \ldots, p$, *where* $\nabla_j = \partial/\partial x_j$. *Then*

$$E_\theta\big\{(X-\theta)^{\mathrm{T}} g(X)\big\} = \mathrm{tr}\big[\Sigma E_\theta\big\{\nabla g(X)\big\}\big],$$

where $\nabla g(x)$ *is the matrix with* (i, j)*th element* $\nabla_j g_i(x)$.

Proof. A direct generalization of Lemma 2 in Stein [27] to a normal random vector with non-identity variance matrix gives

$$E_\theta\big\{(X-\theta) g_i(X)\big\} = \Sigma E_\theta^{\mathrm{T}}\big\{\nabla g_i(X)\big\},$$

where $\nabla g_i(x)$ is the row vector with jth element $\nabla_j g_i(x)$. Taking the ith element of both sides of the equation gives

$$E_\theta\big\{(X_i-\theta_i) g_i(X)\big\} = \sum_{j=1}^{p} \sigma_{ij} E_\theta\big\{\nabla_j g_i(X)\big\},$$

where σ_{ij} is the (i, j)th element of Σ. Summing both sides of the preceding equation over i gives the desired result. □

Proof of Theorem 1. By direct calculation, the risk of $\delta_{A,r}$ is

$$R(\delta_{A,r}, \theta) = \mathrm{tr}(\Sigma Q) + E_\theta\left(\frac{r^2}{X^{\mathrm{T}} A^{\mathrm{T}} Q A X}\right) - 2E_\theta\left\{(X-\theta)^{\mathrm{T}} \frac{r Q A X}{X^{\mathrm{T}} A^{\mathrm{T}} Q A X}\right\}.$$

By Lemma 1 and the fact that $\mathrm{tr}(\Sigma Q A X X^{\mathrm{T}} A^{\mathrm{T}} Q A) = X^{\mathrm{T}} A^{\mathrm{T}} Q A \Sigma Q A X$, the third term after the minus sign in $R(\delta_{A,r}, \theta)$ is

$$2E_\theta\left\{r \frac{\mathrm{tr}(\Sigma Q A)}{X^{\mathrm{T}} A^{\mathrm{T}} Q A X}\right\} - 4E_\theta\left\{r \frac{X^{\mathrm{T}} A^{\mathrm{T}} Q A \Sigma Q A X}{(X^{\mathrm{T}} A^{\mathrm{T}} Q A X)^2}\right\} + 4E_\theta\left(r' \frac{X^{\mathrm{T}} A^{\mathrm{T}} Q A \Sigma Q A X}{X^{\mathrm{T}} A^{\mathrm{T}} Q A X}\right).$$

By condition (14), $A^{\mathrm{T}} Q A \Sigma Q A$ is nonnegative definite. By Section 21.14 and Exercise 21.32 in Harville [18], $(x^{\mathrm{T}} A^{\mathrm{T}} Q A \Sigma Q A x)/(x^{\mathrm{T}} A^{\mathrm{T}} Q A x) \le \lambda_{\max}(A\Sigma Q + \Sigma A^{\mathrm{T}} Q)/2$ for $x \neq 0$. Then the preceding expression is bounded from below by

$$2E_\theta\left\{r \frac{\mathrm{tr}(\Sigma Q A) - \lambda_{\max}(A\Sigma Q + \Sigma A^{\mathrm{T}} Q)}{X^{\mathrm{T}} A^{\mathrm{T}} Q A X}\right\},$$

which leads immediately to the upper bound on $R(\delta_{A,r}, \theta)$. □

Proof for condition (17). We show that if condition (17) holds, then there exists a nonsingular matrix B with the claimed properties. The converse is trivially true. Let R be the

unique symmetric, positive definite matrix such that $R^2 = Q$. Then RAR^{-1} is symmetric, that is, $RAR^{-1} = R^{-1}A^{\mathrm{T}}R$, because $QA = A^{\mathrm{T}}Q$. Moreover, $R\Sigma R$ and RAR^{-1} commute, that is, $RAR^{-1}(R\Sigma R) = R\Sigma R(RAR^{-1})^{\mathrm{T}} = R\Sigma R(RAR^{-1})$, because $A\Sigma = \Sigma A^{\mathrm{T}}$ and RAR^{-1} is symmetric. Therefore, $R\Sigma R$ and RAR^{-1} are simultaneously diagonalizable (Harville [18], Section 21.13). There exists an orthogonal matrix O such that $O(R\Sigma R)O^{\mathrm{T}} = D$ and $O(RAR^{-1})O^{\mathrm{T}} = A^*$ for some diagonal matrices D and A^*. Then $B = OR$ satisfies the claimed properties. □

Proof of inequality (20). We show that if $(Z_1, \ldots, Z_p)$ are independent standard normal variables, then $E\{(\sum_{j=1}^p a_j^2 Z_j^2)^{-1}\} \geq \{p/(p-2)\}(\sum_{j=1}^p a_j^2)^{-1}$. Let $S = \sum_{j=1}^p Z_j^2$. Then S and $(Z_1^2/S, \ldots, Z_p^2/S)$ are independent, $S \sim \chi_p^2$, and $(Z_1^2/S, \ldots, Z_p^2/S) \sim \text{Dirichlet}(1/p, \ldots, 1/p)$. The claimed inequality follows because $E\{(\sum_{j=1}^p a_j^2 Z_j^2)^{-1}\} = E\{(\sum_{j=1}^p a_j^2 Z_j^2/S)^{-1}\}E(S^{-1})$, $E(S^{-1}) = 1/(p-2)$, and $E\{(\sum_{j=1}^p a_j^2 Z_j^2/S)^{-1}\} \geq (\sum_{j=1}^p a_j^2/p)^{-1}$ by Jensen's inequality. □

Proofs of Theorem 2 and Corollary 2. Consider the transformation $\delta_j = d_j^2/(d_j + \gamma_j)$ and $\alpha_j = \{(d_j + \gamma_j)/d_j\}a_j$, so that $\delta_j\alpha_j = d_j a_j$ and $\delta_j\alpha_j^2 = (d_j + \gamma_j)a_j^2$. Problem (22) is then transformed to $\max_{\alpha_1,\ldots,\alpha_p}\{\sum_{j=1}^p \delta_j\alpha_j - 2\max(\delta_1\alpha_1, \ldots, \delta_p\alpha_p)\}$, subject to $\alpha_j \geq 0$ $(j = 1, \ldots, p)$ and $\sum_{j=1}^p \delta_j\alpha_j^2 = \sum_{j=1}^p \delta_j$, which is of the form of the special case of (22) with $\gamma_j = 0$ $(j = 1, \ldots, p)$. But it is easy to verify that if the claimed results hold for the transformed problem, then the results hold for original problem (22). Therefore, assume in the rest of proof that $\gamma_j = 0$ $(j = 1, \ldots, p)$.

There exists at least a solution, $A^\dagger$, to problem (22) by boundedness of the constraint set. Let $\mathcal{K} = \{k\colon\ d_k a_k^\dagger = d_\nu a_\nu^\dagger, k = 1, \ldots, p\}$ and $\mathcal{K}^c = \{j\colon\ d_j a_j^\dagger < d_\nu a_\nu^\dagger, j = 1, \ldots, p\}$. A key of the proof is to exploit the fact that, by the setup of problem (22), $(a_1^\dagger, \ldots, a_p^\dagger)$ is automatically a solution to the problem

$$\begin{aligned} &\max_{a_1,\ldots,a_p} \quad \sum_{j=1}^p d_j a_j - 2d_\nu a_\nu, \\ &\text{subject to} \quad a_j \geq 0, \qquad d_j a_j \leq d_\nu a_\nu \qquad (j = 1, \ldots, p), \quad \text{and} \\ &\qquad\qquad \sum_{j=1}^p d_j a_j^2 = \sum_{j=1}^p d_j. \end{aligned} \tag{A.1}$$

The Karush–Kuhn–Tucker condition for this problem gives

$$-1 + 2\lambda a_j^\dagger - d_j^{-1}\rho_j = 0 \qquad \text{for } j \in \mathcal{K}^c, \tag{A.2}$$

$$-1 + 2\lambda a_k^\dagger + \mu_k - d_k^{-1}\rho_k = 0 \qquad \text{for } k\ (\neq \nu) \in \mathcal{K}, \tag{A.3}$$

$$-1 + 2\lambda a_\nu^\dagger + \left(2 - \sum_{k\in\mathcal{K}\setminus\{\nu\}} \mu_k\right) - d_\nu^{-1}\rho_\nu = 0, \tag{A.4}$$

where λ, $\mu_k \geq 0$ ($k \in \mathcal{K} \setminus \{\nu\}$), and $\rho_j \geq 0$ satisfying $\rho_j a_j^\dagger = 0$ ($j = 1, \ldots, p$) are Lagrange multipliers.

First, we show that $a_j^\dagger > 0$ and hence $\rho_j = 0$ for $j = 1, \ldots, p$. If $\mathcal{K}^c = \varnothing$, then either $a_j^\dagger > 0$ for $j = 1, \ldots, p$, or $a_1^\dagger = \cdots = a_p^\dagger = 0$. The latter case is infeasible by the constraint $\sum_{j=1}^p d_j a_j^2 = \sum_{j=1}^p d_j$. Suppose $\mathcal{K}^c \neq \varnothing$. By (A.2), $a_j^\dagger > 0$ for each $j \in \mathcal{K}^c$. Then $a_k^\dagger > 0$ for each $k \in \mathcal{K}$ because $d_k a_k^\dagger > d_j a_j^\dagger$.

Second, we show that $\nu \geq 3$. If $\mathcal{K}^c = \varnothing$, then $\nu = p \geq 3$. Suppose $\mathcal{K}^c \neq \varnothing$. Then $\lambda > 0$ by (A.2). Summing (A.3) over k ($\neq \nu$) $\in \mathcal{K}$ and (A.4) shows that $-\nu + 2\lambda \sum_{k=1}^\nu a_k^\dagger + 2 = 0$. Therefore, $\nu > 2$ or equivalently $\nu \geq 3$.

Third, we show that $\mathcal{K} = \{1, 2, \ldots, \nu\}$ and $\mathcal{K}^c = \{\nu + 1, \ldots, p\}$. For each k ($\neq \nu$) $\in \mathcal{K}$ and $j \in \mathcal{K}^c$, $a_k^\dagger \leq a_j^\dagger$ by (A.2)–(A.3) and then $d_k > d_j$ because $d_k a_k^\dagger > d_j a_j^\dagger$. The inequalities also hold for $k = \nu$, by application of the argument to problem (A.1) with ν replaced by some k ($\neq \nu$) $\in \mathcal{K}$. Then $\mathcal{K}^c = \{\nu + 1, \ldots, p\}$ because $d_\nu > d_j$ for each $j \in \mathcal{K}^c$, $d_1 \geq d_2 \geq \cdots \geq d_p$, and ν is the largest element in $\mathcal{K}$.

Fourth, we show the expressions for $(a_1^\dagger, \ldots, a_p^\dagger)$ and the achieved maximum value. By the definition of $\mathcal{K}$, $a_k^\dagger \propto d_k^{-1}$ for $k = 1, \ldots, \nu$. By (A.2), $a_j^\dagger \propto 1$ for $j = \nu + 1, \ldots, p$. Let $y^\dagger = d_\nu a_\nu^\dagger$ and $z^\dagger = a_{\nu+1}^\dagger$. Then $(y^\dagger, z^\dagger)$ is a solution to the problem

$$\max_{y,z} \quad (\nu - 2)y + \left(\sum_{j=\nu+1}^p d_j\right) z,$$

$$\text{subject to} \quad y \geq 0, \qquad z \geq 0, \qquad y \geq d_{\nu+1} z, \quad \text{and}$$

$$\left(\sum_{k=1}^\nu d_k^{-1}\right) y^2 + \left(\sum_{j=\nu+1}^p d_j\right) z^2 = \sum_{j=1}^p d_j.$$

By the definition of $\mathcal{K}$, $y^\dagger > d_{\nu+1} z^\dagger$ and hence $(y^\dagger, z^\dagger)$ lies off the boundary in the constraint set. Then $(y^\dagger, z^\dagger)$ is a solution to the foregoing problem with the constraint $y \geq d_{\nu+1} z$ removed. The problem is of the form of maximizing a linear function of (y, z) subject to an elliptical constraint. Straightforward calculation shows that

$$y^\dagger = \left(\frac{\sum_{j=1}^p d_j}{M_\nu}\right)^{1/2} \frac{\nu - 2}{\sum_{j=1}^\nu d_j^{-1}}, \qquad z^\dagger = \left(\frac{\sum_{j=1}^p d_j}{M_\nu}\right)^{1/2},$$

and the achieved maximum value is $(\sum_{j=1}^p d_j)^{1/2} M_\nu^{1/2}$, where $M_\nu = (\nu - 2)^2 / (\sum_{j=1}^\nu d_j^{-1}) + \sum_{j=\nu+1}^p d_j$.

Finally, we show that the sequence $(M_3, M_4, \ldots M_p)$ is nonincreasing: $M_k \geq M_{k+1}$, where the equality holds if and only if $k - 2 = \sum_{j=1}^k d_{k+1}/d_j$. Because $y^\dagger > d_{\nu+1} z^\dagger$ or $\nu - 2 > \sum_{j=1}^\nu d_{\nu+1}/d_j$, this result implies that $M_\nu > M_{\nu+1}$ and hence $A^\dagger$ is a unique solution to (22). Let $L_k = \{(\sum_{j=1}^k d_j)(\sum_{j=1}^k d_j^{-1}) - (k-2)^2\} / \sum_{j=1}^k d_j^{-1}$ so that $M_k = \sum_{j=1}^p d_j - L_k$. By the

identity $(b+\beta)/(a+\alpha) - b/a = (\beta/\alpha - b/a)\{\alpha/(a+\alpha)\}$ and simple calculation,

$$L_{k+1} - L_k = \left[\frac{\sum_{j=1}^{k}(d_j/d_{k+1} + d_{k+1}/d_j) - 2k + 4}{d_{k+1}^{-1}} - \left\{ \sum_{j=1}^{k} d_j - \frac{(k-2)^2}{\sum_{j=1}^{k} d_j^{-1}} \right\} \right] \frac{d_{k+1}^{-1}}{\sum_{j=1}^{k+1} d_j^{-1}}$$
$$= d_{k+1} \frac{\{r_k - (k-2)\}^2}{r_k(r_k+1)}, \tag{A.5}$$

where $r_k = \sum_{j=1}^{k} d_{k+1}/d_j$. Therefore, $L_k \le L_{k+1}$. Moreover, $L_k = L_{k+1}$ if and only if $r_k = k-2$, that is, $\sum_{j=1}^{k} d_{k+1}/d_j = k-2$. □

Proof of Corollary 3. It suffices to show (25). By Corollary 2, $\sum_{k=1}^{\nu-1} d_\nu^*/d_k^* \ge \nu - 3$ and hence $\sum_{k=1}^{\nu} d_\nu^*/d_k^* \ge \nu - 2$. Then for $j = 1, \ldots, \nu$.

$$a_j^\dagger = \frac{(\nu-2)d_j^{*-1}}{\sum_{k=1}^{\nu} d_k^{*-1}} \frac{d_j}{d_j+\gamma_j} \le \frac{(\nu-2)d_\nu^{*-1}}{\sum_{k=1}^{\nu} d_k^{*-1}} \frac{d_j}{d_j+\gamma_j} \le \frac{d_j}{d_j+\gamma_j},$$

because $d_j^* \ge d_\nu^*$ for $j \le \nu$. □

Proof of Theorem 3. Let $L_k = \sum_{j=1}^{k} d_j^* - (k-2)^2/\sum_{j=1}^{k} d_j^{*-1}$ so that $M_k = \sum_{j=1}^{p} d_j^* - L_k$, similarly as in the proof of Theorem 2. By equation (A.5) with $r_k = \sum_{j=1}^{k} d_{k+1}^*/d_j^*$ and d_{k+1} replaced by d_{k+1}^*,

$$L_\nu = L_3 + \sum_{k=3}^{\nu-1}(L_{k+1} - L_k) = L_3 + \sum_{k=3}^{\nu-1} d_{k+1}^* \frac{\{r_k - (k-2)\}^2}{r_k(r_k+1)}.$$

By the relationship $r_k = (d_{k+1}^*/d_k^*)(1 + r_{k-1})$ and simple calculation,

$$L_3 = d_1^* + d_2^* + d_3^* - \frac{1}{d_1^{*-1} + d_2^{*-1} + d_3^{*-1}}$$
$$= d_1^* + d_2^* + d_3^* - \sum_{k=3}^{\nu-1} d_{k+1}^* \left(\frac{1}{r_k} - \frac{1}{r_k+1} \right) - \frac{d_\nu^*}{r_{\nu-1}+1}.$$

If $\nu \ge 4$, combining the two preceding equation gives

$$L_\nu = d_1^* + d_2^* + d_3^* + \sum_{k=3}^{\nu-1} d_{k+1}^* \frac{\{r_k - (k-2)\}^2 - 1}{r_k(r_k+1)} - \frac{d_\nu^*}{r_{\nu-1}+1}$$
$$\le d_1^* + d_2^* + d_3^* + \sum_{k=3}^{\nu-1} d_{k+1}^* \frac{3}{k(k+1)} - \frac{d_\nu^*}{\nu}$$

$$= d_1^* + d_2^* + d_3^* + d_4^* - 3\sum_{k=3}^{\nu-2} \frac{d_{k+1}^* - d_{k+2}^*}{k+1} - 4\frac{d_\nu^*}{\nu}$$

$$\leq d_1^* + d_2^* + d_3^* + d_4^* - 4\frac{d_\nu^*}{\nu}.$$

The first inequality follows because $k-2 \leq r_k \leq k$ for $k = 3, \ldots, \nu-1$ and $\{t-(k-2)\}^2/\{t(t+1)\}$ is increasing for $k-2 \leq t \leq k$ with a maximum at $t = k$. The second inequality follows because $d_1^* \geq d_2^* \geq \cdots \geq d_p^*$. Therefore, if $\nu \geq 4$ then

$$\frac{p}{p-2} M_\nu \geq \frac{p}{p-2}\left\{\sum_{j=1}^{p} d_j^* - \left(d_1^* + d_2^* + d_3^* + d_4^* - 4\frac{d_\nu^*}{\nu}\right)\right\}$$

$$= \sum_{j=3}^{p} d_j^* - \left(d_3^* + d_4^* - \frac{2}{p-2}\sum_{j=5}^{p} d_j^* - \frac{4p}{p-2}\frac{d_\nu^*}{\nu}\right).$$

If $\nu = 3$, then $L_\nu \leq d_1^* + d_2^* + d_3^* - d_3^*/3$ and hence

$$\frac{p}{p-2} M_\nu \geq \frac{p}{p-2}\left\{\sum_{j=1}^{p} d_j^* - \left(d_1^* + d_2^* + d_3^* - d_3^*/3\right)\right\}$$

$$= \sum_{j=3}^{p} d_j^* - \left(d_3^* - \frac{2}{p-2}\sum_{j=4}^{p} d_j^* - \frac{p}{p-2}\frac{d_3^*}{3}\right).$$

This completes the proof. □

Acknowledgements

The author thanks Bill Strawderman and Cunhui Zhang for helpful discussions.

Supplementary Material

Supplementary Material for "Improved minimax estimation of a multivariate normal mean under heteroscedasticity" (DOI: 10.3150/13-BEJ580SUPP; .pdf). We present additional results from the simulation study in Section 4.

References

[1] Baranchik, A.J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report 51, Dept. Statistics, Stanford Univ.

[2] Baranchik, A.J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.* **41** 642–645. MR0253461
[3] Berger, J.O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.* **4** 223–226. MR0397940
[4] Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8** 716–761. MR0572619
[5] Berger, J.O. (1982). Selecting a minimax estimator of a multivariate normal mean. *Ann. Statist.* **10** 81–92. MR0642720
[6] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR0804611
[7] Berger, J.O. and Robert, C. (1990). Subjective hierarchical Bayes estimation of a multivariate normal mean: On the frequentist interface. *Ann. Statist.* **18** 617–651. MR1056330
[8] Berger, J.O. and Srinivasan, C. (1978). Generalized Bayes estimators in multivariate problems. *Ann. Statist.* **6** 783–801. MR0478426
[9] Bhattacharya, P.K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. *Ann. Math. Statist.* **37** 1819–1824. MR0201026
[10] Bock, M.E. (1975). Minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* **3** 209–218. MR0381064
[11] Brandwein, A.C. and Strawderman, W.E. (1990). Stein estimation: The spherically symmetric case. *Statist. Sci.* **5** 356–369. MR1080957
[12] Brown, L.D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters). *J. Amer. Statist. Assoc.* **70** 417–427. MR0373082
[13] Brown, L.D. and Zhao, L.H. (2009). Estimators for Gaussian models having a block-wise structure. *Statist. Sinica* **19** 885–903. MR2536135
[14] Cai, T.T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statist. Sci.* **27** 31–50. MR2953494
[15] Casella, G. (1985). Condition numbers and minimax ridge regression estimators. *J. Amer. Statist. Assoc.* **80** 753–758. MR0803264
[16] Donoho, D.L., Liu, R.C. and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437. MR1062717
[17] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597
[18] Harville, D.A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.
[19] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Berkeley, CA: Univ. California Press. MR0133191
[20] Jones, M.C. (1986). Expressions for inverse moments of positive quadratic forms in normal variables. *Austral. J. Statist.* **28** 242–250. MR0860469
[21] Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. New York: Springer. MR1639875
[22] Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. With discussion. MR0696849
[23] Morris, C.N. and Lysy, M. (2012). Shrinkage estimation in multilevel normal models. *Statist. Sci.* **27** 115–134. MR2953499
[24] Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics* **2** 110–114.
[25] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Berkeley and Los Angeles: Univ. California Press. MR0084922

[26] Stein, C.M. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. Ser. B* **24** 265–296. MR0148184

[27] Stein, C.M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[28] Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388. MR0397939

[29] Strawderman, W.E. (2010). Bayesian decision based estimation and predictive inference. In *Frontiers of Statistical Decision Making and Bayesian Analysis*: *In Honor of James O. Berger* (M.-H. Chen et al., eds.) 69–82. New York: Springer.

[30] Tan, Z. (2014). Supplement to "Improved minimax estimation of a multivariate normal mean under heteroscedasticity". DOI:10.3150/13-BEJ580SUPP.

[31] Xie, X., Kou, S.C. and Brown, L.D. (2012). SURE estimates for a heteroscedastic hierarchical model. *J. Amer. Statist. Assoc.* **107** 1465–1479. MR3036408

Received May 2013 and revised November 2013