2011

Ensemble Minimax Estimation for Multivariate Normal Means

Lawrence D. Brown
University of Pennsylvania

Hui Nie
University of Pennsylvania

Xianchao Xie
Harvard University

Ensemble Minimax Estimation for Multivariate Normal Means

Abstract

This article discusses estimation of a heteroscedastic multivariate normal mean in terms of the ensemble risk. We first derive the ensemble minimaxity properties of various estimators that shrink towards zero. We then generalize our results to the case where the variances are given as a common unknown but estimable chi-squared random variable scaled by different known factors. We further provide a class of ensemble minimax estimators that shrink towards the common mean

Keywords

shrinkage, empirical bayes, ensemble minimax, James-Stein Estimation, Random Effects Models

Disciplines

Physical Sciences and Mathematics

Submitted to the Annals of Statistics
arXiv: math.PR/0000000

ENSEMBLE MINIMAX ESTIMATION FOR MULTIVARIATE NORMAL MEANS

By Lawrence D. Brown[*,†,‡], Hui Nie[‡] and Xianchao Xie[§]

University of Pennsylvania[‡] and Harvard University[§]

This article discusses estimation of a heteroscedastic multivariate normal mean in terms of the ensemble risk. We first derive the ensemble minimaxity properties of various estimators that shrink towards zero. We then generalize our results to the case where the variances are given as a common unknown but estimable chi-squared random variable scaled by different known factors. We further provide a class of ensemble minimax estimators that shrink towards the common mean.

1. Introduction. We consider the problem of simultaneously estimating the mean parameter $\theta = (\theta_1, \cdots, \theta_p)$ from independent normal observations $X \sim N(\theta, \Sigma)$, where $\Sigma = diag\{\sigma_1^2, \cdots, \sigma_p^2\}$. For any estimator $\hat{\theta}$, our loss function is the ordinary squared error loss

$$L(\hat{\theta}, \theta) = \sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2.$$

The conventional risk function is the expected value of the loss function with respect to θ. That is,

$$R(\theta, \hat{\theta}) = E_\theta(L(\hat{\theta}, \theta)) = \sum_{i=1}^{p} E_\theta(\hat{\theta}_i - \theta_i)^2.$$

James and Stein (1961) study the homoscedastic case in which $\sigma^2 = \sigma_1^2 = \cdots = \sigma_p^2$. In that case they prove the astonishing result that the James-Stein shrinkage estimator

$$\delta_{J-S}(X) = \left(1 - \frac{C\sigma^2}{\|X\|^2}\right) X \tag{1.1}$$

and its positive part

$$\delta_{J-S}^{+}(X) = \left(1 - \frac{C\sigma^2}{\|X\|^2}\right)_{+} X \tag{1.2}$$

[*]AMS 2000 subject classifications: 62C12, 62C20

[†]Supported in part by NSF grant DMS-07-07033

Keywords and phrases: Shrinkage, Empirical Bayes, Ensemble Minimax, James-Stein Estimation, Random Effects Models

dominate the MLE $\hat{\theta}_{mle} = X$ for $0 \leq C \leq 2(p-2)$ and $p \geq 3$. The discovery by James and Stein has led to a wide application of shrinkage techniques in many important problems. References include Efron and Morris (1975), Fay and Herriot (1979), Rubin (1981), Morris (1983a), Green and Strawderman (1985), Jones (1991) and Brown (2008). The theoretical properties of shrinkage estimators have also been extensively studied in the literature under the homoscedastic Gaussian model. Since Stein's discovery, "shrinkage" has been developed as a broad statistical framework in many aspects (Stein, 1962; Strawderman, 1971; Efron and Morris, 1971,1972a,1972b and 1973; Casella, 1980; Hastie et al., 2003).

There is also some literature discussing the properties of the James-Stein shrinkage estimators under heteroscedasticity. James and Stein (1961) discuss the estimation problem under heteroscedasticity where the loss function is weighted by the inverse of the variances. This problem can be transformed to the homoscedastic case under ordinary squared error loss. Brown (1975) shows that the James-Stein estimator is not always minimax and hence does not necessarily dominate the usual MLE under ordinary squared error loss when the variances are not equal. Specifically, the James-Stein shrinkage estimator does not dominate the usual MLE when the largest variance is larger than the sum of the rest. Moreover, Casella (1980) argues that the James-Stein shrinkage estimator may not be a desirable shrinkage estimator under heteroscedasticity even if it is minimax. Minimax estimators in general shrink most on the coordinates with smaller variances, while Bayes estimators shrink most on large variance coordinates.

In many applications, θ_i are thought to follow some exchangeable prior distribution π. It is then natural to consider the compound risk function which is then the Bayes risk with respect to the prior π

$$\overline{R}(\pi, \hat{\theta}) = E_\pi(R(\theta, \hat{\theta})) = \int R(\theta, \hat{\theta})\pi(d\theta).$$

Efron and Morris (1971, 1972a, 1972b and 1973) address this problem from both the Bayes and empirical Bayes perspective. They extensively develop this framework. Especially, they consider a prior distribution of the form $\theta \sim N_p(0, \tau^2 I)$ with $\tau \in [0, \infty)$, and they use the term "ensemble risk" for the compound risk. Morris and Lysy (2009) discuss the motivation and importance of shrinkage estimation in this multi-level normal model. The ensemble risk is described as the level-II risk in Morris and Lysy (2009).

By introducing a set of ensemble risks $\overline{R}(\pi, \hat{\theta})$ ($\pi \in \mathcal{P}$), we can then discuss ensemble minimaxity and other properties with respect to a set of prior distributions $\mathcal{P}$. We elaborate the definitions of ensemble minimaxity and other properties in Section 2. The previously cited papers (and others)

discuss the desirability of the ensemble risks with respect to the normal priors $\theta \sim N_p(0, \tau^2 I)$ with $\tau \in [0, \infty)$. In this paper, we will concentrate on the ensemble minimaxity of various estimators in this respect.

Brown (2008) discusses the connection between the parametric empirical Bayes estimator and the random effects model. In fact, the estimation problem of group means in a one way random effects model with infinite degrees of freedom for errors (and hence known error variance) is equivalent to the above problem. Our ensemble risk then corresponds to the ordinary risk function in the random effects model.

The more familiar unbalanced one-way random effects model is exactly equivalent to the generalization considered in Section 5. Again, ensemble risk in the empirical Bayes sense corresponds to ordinary prediction risk for the random effects model. We close Section 5 with a summary statement describing estimators proven to dominate the ordinary least squares group means in the random effect model.

Our article is organized as follows. In Section 2, we introduce necessary definitions and notations related to ensemble minimaxity. In Section 3, we discuss the ensemble minimaxity of various shrinkage estimators under heteroscedasticity. These include generalizations of the James-Stein estimator as well as versions of the empirical Bayes estimators proposed in Carter and Rolph (1974) and Brown (2008). In Section 4, We generalize our results to the case where the variances are given as a common unknown but estimable chi-squared random variable scaled by different known factors. In Section 5, we provide a class of ensemble minimax estimators that shrink towards the common mean.

2. Ensemble Minimaxity. As discussed above, we study in this paper the behavior of shrinkage estimators based on the ensemble risk

$$\overline{R}(\pi, \hat{\theta}) = E_\pi(R(\theta, \hat{\theta})) = \int R(\theta, \hat{\theta}) \pi(d\theta) \ .$$

If the prior $\pi(\theta)$ is known, the resulting posterior mean $E_\pi(\theta|x)$ is then the optimal estimate under the sum of the squared error loss. However, it is often infeasible to exactly specify the prior. To avoid excessive dependence on the choice of prior, it is natural to consider a set of priors $\mathcal{P}$ on θ and study the properties of various estimators based on the corresponding set of ensemble risks. As in the classic decision theory, there rarely exists an estimator that achieves the minimum ensemble risk uniformly for all $\pi \in \mathcal{P}$. A more realistic goal as pursued in this paper is to study the ensemble minimaxity (defined shortly) of familiar shrinkage estimators.

Recall that with ordinary risk $R(\theta, \delta)$, an estimator δ is said to dominate another estimator δ' if

$$R(\theta, \delta) \leq R(\theta, \delta^{'})$$

holds for each $\theta \in \Theta$ with strict inequality for at least one θ. The estimator δ is inadmissible if there exists another procedure which dominates δ; otherwise δ is admissible. δ is said to be minimax if

$$\sup_{\theta \in \Theta} R(\theta, \delta) = \inf_{\delta^{'}} \sup_{\theta \in \Theta} R(\theta, \delta^{'}) \ ,$$

that is, the estimator attains the minimum worst-case risk. Similarly for the case of ensemble risk we have the following definitions.

Ensemble admissibility and minimaxity. Given a set of priors $\mathcal{P}$, an estimator δ is said to dominate another estimator δ' with respect to $\mathcal{P}$ if

$$\overline{R}(\pi, \delta) \leq \overline{R}(\pi, \delta^{'})$$

holds for each $\pi \in \mathcal{P}$ with strict inequality for at least one π. The estimator δ is ensemble inadmissible with respect to $\mathcal{P}$ if there exists another procedure which dominates δ, otherwise δ is ensemble admissible. The estimator δ is ensemble minimax with respect to $\mathcal{P}$ if

$$\sup_{\pi \in \mathcal{P}} \overline{R}(\pi, \delta) = \inf_{\delta^{'}} \sup_{\pi \in \mathcal{P}} \overline{R}(\pi, \delta^{'}) \ .$$

The motivation for the above definitions comes from the use of the empirical Bayes method in simultaneous inference. Efron and Morris (1972a), building from Stein (1962), derive the James-Stein estimator through the parametric empirical Bayes model with $\theta_i \sim N(0, \tau^2)$. Note that in such an empirical Bayes model, τ^2 is the unknown parameter. (Parameter here refers to an unknown non-random quantity.) Ensemble admissibility and minimaxity with respect to $\mathcal{P} = \{\theta_i \sim N(0, \tau^2) : 0 < \tau^2 < \infty\}$ is then exactly the counterpart of ordinary admissibility and minimaxity in the empirical Bayes model. Consistent with this, we also confine $\mathcal{P}$ to be the one given above. Another reason for preferring such a set $\mathcal{P}$ is because it enjoys the conjugate minimaxity property (Morris, 1983a). From now on, mention of this underlying set $\mathcal{P}$ will be omitted whenever confusion is unlikely. As a explicit notation in this setting, we define $\overline{R}_{\tau^2}(\delta) = \overline{R}(\pi, \delta)$ for $\pi = N(0, \tau^2)$.

Note that ensemble minimaxity can also be interpreted as a particular case of Gamma minimaxity studied in the context of robust Bayes analysis (Good, 1952; Berger, 1979). However, in such studies, a "large" set consisting of many diffuse priors are usually included in the analysis. Since this is quite different from our formulation of the problem, we use the term ensemble minimaxity throughout our paper, following the Efron and Morris papers cited above.

3. Main Results on Ensemble Minimaxity. In this section, we discuss the ensemble minimaxity of various shrinkage estimators. We first present a general theorem characterizing a class of shrinkage estimators that are ensemble minimax. We then study the ensemble minimaxity of James-Stein-type shrinkage estimators, along with several supplementary theorems highlighting the difference and similarity between our results and those obtained in the homoscedastic case. Finally, we investigate the ensemble minimaxity of the parametric empirical Bayes estimator via method of moment estimation, a case with several open problems unresolved during our study. Throughout the current discussion, the variances σ_i^2 are assumed to be known; the case of unknown σ_i^2 is addressed in the next section.

3.1. *General Theory.* As discussed in Section 1, when $p \geq 3$ and $0 \leq C \leq 2(p-2)$, both δ_{J-S} in (1.1) and δ_{J-S}^+ in (1.2) are known to be minimax under the homoscedastic model. However, this is not always the case under the heteroscedastic model. Brown (1975) shows that for any $C > 0$, if $\sum \sigma_i^2 \leq 2\max\{\sigma_i^2\}$, both δ_{J-S} in (3.1) and δ_{J-S}^+ in (3.2) are no longer minimax in the ordinary sense. This is one motivation for instead studying the ensemble minimaxity for these shrinkage estimators. Before presenting our main result, we first give a lemma concerning the evaluation of the ensemble risk $\overline{R}_{\tau^2}(\delta)$ that will be repeatedly used in subsequent discussion.

LEMMA 3.1. *The ensemble risk of any estimator δ with the form $\delta_i(X) = (1 - h_i(X))X_i$ can be written as*

$$\overline{R}_{\tau^2}(\delta) = \sum_{i=1}^{p} \left[E_x \left(\frac{\sigma_i^2}{\tau^2 + \sigma_i^2} X_i - h_i(X) X_i \right)^2 + \frac{\tau^2 \sigma_i^2}{\tau^2 + \sigma_i^2} \right] ,$$

where the expectation is taken with respect to the joint distribution of X such that $X_i \sim N(0, \tau^2 + \sigma_i^2)$ and all the coordinates are jointly independent.

PROOF. By definition, we have

$$\overline{R}_{\tau^2}(\delta) = \int \int L(\theta, \delta(x)) P_{x|\theta}(dx) \pi_\tau(d\theta) = \sum_{i=1}^{p} E(X_i - \theta_i)^2 .$$

Note that

$$\begin{pmatrix} \theta_i \\ X_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2 \\ \tau^2 & \tau^2 + \sigma_i^2 \end{pmatrix} \right)$$

and (θ_i, X_i) are jointly independent for different i, we have from the property of conditional expectation

$$\begin{aligned}
\overline{R}_{\tau^2}(\delta) &= \sum_{i=1}^{p}[E_x(\delta_i(X) - E(\theta_i|X))^2 + E_x(E((\theta_i - E(\theta_i|X_i))^2|X))] \\
&= \sum_{i=1}^{p}\left[E_x\left((1-h_i(X))X_i - \frac{\tau^2}{\tau^2+\sigma_i^2}X_i\right)^2 + \frac{\tau^2\sigma_i^2}{\tau^2+\sigma_i^2}\right] \\
&= \sum_{i=1}^{p}\left[E_x\left(\frac{\sigma_i^2}{\tau^2+\sigma_i^2}X_i - h_i(X)X_i\right)^2 + \frac{\tau^2\sigma_i^2}{\tau^2+\sigma_i^2}\right]
\end{aligned}$$

where E_x is used to emphasize that the expectation is taken with respect to the marginal distribution of X, i.e., each coordinate X_i has the normal distribution $N(0, \tau^2 + \sigma_i^2)$ and they are jointly independent. □

Under the heteroscedastic model, we define the James-Stein-type estimator δ_{J-S} as

$$(\delta_{J-S}(X))_i = \left(1 - \frac{C\sigma_i^2}{\|X\|^2}\right) X_i \ . \tag{3.1}$$

Its positive part δ^+_{J-S} is then

$$(\delta^+_{J-S}(X))_i = \left(1 - \frac{C\sigma_i^2}{\|X\|^2}\right)_+ X_i \ . \tag{3.2}$$

Estimators of this general form appear as a generalization of the original James-Stein proposal in Brown (1966). See also Efron and Morris (1971). To study the ensemble minimaxity of these two estimators, we first present a general result that characterizes a class of shrinkage estimators that are ensemble minimax.

The general result refers to estimators with the form $\delta_i(X) = (1 - h_i(X))X_i$, as in Lemma 3.1, and in which h_i is symmetric in the sense that

$$h_i(X) = \mathfrak{h}_i(X_1^2, \cdots, X_p^2). \tag{3.3}$$

In addition, we define

$$W = \sum_{j=1}^{p} \frac{X_j^2}{\tau^2 + \sigma_j^2} \tag{3.4}$$

$$T_i = \frac{X_i^2}{W(\tau^2 + \sigma_i^2)}, \ i = 1, \cdots, p. \tag{3.5}$$

In this way $X_i^2 = (\tau^2 + \sigma_i^2)WT_i$, and $\mathfrak{h}$ can be rewritten as a function of $\underline{T} = (T_1, \cdots, T_p)$ and W. With a minor extension of the notation, write $\mathfrak{h}(\underline{T}, W) = \mathfrak{h}((\tau^2 + \sigma_1^2)WT_1, \cdots, (\tau^2 + \sigma_p^2)WT_p)$.

THEOREM 3.1. *An estimator δ with the form $\delta_i(X) = (1 - h_i(X))X_i$ is ensemble minimax if each shrinkage factor $h_i(X)$ satisfies the following conditions:*

(1) $h_i(X) \geq 0$, $\forall X$.
(2) $h_i(X)$ *can be written in the form (3.3).*
(3) $\mathfrak{h}_i(\underline{T}, W)$ *is decreasing in* W *for fixed* $\underline{T}$.
(4) $\mathfrak{h}_i(\underline{T}, W)W$ *is increasing in* W *for fixed* $\underline{T}$.
(5)

$$E\left[\sup_{\underline{T}} \mathfrak{h}_i(\underline{T}, W)\right] \leq \frac{2\sigma_i^2}{\sigma_i^2 + \tau^2} \ .$$

PROOF. From Lemma 3.1. and the fact that

$$\overline{R}_{\tau^2}(\delta_0) = \sum_{i=1}^{p} \sigma_i^2,$$

it suffices to show that for each i,

$$E\left(\frac{\sigma_i^2}{\sigma_i^2 + \tau^2} X_i - h_i(X)X_i\right)^2 \leq \frac{\sigma_i^4}{\tau^2 + \sigma_i^2} \ , \tag{3.6}$$

which is equivalent to

$$E\left[h_i(X)^2 X_i^2\right] \leq \frac{2\sigma_i^2}{\tau^2 + \sigma_i^2} E\left[h_i(X)X_i^2\right] \ .$$

To prove the above inequality, first note that condition (2) indicates

$$\begin{aligned} E\left[h_i(X)^2 X_i^2\right] &= E\left[\mathfrak{h}_i(\underline{T}, W)^2(\tau^2 + \sigma_i^2)T_i W\right] \\ &= E\left[E(\mathfrak{h}_i(\underline{T}, W)(\tau^2 + \sigma_i^2)T_i W \times \mathfrak{h}_i(\underline{T}, W)|\underline{T})\right] \ . \end{aligned}$$

From condition (3), (4) and the covariance inequality, we then have

$$E\left[h_i(X)^2 X_i^2\right] \leq E[E(\mathfrak{h}_i(\underline{T}, W)(\tau^2 + \sigma_i^2)T_i W|\underline{T}) \times E(\mathfrak{h}_i(\underline{T}, W)|\underline{T})] \ ,$$

which implies

$$E\left[h_i(X)^2 X_i^2\right] \leq E\left[E(\mathfrak{h}_i(\underline{T},W)(\tau^2+\sigma_i^2)T_iW|\underline{T}) \times E\left(\sup_{\underline{T}} \mathfrak{h}_i(\underline{T},W)|\underline{T}\right)\right] .$$

Based on the independence of $\underline{T}$ and W, we have

$$\begin{aligned} E\left[h_i(X)^2 X_i^2\right] &\leq E\left[E(\mathfrak{h}_i(\underline{T},W)(\tau^2+\sigma_i^2)T_iW|\underline{T}) \times E\left(\sup_{\underline{T}} \mathfrak{h}_i(\underline{T},W)\right)\right] \\ &= E\left(\sup_{\underline{T}} \mathfrak{h}_i(\underline{T},W)\right) \times E[\mathfrak{h}_i(\underline{T},W)T_iW] , \end{aligned}$$

which along from condition (5) shows

$$E\left[h_i(X)^2 X_i^2\right] \leq \frac{2\sigma_i^2}{\sigma_i^2+\tau^2} E(\mathfrak{h}_i(\underline{T},W)(\tau^2+\sigma_i^2)T_iW) = \frac{2\sigma_i^2}{\sigma_i^2+\tau^2} E(h_i(X)X_i^2) .$$

This proves the ensemble minimaxity of δ.

□

Note that most of the conditions in Theorem 3.1. are rather intuitive to understand. Condition (1) simply means that the estimator is indeed a genuine shrinkage estimator, and never an expander. Condition (2) implies the shrinkage estimator has a certain natural symmetry property. Condition (3) requires the amount of shrinkage to decrease when the distance of the data vector is further away from the origin. Condition (5) controls the expected overall amount of shrinkage according to the ratio of the variability of the observation and that of the prior, but this condition is less intuitive than the others.

Let $\mu \in \mathcal{R}^p$. Consider estimation of the linear combination $\mu^t\theta$ under squared error loss $L_{lc}(d,\theta) = (d-\mu^t\theta)^2$. Ordinary minimaxity and ensemble minimaxity can be defined for this loss. As a Corollary to Theorem 3.1, we have

COROLLARY 3.1. *Assume conditions (1)-(5) of Theorem 3.1. Then the estimator $\hat{\eta} = \mu^t\delta$ is an ensemble minimax estimator of $\eta = \mu^t\theta$.*

PROOF. From the proof of Theorem 3.1., we see that ensemble minimaxity is actually achieved for each coordinate, that is, for any $i = 1, \cdots, p$,

$$\overline{R}_{lc}(\delta_i, \theta_i) \leq \sigma_i^2 .$$

This proves validity of the corollary. □

3.2. *Ensemble Minimaxity of James-Stein-type Shrinkage Estimators.* With Theorem 3.1., we then proceed to study the ensemble minimaxity of certain shrinkage estimators. These estimators include the original James-Stein estimators. As we will show, the original James-Stein estimator is often-but not always-ensemble minimax. Consider the estimator δ_{GS} with the form

$$(\delta_{GS}(X))_i = \left(1 - \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \|X\|^2}\right) X_i \ , \tag{3.7}$$

where λ_i and ν_i are properly chosen constants. Consider also its positive part version δ_{GS}^+ given by

$$(\delta_{GS}^+(X))_i = \left(1 - \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \|X\|^2}\right)_+ X_i \ . \tag{3.8}$$

Note that these forms are generalizations of the original James-Stein forms, as can be seen by setting $\nu_i = 0$, $\lambda_i = C$.

The following two corollaries state conditions under which δ_{GS} in (3.7) and δ_{GS}^+ in (3.8) are ensemble minimax.

COROLLARY 3.2. *δ_{GS} in (3.7) is ensemble minimax if $p \geq 3$ and for any $i = 1, \cdots, p$, $0 \leq \lambda_i \leq 2(p-2)$ and $\nu_i \geq (\lambda_i/2 - (p-2) \cdot \sigma_{min}^2/\sigma_i^2)_+$ with $\sigma_{min}^2 = \min_i\{\sigma_i^2\}$.*

COROLLARY 3.3. *δ_{GS}^+ in (3.8) is ensemble minimax if $p \geq 3$ and for any $i = 1, \cdots, p$, $0 \leq \lambda_i \leq 2(p-2)$ and $\nu_i \geq [\lambda_i - (p-2)(1 + \sigma_{min}^2/\sigma_i^2)]_+$ with $\sigma_{min}^2 = \min_i\{\sigma_i^2\}$.*

Remarks: When $\nu_i = 0$ and $\lambda_i = C$, δ_{GS} in (3.7) and δ_{GS}^+ in (3.8) reduce to the James-Stein estimators in (3.1) and (3.2). In the case where σ_i^2 are all equal, Corollary 3.2 and 3.3 show that δ_{J-S} in (3.1) and δ_{J-S}^+ in (3.2) are each ensemble minimax when $0 \leq C \leq 2(p-2)$. This reaches the same conclusion as James and Stein (1961).

When the values of σ_i^2 are not all equal, the results in Corollary 3.2 and 3.3 do not always establish ensemble minimaxity of δ_{J-S} in (3.1) and δ_{J-S}^+ in (3.2) for the entire range $0 \leq C \leq 2(p-2)$. Specializing the conditions of Corollary 3.2 to the case where $\lambda_i = C$ and $\nu_i = 0$ yields that δ_{J-S} in (3.1) is ensemble minimax if

$$C \leq 2(p-2)\frac{\sigma_{min}^2}{\sigma_{max}^2}. \tag{3.9}$$

Thus, for any $C > 0$, there are configurations of $\sigma_1^2, \cdots, \sigma_p^2$ for which the conditions in Corollary 3.2 fail to prove δ_{J-S} in (3.1) is ensemble minimax.

For δ_{J-S}^+ in (3.2), the situation is a little different. Specializing the conditions of Corollary 3.3 to the case $\lambda_i = C$ and $\nu_i = 0$ yields that δ_{J-S}^+ in (3.2) is ensemble minimax if

$$C \leq (p-2)(1 + \frac{\sigma_{min}^2}{\sigma_{max}^2}). \tag{3.10}$$

Hence, when $C \leq p-2$, the conditions in Corollary 3.3 are always satisfied by δ_{J-S}^+ in (3.2). However, for any $C > p-2$, there are configurations of $\sigma_1^2, \cdots, \sigma_p^2$ for which Corollary 3.3 fails to prove δ_{J-S}^+ in (3.2) is minimax.

Theorem 3.2 and 3.3 below address ensemble minimaxity of δ_{J-S} in (3.1) when $C > 0$ and δ_{J-S}^+ in (3.2) when $C > p-2$. They state conditions under which δ_{J-S} in (3.1) and δ_{J-S}^+ in (3.2) can fail to be ensemble minimax when $C > 0$ or $C > p-2$, respectively. There is a gap between the conditions in Corollaries 3.2 and 3.3. We do not as yet have a formulation of a sharp necessary and sufficient condition for ensemble minimaxity of δ_{J-S} in (3.1) and δ_{J-S}^+ in (3.2) in the case of general $\sigma_1^2, \cdots, \sigma_p^2$.

PROOF. **Proof of Corollary 3.2**

It is sufficient for us to verify that the conditions in Theorem 3.1 are satisfied by

$$\begin{aligned} h_i(X) &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \|X\|^2} \\ &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + \sum_{j=1}^p (\sigma_i^2 + \tau^2) T_i W} \\ &= \mathfrak{h}_i(\underline{T}, W) \ . \end{aligned}$$

Clearly, the shrinkage factor $h_i(X)$ satisfies conditions (1)-(4). For (5), define $g_i(W)$ as

$$g_i(W) = \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + (\sigma_{min}^2 + \tau^2) W} \ .$$

Then, $\sup_{\underline{T}} h_i(\underline{T}, W) \leq g_i(W)$. Using the covariance inequality, we have

$$\begin{aligned} E[g_i(W)] &= E\left[\frac{\lambda_i \sigma_i^2 / W}{\nu_i \sigma_i^2 / W + \sigma_{min}^2 + \tau^2}\right] \leq \frac{E[\lambda_i \sigma_i^2 / W]}{E[\nu_i \sigma_i^2 / W + \sigma_{min}^2 + \tau^2]} \\ &= \frac{\lambda_i \sigma_i^2}{\nu_i \sigma_i^2 + (p-2)(\sigma_{min}^2 + \tau^2)} \ . \end{aligned}$$

From the condition $0 \le \lambda_i \le 2(p-2)$ and $\nu_i \ge (\lambda_i/2 - (p-2)\cdot\sigma^2_{min}/\sigma^2_i)_+$, it is then easy to verify

$$\frac{\lambda_i\sigma_i^2}{\nu_i\sigma_i^2 + (p-2)(\sigma^2_{min}+\tau^2)} \le \frac{2\sigma_i^2}{\sigma_i^2+\tau^2} \ ,$$

which completes the proof. □

PROOF. **Proof of Corollary 3.3**

As in the proof of Corollary 3.2., it is sufficient for us to verify that conditions in Theorem 3.1 are satisfied by $h_i^+(X) = h_i(X) \wedge 1$, where

$$\begin{aligned} h_i(X) &= \frac{\lambda_i\sigma_i^2}{\nu_i\sigma_i^2 + \|X\|^2} \\ &= \frac{\lambda_i\sigma_i^2}{\nu_i\sigma_i^2 + \sum_{j=1}^p(\sigma_i^2+\tau^2)T_iW} \\ &= \mathfrak{h}_i(\underline{T}, W) \ . \end{aligned}$$

Conditions (1)-(4) are straightforward. If $\tau^2 \le \sigma_i^2$, (5) is also automatically satisfied. Assuming $\tau^2 > \sigma_i^2$, define $g_i(W)$ as

$$g_i(W) = \frac{\lambda_i\sigma_i^2}{\nu_i\sigma_i^2 + (\sigma^2_{min}+\tau^2)W} \ .$$

Note that $\sup_{\underline{T}} \mathfrak{h}_i(\underline{T}, W) \le g_i(W)$. Using the covariance inequality we have

$$\begin{aligned} E[g_i(W)] &= E\left[\frac{\lambda_i\sigma_i^2/W}{\nu_i\sigma_i^2/W + \sigma^2_{min}+\tau^2}\right] \le \frac{E[\lambda_i\sigma_i^2/W]}{E[\nu_i\sigma_i^2/W + \sigma^2_{min}+\tau^2]} \\ &= \frac{\lambda_i\sigma_i^2}{\nu_i\sigma_i^2 + (p-2)(\sigma^2_{min}+\tau^2)} \ . \end{aligned}$$

Now we only need to show

$$\frac{\lambda_i\sigma_i^2}{\nu_i\sigma_i^2 + (p-2)(\sigma^2_{min}+\tau^2)} \le \frac{2\sigma_i^2}{\sigma_i^2+\tau^2}$$

for $\tau^2 > \sigma_i^2$, which is equivalent to

$$2\sigma_i^2((p-2)-\nu_i) \le (\sigma_i^2+\tau^2)(2(p-2)-\lambda_i) + 2(p-2)\sigma^2_{min} \ .$$

Since $0 \le \lambda_i \le 2(p-2)$ and $\nu_i \ge [\lambda_i - (p-2)(1+\sigma^2_{min}/\sigma^2)]_+$, we have

$$\begin{aligned} 2\sigma_i^2((p-2)-\nu_i) &\le 2\sigma_i^2(2(p-2)-\lambda_i) + 2(p-2)\sigma^2_{min} \\ &\le (\sigma_i^2+\tau^2)(2(p-2)-\lambda_i) + 2(p-2)\sigma^2_{min} \ , \end{aligned}$$

which completes the proof. □

THEOREM 3.2. *For any $C = c(p-2)$ with $c > 1$, there exists some sufficiently large $p \geq 3$ and some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ_{J-S}^+ in (3.2) is not ensemble minimax.*

PROOF. Fix $\sigma_1^2 > 0$ and set $\sigma_2^2 = \cdots = \sigma_p^2 = \sigma^2$. Let $\sigma^2 \to 0$. In order to show that δ_{J-S}^+ dominates δ_0 with respect to $\mathcal{P}$, we have to prove

$$E\left[h_1(X)^2 X_1^2\right] \leq \frac{2\sigma_1^2}{\tau^2 + \sigma_1^2} E\left[h_1(X) X_1^2\right] \tag{3.11}$$

with

$$h_1(X) = \frac{c(p-2)\sigma_1^2}{\|X\|^2} \wedge 1 \ .$$

However, the law of large numbers implies that

$$\frac{c(p-2)\sigma_1^2}{\|X\|^2} = \frac{c(p-2)\sigma_1^2}{p} \frac{p}{\|X\|^2} \to \frac{c\sigma_1^2}{\tau^2}$$

as $p \to \infty$. Let $\sigma_1^2 < \tau^2 < c\sigma_1^2$, we then have

$$h_1(X) \to 1$$

as $p \to \infty$. Since $0 < h_1(X) \leq 1$, the dominated convergence theorem implies

$$E(h_1(X)^2 X_1^2) \to \tau^2 + \sigma_1^2$$

and

$$E(h_1(X) X_1^2) \to \tau^2 + \sigma_1^2 \ .$$

However, our choice of σ_1^2 and τ^2 indicates

$$1 > \frac{2\sigma_1^2}{\tau^2 + \sigma_1^2},$$

which means that the inequality (3.11) does not hold for large p. Hence, δ_{J-S}^+ in (3.2) is not ensemble minimax for some p and $\sigma_1^2, \cdots, \sigma_p^2$. □

The above results show that for δ_{J-S}^+ in (3.2) to be ensemble minimax, the constant C must have a much smaller upper bound under the general heteroscedastic model than under the homoscedastic one. Furthermore, it is proved below that δ_{J-S} in (3.1) is not even always ensemble minimax regardless of the choice of C.

THEOREM 3.3. *For any $C > 0$, there exists some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ_{J-S} in (3.1) is not ensemble minimax.*

PROOF. Fix $\sigma_1^2 = 1$, and let $\sigma_2^2 = \cdots = \sigma_p^2 = \sigma^2$. It suffices to show

$$\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \overline{R}_{\tau^2}(\delta_{J-S}) = \infty \ .$$

From Lemma 3.1., we have

$$\begin{aligned}
\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \overline{R}_{\tau^2}(\delta_{J-S}) &\geq \lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} E\left(\frac{1}{\tau^2+1}X_1 - \frac{C}{\|X\|^2}X_1\right)^2 \\
&\geq \lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \frac{1}{2}E\left(\frac{C}{\|X\|^2}X_1\right)^2 - \lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} E\left(\frac{1}{\tau^2+1}X_1\right)^2 \\
&\geq \lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \frac{1}{2}E\left(\frac{C^2X_1^2}{\|X\|^4}\right) - 1 \ .
\end{aligned}$$

Since the last term is finite, it is sufficient to prove

$$\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} E\left(\frac{X_1^2}{\|X\|^4}\right) = \infty \ .$$

Let $X_1 = \sqrt{1+\tau^2}Z_1$, $Z_1 \sim N(0,1)$, and $X_i = \sqrt{\tau^2+\sigma^2}Z_i$, $Z_i \sim N(0,1)$, $\forall i = 2, \cdots, p$. Therefore,

$$E\left(\frac{X_1^2}{\|X\|^4}\right) = (1+\tau^2)E\left(\frac{Z_1^2}{((1+\tau^2)Z_1^2 + (\tau^2+\sigma^2)\sum_{i=2}^p Z_i^2)^2}\right) \ .$$

Note that $\frac{Z_1^2}{((1+\tau^2)Z_1^2+(\tau^2+\sigma^2)\sum_{i=2}^p Z_i^2)^2}$ is increasing when both σ^2 and τ^2 decrease to zero and

$$\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} \frac{X_1^2}{\|X\|^4} = \frac{1}{Z_1^2}.$$

From the monotone convergence theorem, we have

$$\lim_{\substack{\tau^2 \to 0 \\ \sigma^2 \to 0}} E\left(\frac{X_1^2}{\|X\|^4}\right) = E\left(\frac{1}{Z_1^2}\right) = \infty$$

which completes the proof. □

The fact that δ_{J-S} in (3.1) is not in general ensemble minimax may be quite surprising at first glance. However, this is to be expected given the form of δ_{J-S}. Under the heteroscedastic model, δ_{J-S} may have a non-negligible probability of dramatically over-shrinking, which causes the performance of the shrinkage estimator to deteriorate; while such an issue is always well controlled in the homoscedastic case.

Similar to the homoscedastic case, δ^+_{J-S} in (3.2) is always better than δ_{J-S} in (3.1) in terms of ensemble risk under the heteroscedastic model. A general result is given in the following theorem.

THEOREM 3.4. *Let δ be any estimator with the form $\delta_i(X) = (1 - h_i(X))X_i$ and $\delta^+(X)$ be its positive part estimator such that $\delta_i^+(X) = (1 - h_i(X))_+X_i$. If $P(\delta \neq \delta^+) > 0$, then δ^+ dominates δ with respect to $\mathcal{P}$.*

PROOF. First note that δ_i^+ is equivalently written as $\delta_i^+(X) = (1 - h_i^+(X))X$, where $h_i^+(X) = h_i(x) \wedge 1$. From Lemma 3.1., we have

$$\begin{aligned} & \overline{R}_{\tau^2}(\delta^+) - \overline{R}_{\tau^2}(\delta) \\ = & \sum_{i=1}^{p} E\left[\left(h_i^+(X) + h_i(X) - \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}\right)(h_i(X) - h_i^+(X))X_i^2\right] . \end{aligned}$$

Since for any $i = 1, \cdots, p$,

$$h_i(X) - h_i^+(X) = \begin{cases} h_i(X) - 1 & , \text{if } h_i(X) > 1 \\ 0 & , \text{if } h_i(X) \leq 1 \end{cases} ,$$

we then have

$$\begin{aligned} & \overline{R}_{\tau^2}(\delta^+) - \overline{R}_{\tau^2}(\delta) \\ = & \sum_{i=1}^{p} E\left[\left(h_i^+(X) + h_i(X) - \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}\right)(h_i(X) - h_i^+(X))X_i^2\right] \\ = & \sum_{i=1}^{p} E\left[\left(1 + h_i(X) - \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}\right)(h_i(X) - 1)X_i^2 I_{\{h_i(X)>1\}}\right] \\ \geq & \sum_{i=1}^{p} E\left[\left(2 - \frac{2\sigma_i^2}{\tau^2+\sigma_i^2}\right)(h_i(X) - 1)X_i^2 I_{\{h_i(X)>1\}}\right] > 0 \end{aligned}$$

which completes the proof. □

COROLLARY 3.4. *For any constant $C \geq 0$, δ^+_{J-S} in (3.2) dominates δ_{J-S} in (3.1) with respect to $\mathcal{P}$.*

PROOF. Directly from Theorem 3.4.. □

3.3. *Parametric empirical Bayes estimator (via Method of Moments).* Carter and Rolph (1974), Brown (2008) and Efron and Morris (1973, 1975) each derive parametric empirical Bayes estimators for the heteroscedastic problem. The first two papers use method of moments to estimate the hyperparameter τ^2. (Morris and Lysy (2009) also discuss such estimators.) We will discuss here ensemble minimaxity of such empirical Bayes estimators.

In contrast, Efron and Morris (1973, 1975) use a maximum likelihood method for this step. The resulting estimation does not have an explicit closed form, although it is easily calculated numerically. For this reason we have (so far) been less successful in settling the ensemble minimaxity of this empirical Bayes version, and we do not address this issue here.

In this subsection, we treat the special case of shrinkage to 0. The previously cited references (and others) involve shrinkage to a common mean. This generalization is treated in Section 5. While our results in the present subsection shed some light on the ensemble minimaxity of these estimators, they are unfortunately not as nearly complete as are our preceding results about generalized James-Stein estimators.

As mentioned above, if τ^2 is known, the optimal estimator of θ_i ($i = 1, \cdots, p$) would be

$$(\delta_B(X))_i = \left(1 - \frac{\sigma_i^2}{\tau^2 + \sigma_i^2}\right) X_i \ , \tag{3.12}$$

which is the Bayes estimator. However in the empirical Bayes setting, τ^2 is an unknown hyper-parameter to be estimated. The idea of the parametric empirical Bayes method is to use $\{X_i\}$ to obtain an estimate of τ^2 and then substitute the estimate of τ^2 into (3.12) to yield a final estimator of $\{\theta_i\}$. Below we use the method of moments estimator

$$\widetilde{\tau}^2 = \frac{1}{p}\sum_{i=1}^{p}(X_i^2 - \sigma_i^2),$$

and its positive part

$$\widetilde{\tau}_+^2 = \frac{1}{p}\left[\sum_{i=1}^{p}(X_i^2 - \sigma_i^2)\right]_+ .$$

In practice, some other constant "$1/C$" is oftens used in lieu of "$1/p$" above.

The corresponding parametric empirical Bayes estimator is then given by

$$(\delta_{PEB}(X))_i = \left(1 - \frac{\sigma_i^2}{\sigma_i^2 + \frac{1}{C}\sum_{j=1}^{p}(X_j^2 - \sigma_j^2)}\right) X_i \ , \tag{3.13}$$

along with its positive part estimator

$$(\delta^+_{PEB}(X))_i = \left(1 - \frac{\sigma_i^2}{\sigma_i^2 + \frac{1}{C}(\sum_{j=1}^p (X_j^2 - \sigma_j^2))_+}\right) X_i \ . \tag{3.14}$$

Note that the form of the parametric empirical Bayes estimator δ_{PEB} in (3.13) differs from the James-Stein-type estimator δ_{J-S} in (3.1) in the use of the term $C\sigma_i^2 + \sum_{j=1}^p (X_j^2 - \sigma_j^2)$ instead of $\sum_{j=1}^p X_j^2$ in the denominator. Therefore the former denominator can be much smaller than the latter and hence lead to over-shrinkage. Not surprisingly, the conditions needed for ensemble minimaxity appear somewhat more restrictive than in the James-Stein case.

The following corollary contains conditions that guarantee ensemble minimaxity for the parametric empirical Bayes estimators. Simulation results (not reported here) lead us to conjecture that ensemble minimaxity holds under somewhat less restrictive conditions.

COROLLARY 3.5. *Assume*

$$p \le \frac{\sum_{j=1}^p \sigma_j^2}{\sigma_{min}^2} \le C \le 2(p-2) \ . \tag{3.15}$$

Then both δ_{PEB} in (3.13) and δ^+_{PEB} in (3.14) are ensemble minimax.

Remark: In the homoscedastic case, Condition (3.15) requires $p \le 2(p-2)$. This is satisfied if and only if $p \ge 4$. In that case, δ_{PEB} in (3.13) and δ^+_{PEB} in (3.14) are ensemble minimax if $4 \le p \le C \le 2(p-2)$.

PROOF. Set $\lambda_i = C$ and $\nu_i = C - \frac{\sum_{j=1}^p \sigma_j^2}{\sigma_i^2}$. Condition (3.15) guarantees that $\nu_i \ge 0$ and $\lambda_i \le 2(p-2)$. It is evident that $\delta_{PEB} = \delta_{GS}$. A little care with the positive part conditions shows that also $\delta^+_{PEB} = \delta^+_{GS}$.

It then follows from Corollary 3.2 that $\delta_{PEB} = \delta_{GS}$ is ensemble minimax if

$$\text{diff} = \nu_i - [\frac{C}{2} - (p-2)\frac{\sigma_{min}^2}{\sigma_i^2}] \ge 0. \tag{3.16}$$

Substituting and simplifying yields

$$\begin{aligned} \text{diff} &= \frac{C}{2} - [\frac{\sum_{j=1}^p \sigma_j^2}{\sigma_i^2} - (p-2)\frac{\sigma_{min}^2}{\sigma_i^2}] \\ &\ge \frac{C}{2} - \frac{\sum_{j=1}^p \sigma_j^2 - (p-2)\sigma_{min}^2}{\sigma_{min}^2}, \end{aligned}$$

since $\sum_{j=1}^p \sigma_j^2 \geq p\sigma_{min}^2 \geq (p-2)\sigma_{min}^2$. Hence, from (3.15),

$$\begin{aligned} \text{diff} &\geq \frac{C}{2} + p - 2 - \frac{\sum_{j=1}^p \sigma_j^2}{\sigma_{min}^2} \\ &\geq \frac{1}{2}[C - \frac{\sum_{j=1}^p \sigma_j^2}{\sigma_{min}^2}] \\ &\geq 0 \ . \end{aligned}$$

This verifies (3.16) and proves δ_{PEB} is ensemble minimax.

The proof for δ_{PEB}^+ is similar, but easier. λ_i and ν_i are defined as before. Condition (3.15) is still required in order that $0 \leq \lambda_i \leq 2(p-2)$ and $\nu_i \geq 0$. Truth of (3.16) validates the remaining condition in Corollary 3.3, and hence proves δ_{PEB}^+ is ensemble minimax. □

THEOREM 3.5. *Let $p \geq 1$ and $C > 0$. Then there exists some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ_{PEB} in (3.13) is not ensemble minimax.*

PROOF. When $p = 1$, set $\tau^2 = 1$. From Lemma 3.1., we only need to show

$$\lim_{\sigma_1^2 \to 0} E\left(\frac{1}{1+\sigma_1^2}X_1 - \frac{C}{X_1^2 + (C-1)\sigma_1^2}X_1\right)^2 > \frac{1}{1+\sigma_1^2} \ .$$

Since

$$\begin{aligned} & E\left(\frac{1}{1+\sigma_1^2}X_1 - \frac{C}{X_1^2 + (C-1)\sigma_1^2}X_1\right)^2 \\ \geq \ & \frac{1}{2}E\left(\frac{C}{X_1^2 + (C-1)\sigma_1^2}X_1\right)^2 - E\left(\frac{1}{1+\sigma_1^2}X_1\right)^2 \\ = \ & \frac{1}{2}E\left(\frac{C^2X_1^2}{(X_1^2 + (C-1)\sigma_1^2)^2}\right) - \frac{1}{1+\sigma_1^2} \end{aligned}$$

where the last term is finite, it is then sufficient to show

$$\lim_{\sigma_1^2 \to 0} E\left(\frac{X_1^2}{(X_1^2 + (C-1)\sigma_1^2)^2}\right) = \infty \ . \tag{3.17}$$

When $C < 1$, this is trivial. In fact, it holds for any σ_1^2. When $C \geq 1$, let $X_1 = \sqrt{1+\sigma_1^2}Z_1$, $Z_1 \sim N(0,1)$, we have

$$E\left(\frac{X_1^2}{(X_1^2 + (C-1)\sigma_1^2)^2}\right) = E\left(\frac{(1+\sigma_1^2)Z_1^2}{((1+\sigma_1^2)Z_1^2 + (C-1)\sigma_1^2)^2}\right) \ .$$

Since $\frac{Z_1^2}{((1+\sigma_1^2)Z_1^2+(C-1)\sigma_1^2)^2}$ is increasing as $\sigma_1^2 \to 0$, and

$$\lim_{\sigma_1^2 \to 0} \frac{Z_1^2}{((1+\sigma_1^2)Z_1^2+(C-1)\sigma_1^2)^2} = \frac{1}{Z_1^2} ,$$

from monotone convergence theorem, we have

$$\lim_{\sigma_1^2 \to 0} E\left(\frac{Z_1^2}{((1+\sigma_1^2)Z_1^2+(C-1)\sigma_1^2)^2}\right) = E\left(\frac{1}{Z_1^2}\right) = \infty .$$

Note that $1+\sigma_1^2 \to 1$ as $\sigma_1^2 \to 0$, (3.17) is then verified. Hence, when $p=1$, δ_{PEB} in (3.13) is not ensemble minimax.

For the case where $p \geq 2$, let $\sigma_1^2 = 1$ and $\sigma_2^2 = \cdots = \sigma_p^2 = C$. Again from Lemma 3.1. we have

$$\begin{aligned}
\overline{R}_{\tau^2}(\delta_{PEB}) &\geq E\left(\frac{1}{\tau^2+1}x_1 - \frac{C}{C+\|X\|^2-1-(p-1)C}X_1\right)^2 \\
&\geq \frac{1}{2}E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2}\right) - E\left(\frac{1}{(\tau^2+1)^2}X_1^2\right) \\
&= \frac{1}{2}E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2}\right) - \frac{1}{\tau^2+1} \\
&= \frac{1}{2}E\left[E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2} \mid X_1\right)\right] - \frac{1}{\tau^2+1} .
\end{aligned}$$

For any $X_1^2 < 1+(p-2)C$, it is not difficult to see that

$$E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2} \mid X_1\right) = \infty$$

and

$$P(X_1^2 < 1+(p-2)C) > 0 ,$$

we then have

$$E\left(\frac{C^2X_1^2}{(\|X\|^2-1-(p-2)C)^2}\right) = \infty ,$$

which implies

$$\overline{R}_{\tau^2}(\delta_{PEB}) = \infty . \tag{3.18}$$

Therefore, δ_{PEB} in (3.13) is not ensemble minimax. To sum up, there exists some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ_{PEB} in (3.13) is not ensemble minimax. □

Unfortunately, we have been unable to obtain a complete answer on the ensemble minimaxity of the positive part estimators δ^+_{PEB} in (3.14). Nevertheless, the following theorem indicates that unlike in the case of James-Stein-type estimators C can not be too small.

THEOREM 3.6. *For any p, there exists some sufficiently small C and some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ^+_{PEB} in (3.14) is not ensemble minimax.*

PROOF. Let $\sigma_1^2 = \cdots = \sigma_p^2 = 1$ and $\tau^2 = 2$. Similarly as above, to show that δ^+_{PEB} in (3.14) is ensemble minimax, we would need to have

$$E\left(\sum_{i=1}^{p} h_i^2(X) X_i^2\right) \leq \frac{2}{3} E\left(\sum_{i=1}^{p} h_i(X) X_i^2\right) \tag{3.19}$$

with

$$h_i(X) = \frac{1}{1 + \frac{1}{C}(\|X\|^2 - p)_+} .$$

Notice that $h_i(X) \to I_{\{\|X\|^2 \leq p\}}$ as $C \to 0$ and $h_i(X) \leq 1$, from dominant convergence theorem, we have

$$E\left(\sum_{i=1}^{p} h_i^2(X) X_i^2\right) \to E\left[\|X\|^2 I_{\{\|X\|^2 \leq p\}}\right]$$

and

$$E\left(\sum_{i=1}^{p} h_i(X) X_i^2\right) \to E\left[\|X\|^2 I_{\{\|X\|^2 \leq p\}}\right]$$

as $C \to 0$. Hence, as $C \to 0$, (3.19) would no longer always hold. Thus, there exists some sufficiently small C and some $\sigma_1^2, \cdots, \sigma_p^2$ such that δ^+_{PEB} in (3.14) is not ensemble minimax. □

One interesting observation here is that as $C \to 0$, δ^+_{PEB} reduces to the hard-threshold estimator $X 1_{\{\|X\|^2 > p\sigma^2\}}$ under the homoscedastic model. The above theorem simply indicates that the hard-threshold estimator is worse than the ordinary MLE in terms of ensemble risk when $\tau^2 > \sigma^2$.

Similar to the James-Stein estimator, δ^+_{PEB} in (3.14) is better than δ_{PEB} in (3.13) in terms of ensemble risk.

COROLLARY 3.6. *For any constant $C \geq 0$, δ^+_{PEB} in (3.14) dominates δ_{PEB} in (3.13) with respect to $\mathcal{P}$.*

PROOF. Directly from Theorem 3.4.. □

Although simulation results lend support to the conjecture that δ^+_{PEB} in (3.14) is ensemble minimax when $1 \le C \le 2(p-2)$ (the lower bound could be much smaller) and $p \ge 3$, a rigorous proof is still yet to be found.

4. Generalization to the Unknown Variances Case. The discussion so far has been focused on the ensemble minimaxity of shrinkage estimators assuming the variances σ_i^2 to be known. It is common in many circumstances that variances are unknown and have to be estimated from data. Here we consider the case where $X_i \sim N(\theta_i, \sigma^2\gamma_i)$ for $i = 1, \cdots, p$ with unknown σ^2 but known γ_i. Denote $\Gamma = diag\{\gamma_1, \cdots, \gamma_p\}$. We also assume that σ^2 is estimated by $M \sim \sigma^2\chi^2_m/m$ where M is independent of X, an assumption which is satisfied in applications in which a pooled estimate of σ^2 is used. In particular, this setting corresponds to the one-way random effects setting of Section 5.2 with $\gamma_i = 1/J_i$ where J_i is the number of observations in group i. We will discuss the ensemble minimaxity of some shrinkage estimators. First of all, we give two lemmas that will be used in our later proof. The first one is the generalization of Lemma 3.1. to the unknown variances case.

LEMMA 4.1. *The ensemble risk of any estimator δ with the form $\delta_i(X, M) = (1 - h_i(X, M))X_i$ has the following representation*

$$\overline{R}_{\tau^2}(\delta) = \sum_{i=1}^{p} E\left[\left(\frac{\sigma^2\gamma_i}{\tau^2 + \sigma^2\gamma_i}X_i - h_i(X, M)X_i\right)^2 + \frac{\tau^2\sigma^2\gamma_i}{\tau^2 + \sigma^2\gamma_i}\right], \tag{4.1}$$

where the expectation is taken with respect to the joint distribution of (X, M) where each $X_i \sim N(0, \tau^2 + \sigma^2\gamma_i)$ and $M \sim \sigma^2\chi^2_m/m$, and they are jointly independent.

PROOF. The proof follows the same approach as in Lemma 3.1. once we condition on M. First note that

$$\begin{aligned}\overline{R}_{\tau^2}(\delta) &= E\left[\sum_{i=1}^{p}(\delta_i(X, M) - \theta_i)^2\right] \\ &= \sum_{i=1}^{p} E[E[(\delta_i(X, M) - \theta_i)^2|M]]\end{aligned}$$

Since given M, (θ_i, X_i) is an independent array whose distribution is

$$\begin{pmatrix} \theta_i \\ X_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2 \\ \tau^2 & \tau^2 + \sigma^2\gamma_i \end{pmatrix}\right).$$

Conditioning on M, we have, as in Lemma 3.1.,

$$\begin{aligned} & E[(\delta_i(X,M)-\theta_i)^2|M] \\ = \; & E\left[\left(\frac{\sigma^2\gamma_i}{\tau^2+\sigma^2\gamma_i}X_i - h_i(X,M)X_i\right)^2|M\right] + \frac{\tau^2\sigma^2\gamma_i}{\tau^2+\sigma^2\gamma_i} \ , \end{aligned}$$

which then implies

$$\begin{aligned} \overline{R}_{\tau^2}(\delta) &= \sum_{i=1}^{p} E[E[(\delta_i(X,M)-\theta_i)^2]|M] \\ &= E\left[\sum_{i=1}^{p}\left(\frac{\sigma^2\gamma_i}{\tau^2+\sigma^2\gamma_i}X_i - h_i(X,M)X_i\right)^2 + \frac{\tau^2\sigma^2\gamma_i}{\tau^2+\sigma^2\gamma_i}\right] \ . \end{aligned}$$

□

The second Lemma is an inequality concerning expectations of non-negative random variables.

LEMMA 4.2. *For a non-negative random variable M and two non-negative functions $\mu(M)$ and $\mu'(M)$, if the ratio $r(M)=\mu(M)/\mu'(M)$ is non-decreasing in M, we then have*

$$\frac{E(M\mu(M))}{E(\mu(M))} \geq \frac{E(M\mu'(M))}{E(\mu'(M))}$$

assuming all expectations are finite and non-zero.

PROOF. First we use $\mu'(M)$ to induce a new probability distribution

$$P_{\mu'}(M\in A) = \int_A \frac{\mu'(m)}{E(\mu'(M))}dm \ .$$

Using this change of measure, we have

$$\frac{E[M\mu(M)]}{E[\mu(M)]} = E_{\mu'}[M\cdot r(M)] \times \frac{E[\mu'(M)]}{E[\mu(M)]}$$

and

$$\frac{E(M\mu'(M))}{E(\mu'(M))} = E_{\mu'}(M) \ ,$$

the original inequality then becomes a direct application of covariance inequality under the new probability $P_{\mu'}$. □

Define δ_{GSV} with the form

$$(\delta_{GSV}(X))_i = \left(1 - \frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right) X_i \ . \tag{4.2}$$

We have the following theorem characterizing the ensemble minimaxity of $\delta_{GSV}(X)$ in (4.2). The upper bound for λ_i is slightly smaller since we are now estimating σ_i^2, a phenomenon observed in similar studies under the homoscedastic model.

THEOREM 4.1. *δ_{GSV} in (4.2) is ensemble minimax if $p \geq 3$, $m \geq 3$ and for any $i = 1, \cdots, p$, $0 \leq \lambda_i \leq \frac{2m(p-2)}{m+2}$ and $\nu_i \geq (\frac{m+2}{2(m-2)}\lambda_i - \frac{m\gamma_{min}(p-2)}{\gamma_i(m-2)})_+$.*

PROOF. As in the proof for the known variance case, based on Lemma 4.1., it suffices to show

$$E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)^2 X_i^2\right] \leq \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i + \tau^2} E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2} X_i^2\right) \ .$$

Conditioning on M and following the proof in Theorem 3.1. and Corollary 3.2., we know

$$\begin{aligned} & E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)^2 X_i^2\right] \\ \leq \ & E\left[E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2} X_i^2 | M\right) \times E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W} | M\right)\right] \ . \end{aligned}$$

The difficulty here is that a direct application of the covariance inequality on the two conditional expectation is no longer helpful since they are both increasing in M. However, by moving the M in the numerator of the second conditional expectation to the first one, the covariance inequality can then be applied, i.e.,

$$\begin{aligned} & E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)^2 X_i^2\right] \\ \leq \ & E\left[E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2} X_i^2 | M\right) \times E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W} | M\right)\right] \\ = \ & E\left[E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2} X_i^2 | M\right) \times E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W} | M\right)\right] \\ = \ & E\left[E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2} X_i^2 | M\right)\right] \times E\left[E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W} | M\right)\right] \\ = \ & E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2} X_i^2\right) \times E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}\right) \ . \end{aligned}$$

Now let

$$\mu_s(M) = \frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + s} \ ,$$

notice that the ratio $r(M) = \mu_s(M)/\mu_{s'}(M)$ is non-decreasing in M for $s > s'$, from Lemma 4.2. we then have

$$\frac{E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i+\|X\|^2}|X\right)}{E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i+\|X\|^2}|X\right)} \leq \lim_{\|X\|^2\to\infty} \frac{E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i+\|X\|^2}|X\right)}{E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i+\|X\|^2}|X\right)} = \lim_{\|X\|^2\to\infty} \frac{E\left(\frac{\lambda_i M^2\gamma_i\cdot\|X\|^2}{\nu_i M\gamma_i+\|X\|^2}|X\right)}{E\left(\frac{\lambda_i M\gamma_i\cdot\|X\|^2}{\nu_i M\gamma_i+\|X\|^2}|X\right)} \ .$$

Applying monotone convergence theorem gives us

$$\lim_{\|X\|^2\to\infty} \frac{E\left(\frac{\lambda_i M^2\gamma_i\cdot\|X\|^2}{\nu_i M\gamma_i+\|X\|^2}|X\right)}{E\left(\frac{\lambda_i M\gamma_i\cdot\|X\|^2}{\nu_i M\gamma_i+\|X\|^2}|X\right)} = \frac{E(M^2)}{E(M)} = \frac{(m+2)\sigma^2}{m} \ ,$$

which along with the previous inequality implies

$$E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}|X\right) \leq \frac{(m+2)\sigma^2}{m} E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}|X\right) \ .$$

Multiplying both sides by X_i^2 and taking expectation leads to

$$E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2\right) \leq \frac{(m+2)\sigma^2}{m} E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2\right) \ .$$

Since we have already shown that

$$E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \tau^2 W}\right)^2 X_i^2\right] \leq E\left(\frac{\lambda_i M^2\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2\right) \times E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}\right) \ ,$$

in order to prove

$$E\left[\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)^2 X_i^2\right] \leq \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i + \tau^2} E\left(\frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}X_i^2\right) \ ,$$

it is then sufficient to show

$$\frac{(m+2)\sigma^2}{m} E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}\right) \leq \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i + \tau^2} \ .$$

As in the proof of Corollary 3.2., using the covariance inequality twice, we have

$$\begin{aligned}
& \frac{(m+2)\sigma^2}{m} E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min}+\tau^2)W}\right) \\
= & \frac{(m+2)\sigma^2}{m} E\left[E\left(\frac{\lambda_i\gamma_i/W}{\nu_i M\gamma_i/W + (\sigma^2\gamma_{min}+\tau^2)}\right)|M\right] \\
= & \frac{(m+2)\sigma^2}{m} E\left[\frac{E(\lambda_i\gamma_i/W|M)}{E[(\nu_i M\gamma_i/W + (\sigma^2\gamma_{min}+\tau^2))|M]}\right] \\
= & \frac{(m+2)\sigma^2}{m} E\left[\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (p-2)(\sigma^2\gamma_{min}+\tau^2)}\right] \\
= & \frac{(m+2)\sigma^2}{m} E\left[\frac{\lambda_i\gamma_i/M}{\nu_i\gamma_i + (p-2)(\sigma^2\gamma_{min}+\tau^2)/M}\right] \\
\leq & \frac{(m+2)\sigma^2}{m} \frac{E(\lambda_i\gamma_i/M)}{E(\nu_i\gamma_i + (p-2)(\sigma^2\gamma_{min}+\tau^2)/M)} \\
= & \frac{(m+2)\sigma^2}{m} \frac{\lambda_i\gamma_i\cdot m/(m-2)}{\nu_i\sigma^2\gamma_i + (p-2)(\sigma^2\gamma_{min}+\tau^2)\cdot m/(m-2)} \\
= & \frac{(m+2)\lambda_i\gamma_i\cdot\sigma^2}{(m-2)\nu_i\sigma^2\gamma_i + m(p-2)(\sigma^2\gamma_{min}+\tau^2)} .
\end{aligned}$$

Now applying the condition $0 \leq \lambda_i \leq \frac{2m(p-2)}{m+2}$ and $\nu_i \geq (\frac{m+2}{2(m-2)}\lambda_i - \frac{m\gamma_{min}(p-2)}{\gamma_i(m-2)})_+$, we finally have

$$\frac{(m+2)\lambda_i\gamma_i\cdot\sigma^2}{(m-2)\nu_i\sigma^2\gamma_i + m(p-2)(\sigma^2\gamma_{min}+\tau^2)} \leq \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i+\tau^2}$$

which completes the proof. □

For the corresponding positive part estimator δ^+_{GSV} given by

$$(\delta^+_{GSV}(X))_i = \left(1 - \frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right)_+ X_i \ , \tag{4.3}$$

as in the case of known variance, a slightly stronger result holds.

THEOREM 4.2. *δ^+_{GSV} in (4.3) is ensemble minimax if $p \geq 3$, $m \geq 3$ and for any $i = 1, \cdots, p$, $0 \leq \lambda_i \leq \frac{2m(p-2)}{m+2}$ and $\nu_i \geq (\frac{m+2}{m-2}\lambda_i - \frac{m(p-2)}{m-2}(1+\frac{\gamma_{min}}{\gamma_i}))_+$.*

PROOF. The proof follows similar steps in the proofs of Corollary 3.3. and Theorem 4.1., therefore, we will skip most of the details and only highlight

the parts that are substantially different. First let the shrinkage factor

$$h_i^+(X,M) = \min(1, h_i(X,M)) = \min\left(1, \frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + \|X\|^2}\right) .$$

As before, we have to prove

$$E\left[h_i^+(X,M)^2 X_i^2\right] \le \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i + \tau^2} E\left[h_i^+(X,M) X_i^2\right] .$$

When $\sigma^2\gamma_i \ge \tau^2$, the above inequality is trivial. From now on, assume $\sigma^2\gamma_i < \tau^2$. As in the proof of Theorem 4.1., we have

$$E\left[h_i^+(X,M)^2 X_i^2\right] \le E\left(h_i^+(X,M) M X_i^2\right) \times E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}\right) .$$

Define

$$\mu_s(M) = \min\left(1, \frac{\lambda_i M\gamma_i}{\nu_i M\gamma_i + s}\right) .$$

Note that the ratio $r(M) = \mu_s(M)/\mu'_s(M)$ is still non-decreasing in M for $s > s'$. As in Theorem 4.1., applying Lemma 4.2. and monotone convergence theorem leads to

$$E\left(h_i^+(X,M) M X_i^2\right) \le \frac{(m+2)\sigma^2}{m} E\left(h_i^+(X,M) X_i^2\right) .$$

It is then sufficient to show

$$\frac{(m+2)\sigma^2}{m} E\left(\frac{\lambda_i\gamma_i}{\nu_i M\gamma_i + (\sigma^2\gamma_{min} + \tau^2)W}\right) \le \frac{2\sigma^2\gamma_i}{\sigma^2\gamma_i + \tau^2}$$

whose proof follows exactly the same argument used in the last part of the proof of Corollary 3.3. □

When $\lambda_i = C$ and $\nu_i = 0$, δ_{GSV} in (4.2) and δ^+_{GSV} in (4.3) reduce to the James-Stein estimator and its positive part for the unknown variance case. Similar to the known variances case, the choice of C in the above theorems is different from that in the homoscedastic case. For the homoscedastic case and ordinary minimaxity, the upper bound of the constant C can be chosen to be as large as $\frac{2m(p-2)}{m+2}$ for the original James-Stein estimators. While for our case, the upper bound becomes $\frac{m(p-2)}{m+2}$. Like in Theorem 3.2., it can be shown that the bound can not be easily improved. However, we omit the result here for simplicity.

We can also extend the parametric Bayes estimator δ_{PEB} in (3.13) and δ^+_{PEB} in (3.14) to the unknown variance case. Consider δ_{PEBV} with the form

$$(\delta_{PEBV}(X))_i = \left(1 - \frac{CM\gamma_i}{CM\gamma_i + (\sum_{j=1}^p X_j^2 - \sum_{j=1}^p M\gamma_j)}\right) X_i \tag{4.4}$$

and δ^+_{PEBV} with the form

$$(\delta^+_{PEBV}(X))_i = \left(1 - \frac{CM\gamma_i}{CM\gamma_i + (\sum_{j=1}^p X_j^2 - \sum_{j=1}^p M\gamma_j)_+}\right) X_i. \tag{4.5}$$

The following corollary gives the conditions that guarantee the ensemble minimaxity of δ_{PEBV} in (4.4) and δ^+_{PEBV} in (4.5).

COROLLARY 4.1. *Assume $m \geq 6$, $p \geq 3$ and*

$$p \leq \frac{\sum_{j=1}^p \gamma_i}{\gamma_{min}} \leq C \leq \frac{2m(p-2)}{m+2}. \tag{4.6}$$

Then both δ_{PEBV} in (4.4) and δ^+_{PEBV} in (4.5) are ensemble minimax.

PROOF. Set $\lambda_i = C$ and $\nu_i = C - \frac{\sum_{j=1}^p \gamma_i}{\gamma_{min}}$. Condition (4.6) guarantees that $\nu_i \geq 0$ and $\lambda_i \leq \frac{2m(p-2)}{m+2}$. It is evident that $\delta_{PEBV} = \delta_{GSV}$. A little care with the positive part conditions shows that also $\delta^+_{PEBV} = \delta^+_{GSV}$.

It then follows from Theorem 4.1 that $\delta_{PEBV} = \delta_{GSV}$ is ensemble minimax if

$$\text{diff} = \nu_i - [\frac{m+2}{2(m-2)}C - \frac{m(p-2)\gamma_{min}}{(m-2)\gamma_i}] \geq 0. \tag{4.7}$$

Substituting and simplifying yields

$$\begin{aligned}
\text{diff} &= \frac{m-6}{2(m-2)}C + \frac{m(p-2)\gamma_{min}}{(m-2)\gamma_i} - \frac{\sum_{j=1}^p \gamma_j}{\gamma_i} \\
&\geq \frac{m-6}{2(m-2)}\frac{\sum_{j=1}^p \gamma_j}{\gamma_i} + \frac{m(p-2)\gamma_{min}}{(m-2)\gamma_i} - \frac{\sum_{j=1}^p \gamma_j}{\gamma_i} \\
&= \frac{\gamma_{min}}{\gamma_i}[\frac{m(p-2)}{m-2} - \frac{m+2}{2(m-2)}\frac{\sum_{j=1}^p \gamma_j}{\gamma_{min}}] \\
&\geq \frac{\gamma_{min}}{\gamma_i}[\frac{m(p-2)}{m-2} - \frac{m+2}{2(m-2)}\frac{2m(p-2)}{m+2}] \\
&= 0\ .
\end{aligned}$$

This verifies (4.7) and proves δ_{PEBV} is ensemble minimax.

The proof for δ^+_{PEBV} is similar, but easier. λ_i and ν_i are defined as before. Condition (4.6) is still required in order that $0 \le \lambda_i \le \frac{2m(p-2)}{m+2}$ and $\nu_i \ge 0$. Truth of (4.7) validates the remaining condition in Theorem 4.2, and hence proves δ^+_{PEBV} is ensemble minimax. □

5. Shrinkage towards the Common Mean. In the sections above, we discuss the ensemble minimaxity properties of the estimators that shrink towards zero under the heteroscedastic model. We will generalize our method to provide a class of ensemble minimax estimators that shrink towards the common mean in this section. Assume that $X \sim N(\theta, \Sigma)$ and $\theta \sim N(\mu 1, \tau^2 I)$, where Σ is the covariance matrix. We first present a Lemma whose proof is sufficiently simple to be omitted.

5.1. *General Theory.*

LEMMA 5.1. *There exists an orthogonal matrix Q with the form*

$$Q = \begin{pmatrix} \frac{1}{\sqrt{p}} 1^T \\ Q_2 \end{pmatrix},$$

such that $T = Q\Sigma Q^T$ can be written in the block matrix form

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$$

where T_{11} is 1×1, and $T_{22} = diag\{t_{22}, \cdots, t_{pp}\}$ is a $(p-1) \times (p-1)$ diagonal matrix.

From the fact that Q is orthogonal, we have

$$\begin{aligned} Q_2 1 &= 0 \\ Q_2 Q_2^T &= I_{p-1} \\ Q_2^T Q_2 &= I_p - \frac{1}{p} 1 1^T . \end{aligned}$$

Moreover, we also have $T_{22} = Q_2 \Sigma Q_2^T$. Since Σ is positive definite, we can easily verify that T_{22} is also positive definite. Therefore, $t_{ii} > 0$ for all $i = 2, \cdots, p$. Assume $p \ge 4$. Let $Y = QX$,$\eta = Q\theta$ and $Y_{(2)} = (Y_2, \cdots, Y_p)^T$. Then we have

$$Y = \begin{pmatrix} \frac{1}{\sqrt{p}} 1^T \\ Q_2 \end{pmatrix} (\overline{X} 1 + (X - \overline{X} 1)) = \begin{pmatrix} \sqrt{p} \overline{X} \\ Q_2 (X - \overline{X} 1) \end{pmatrix},$$

which implies $Y_1 = \sqrt{p}\overline{X}$ and $Y_{(2)} = Q_2(X - \overline{X}1)$. Note that $Q\mu 1 = (\sqrt{p}\mu, 0, \cdots, 0)^T$ and $Q\text{diag}(\tau^2 I_p)Q^T = \tau^2 I_p$. Consider the estimator δ_{cm} with the form

$$\delta_{cm}(X) = Q^T \left(Y_1, \xi_2(Y_{(2)}), \cdots, \xi_p(Y_{(2)})\right)^T , \tag{5.1}$$

where $\xi_i(Y_{(2)})$ is any ensemble minimax estimator for $\eta_{(2)}$, $\forall i = 2, \cdots, p$. We then have the following result.

THEOREM 5.1. *For $p \geq 4$, δ_{cm} in (5.1) is ensemble minimax.*

PROOF. Since $\xi(Y_{(2)})$ is an ensemble minimax estimator for $\eta_{(2)}$, we have that

$$E\left[\sum_{i=2}^{p}\left(\xi_i(Y_{(2)}) - \eta_i\right)^2\right] \leq \text{trace}(T_{22}) ,$$

which along with

$$E[(Y_1 - \eta_1)^2] = T_{11}$$

and $\text{trace}(T) = \text{trace}(\Sigma)$ implies

$$E\left[(Y_1 - \eta_1)^2 + \sum_{i=2}^{p}\left(\xi_i(Y_{(2)}) - \eta_i\right)^2\right] \leq \text{trace}(\Sigma) .$$

Therefore, we have

$$\begin{aligned} & E\left[\sum_{i=1}^{p}((\delta_c(X))_i - \theta_i)^2\right] \\ = \; & E\left[(Q^T(Y_1 - \eta_1, \xi_2(Y_{(2)}) - \eta_2, \cdots, \xi_p(Y_{(2)}) - \eta_p)^T)^T\right. \\ & \left. \cdot\, (Q^T(Y_1 - \eta_1, \xi_2(Y_{(2)}) - \eta_2, \cdots, \xi_p(Y_{(2)}) - \eta_p)^T)\right] \\ = \; & E\left[(Y_1 - \eta_1)^2 + \sum_{i=2}^{p}\left(\xi_i(Y_{(2)}) - \eta_i\right)^2\right] \leq \text{trace}(\Sigma) \end{aligned}$$

which completes the proof. □

Note that δ_{cm} in (5.1) can be interpreted as "shrinking" towards the overall mean since it can be written as $\delta_{cm}(X) = \overline{X}1 + Q_2^T\xi(Q_2(X - \overline{X}1))$, which is a generalized shrinkage estimator.

Furthermore, if we assume that $\xi_i(Y_{(2)}) = (1 - h_i(\|Y_{(2)}\|^2))Y_i$ for $i = 2, \cdots, p$, we have

$$\begin{aligned}\delta_{cm}(X) &= \overline{X}1 + Q_2^T \mathrm{diag}\{1 - h_2(\|Y_{(2)}\|^2), \cdots, 1 - h_p(\|Y_{(2)}\|^2)\}Y_{(2)} \\ &= \overline{X}1 + Q_2^T \mathrm{diag}\{1 - h_2(\|Y_{(2)}\|^2), \cdots, 1 - h_p(\|Y_{(2)}\|^2)\}Q_2(X - \overline{X}1) \ ,\end{aligned}$$

which, along with the fact that $\|Y_{(2)}\|^2 = \|Q_2(X - \overline{X}1)\|^2 = \|X - \overline{X}1\|^2$, implies

$$\delta_{cm}(X) = \overline{X}1 + D \cdot (X - \overline{X}1)$$

with $D = Q_2^T \mathrm{diag}\{1 - h_2(\|X - \overline{X}1\|^2), \cdots, 1 - h_p(\|X - \overline{X}1\|^2)\}Q_2$.

5.2. *Random Effects Models.* The standard one-way random effects model involves observations of independent variables Y_{ij}, $i = 1, \cdots, p$, $j = 1, \cdots, J_i$ under the distributional assumptions

$$\begin{aligned}Y_{ij}|\mu, \tau^2 &\sim N(\theta_i, \sigma^2) \\ \theta_i &\sim N(\mu, \tau^2) \text{ (independent)} \ .\end{aligned}$$

Here, the unknown parameters are σ^2, μ, τ^2. To fit with previous notation, let

$$\begin{aligned}X_i &= Y_{i\cdot} = \frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} \\ M &= \frac{1}{n-p}\sum_{i}^{p}\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\cdot})^2, \ m = \sum_{i=1}^{p}(J_i - 1) = n - p \ .\end{aligned}$$

Thus, M denotes the usual sum of squared errors and m denotes the degrees of freedom for error.

The goal in random effects estimation (sometimes referred to as "prediction") is to estimate (predict) the values of θ_i under ordinary squared error loss

$$L(\delta, \theta) = \sum_{i=1}^{p}(\delta_i - \theta_i)^2 \ .$$

The usual estimator (predictor) is of course $\delta_i(X) = X_i$. This problem is clearly mathematically equivalent to the ensemble minimaxity formulation. Hence, ensemble minimaxity in the hierarchical formulation is identical to ordinary minimaxity for the estimation of $\{\theta_i\}$ in the random effects model.

We construct a class of ensemble minimax generalized shrinkage estimators following the approach in Section 5.1. Let $\Gamma = diag\{1/J_1, \cdots, 1/J_p\}$. From Lemma 5.1, there exists an orthogonal matrix Q with the form

$$Q = \begin{pmatrix} \frac{1}{\sqrt{p}} 1^T \\ Q_2 \end{pmatrix},$$

such that $T = Q\Gamma Q^T$ can be written in the block matrix form

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$$

where T_{11} is 1×1, and $T_{22} = \text{diag}\{t_{22}, \cdots, t_{pp}\}$ is a $(p-1) \times (p-1)$ diagonal matrix.

Assume $p \geq 4$. Let $Y = QX$,$\eta = Q\theta$ and $Y_{(2)} = (Y_2, \cdots, Y_p)^T$. Consider the estimator δ_{cmv} with the form

$$\delta_{cmv}(X) = Q^T \left(Y_1, \xi_2(Y_{(2)}, M), \cdots, \xi_p(Y_{(2)}, M)\right)^T , \tag{5.2}$$

where $\xi_i(Y_{(2)}, M)$ is any ensemble minimax estimator for $\eta_{(2)}$, $\forall i = 2, \cdots, p$. Note that δ_{cmv} in (5.2) can be interpreted as "shrinking" towards the overall mean since it can be written as $\delta_{cmv}(X) = \overline{X}1 + Q_2^T \xi(Q_2(X - \overline{X}1), M)$, which is a generalized shrinkage estimator. Following the similar approach in Theorem 5.1, it is easy to verify that δ_{cmv} in (5.2) is ensemble minimax.

Especially, if we choose δ^+_{GSV} as ξ here, we get the estimator $\delta^+_{GSV;re}$ with the form

$$(\delta^+_{GSV;re}(X))_i = \overline{X}1 + Q_2^T A Q_2 (X - \overline{X}1) , \tag{5.3}$$

where $A = diag\{(1 - \frac{\lambda_2 M t_{22}}{\nu_2 M t_{22} + ||X - \overline{X}1||^2})_+, \cdots, (1 - \frac{\lambda_p M t_{pp}}{\nu_p M t_{pp} + ||X - \overline{X}1||^2})_+\}$. We have the following corollary that shows its ensemble mimimax properties.

COROLLARY 5.1. *$\delta^+_{GSV;re}$ in (5.3) is ensemble minimax if $p \geq 4$, $m \geq 3$ and for any $i = 2, \cdots, p$, $0 \leq \lambda_i \leq \frac{2m(p-3)}{m+2}$ and $\nu_i \geq (\frac{m+2}{m-2}\lambda_i - \frac{m(p-3)}{m-2}(1 + \frac{t_{min}}{t_{ii}}))_+$. Hence, $\delta^+_{GSV;re}$ is minimax for the random effects model under these conditions and dominates the usual estimator $\delta_i(X) = X_i$.*

In the interest of space we omit the formal proof. If a single version of the above estimators is to be used for all $J_1, \cdots, J_p$, the preferred and simple choice would be $\delta^+_{GSV;re}$ in (5.3) with $\lambda_i = \frac{m(p-3)}{m+2}$ and $\nu_i = 0$.

References.

[1] Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya, Series A*, **15**, 377–380.

[2] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer, New York. MR1700749

[3] Berger, R.L. (1979). Gamma minimax robustness of Bayes rules. *Communications in Statistics: Theory and Methods*, **8**, 543–560.

[4] Brown, L.D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ananals of Mathematical Statistics*, **37**, 1087–1136.

[5] Brown, L.D. (1975). Estimation with incomplete specified loss functions (the case with several location parameters). *Journal of the American Statistical Association*, **70**, 417–427.

[6] Brown, L.D. (2008). In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Annals of Applied Statistics*, **2**, 113–152.

[7] Carter, G.M. and Rolph, J.E. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, **69**, 880–885.

[8] Casella, G. (1980). Minimax ridge regression estimation. *Ananals of Statistics*, **8**, 1036–1056.

[9] Copas, J.B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series B (Methodological)*, **45**, 311–354.

[10] Efron, B. and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators - Part I: Bayes case. *Journal of the American Statistical Association*, **66**, 807–815.

[11] Efron, B. and Morris, C. (1972a). Limiting the risk of Bayes and empirical Bayes estimators - Part II: The empirical Bayes case. *Journal of the American Statistical Association*, **67**, 130–139.

[12] Efron, B. and Morris, C. (1972b). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, **59**, 335–347.

[13] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – An empirical Bayes approach. *Journal of the American Statistical Association*, **68**, 117–130.

[14] Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70**, 311–319.

[15] Fay, R.E.III and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

[16] Good, I.J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B (Methodological)*, **14**, 107–114.

[17] Green, J. and Strawderman, W.E. (1985). The use of Bayes/empirical Bayes estimation in individual tree volume development. *Forest Science*, **31**, 975–990.

[18] Hastie, T., Tibshirani, R., C. and Friedman, J. (2001). *The elements of statsitical learning: data mining, inference and prediction*, Springer, New York.

[19] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of 4th Berkeley Symposium on Probability and Statistics*, **I**, 367–379.

[20] Jones, K. (1991). Specifying and estimating multi-level models for gerographical research *Transactions of the institute of british geographers*, **16**, 148–159.

[21] Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B (Methodological)*, **34**, 1–41.

[22] Morris, C. (1983a). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, **78**, 47–55.

[23] Morris, C. (1983b). Parametric empirical Bayes confidence intervals. in *Scientific Inference, Data Analysis, and Robustness*, eds. Box, G.E.P., Leonard, T. and Wu, C.F.J., New York: Academic Press, 25–50.

[24] Morris, C. and Lysy, M. (2009). Shrinkage estimation in multi-level normal models. *preprint*

[25] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, **1**, University of California Press, Berkeley.

[26] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, **35**, 1–20.

[27] Rubin, D. (1981). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association*, **75**, 801–827.

[28] Strawderman, W. (1971). Proper Bayes estimators of the multivariate normal mean. *Annals of Mathematical Statistics*, **42**, 385–388.

Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104
E-mail: lbrown@wharton.upenn.edu
E-mail: niehui@wharton.upenn.edu

Department of Statistics
Harvard University
Cambridge, MA 02138
E-mail: xxie@fas.harvard.edu