Statistica Sinica **26** (2016), 1219-1248
doi:http://dx.doi.org/10.5705/ss.202014.0069

STEINIZED EMPIRICAL BAYES ESTIMATION FOR HETEROSCEDASTIC DATA

Zhiqiang Tan

Rutgers University

Abstract: Consider the problem of estimating normal means from independent observations with known variances, possibly different from each other. Suppose that a second-level normal model is specified on the unknown means, with the prior means depending on a vector of covariates and the prior variances constant. For this two-level normal model, existing empirical Bayes methods are constructed from the Bayes rule with the prior parameters selected either by maximum likelihood or moment equations or by minimizing Stein's unbiased risk estimate. Such methods tend to deteriorate, sometimes substantially, when the second-level model is misspecified. We develop a Steinized empirical Bayes approach for improving the robustness to misspecification of the second-level model, while preserving the effectiveness in risk reduction when the second-level model is appropriate in capturing the unknown means. The proposed methods are constructed from a minimax Bayes estimator or, interpreted by its form, a Steinized Bayes estimator, which is not only globally minimax but also achieves close to the minimum Bayes risk over a scale class of normal priors including the specified prior. The prior parameters are then estimated by standard moment methods. We provide formal results showing that the proposed methods yield no greater asymptotic risks than existing methods using the same estimates of prior parameters, but without requiring the second-level model to be correct. We present both an application for predicting baseball batting averages and two simulation studies to demonstrate the practical advantage of the proposed methods.

Key words and phrases: Bayes estimation, empirical Bayes, Fay–Herriot model, minimax estimation, small-area estimation, Stein's unbiased risk estimate, subspace shrinkage, unequal variance.

1. Introduction

Consider the following problem of estimating normal means with heteroscedastic observations. Assume that $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is a $n \times 1$ vector of independent and normally distributed observations:

$$Y_j|\theta_j \sim \mathrm{N}(\theta_j, d_j), \quad j = 1, \ldots, n, \tag{1.1}$$

where the means $\theta = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$ are unknown and the variances $(d_1, \ldots, d_n)$ are known but possibly different from each other. In matrix notation, the model says

$Y \sim \mathrm{N}(\theta, D)$, where $D = \mathrm{diag}(d_1, \ldots, d_n)$. In addition, let $X = (x_1, \ldots, x_n)^{\mathrm{T}}$, where x_j is a $q \times 1$ vector of covariates, possibly depending on d_j but considered to be non-random. The problem of interest is to estimate the mean vector θ from the observation vector Y, with the assistance of the covariate matrix X.

We adopt a decision-theoretic framework (e.g., Lehmann and Casella (1998)) for evaluating the performance of an estimator $\delta = (\delta_1, \ldots, \delta_n)^{\mathrm{T}}$ for θ, where δ_j can depend on all $(Y_1, \ldots, Y_n)$ for each j. The (pointwise) risk of δ is defined as

$$R(\delta, \theta) = \sum_{j=1}^{n} E_\theta\{(\delta_j - \theta_j)^2\},$$

where $E_\theta(\cdot)$ denotes the expectation with respect to the distribution of Y given θ. It is desirable to construct an estimator δ such that $R(\delta, \theta)$ is small in some overall sense. Two standard optimality criteria are to minimize the maximum risk over $\theta \in \mathbb{R}^n$ and to minimize the Bayes risk with a prior distribution on θ, corresponding to, respectively, minimax estimation and Bayes estimation. There is also a substantial literature, not considered here, on minimax estimation of normal means restricted to a subset of $\mathbb{R}^n$. For example, the restricted space of θ can be defined as an ℓ_p ball, $\{\theta \in \mathbb{R}^n : \sum_{j=1}^n |\theta_j|^p \le C^p\}$, and this framework is useful for estimation of θ assumed to be sparse and for nonparametric function estimation (e.g., Donoho and Johnstone (1994); Johnstone (2013)). See also Ben-Hain and Eldar (2007) for an estimation approach based on minimax estimators over a bounded parameter set.

For the present problem, minimaxity over $\theta \in \mathbb{R}^n$ is equivalent to a simple property: for any $\theta \in \mathbb{R}^n$, the risk $R(\delta, \theta)$ is no greater than $\sum_{j=1}^n d_j$, the risk of the usual estimator $\delta_0 = Y$. For $n \ge 3$, minimax estimators different from and hence dominating δ_0 were first discovered by Stein (1956) and James and Stein (1961) in the homoscedastic case (i.e., $d_1 = \cdots = d_n$). Minimax estimation has since been extensively studied (e.g., Lehmann and Casella (1998); Strawderman (2010)). In the general heteroscedastic case, all existing estimators, except those in Berger (1982) and Tan (2015), fall into two categories, each with some limitations. Estimators in the first category are minimax over $\theta \in \mathbb{R}^n$ only under some restrictive conditions on how much $(d_1, \ldots, d_n)$ can differ from each other (e.g., Bock (1975); Brown (1975)). Estimators in the second category (e.g., Berger (1976)) are minimax regardless of differences between $(d_1, \ldots, d_n)$, but the observations Y_j are shrunk inversely in proportion to the variances d_j so that the risk reduction achieved over δ_0 is insubstantial unless all the observations have similar variances.

The Bayes approach requires a prior distribution to be specified on θ, but leads directly to a Bayes rule that achieves the minimum Bayes risk. Consider a

class of normal priors $\pi_{\gamma,\beta}$ such that $(\theta_1, \ldots, \theta_n)$ are independent and

$$\theta_j \sim \mathrm{N}(x_j^{\mathrm{T}}\beta, \gamma), \quad j = 1, \ldots, n, \tag{1.2}$$

where β and γ are hyper-parameters, with β a vector of regression coefficients and γ a prior variance. Recall that θ_j is allowed to depend on d_j through x_j. In matrix notation, the prior $\pi_{\gamma,\beta}$ says $\theta \sim \mathrm{N}(X\beta, \gamma I)$ with I the identity matrix. For any fixed (γ, β), the Bayes risk of δ is defined as $R(\delta, \pi_{\gamma,\beta}) = \int R(\delta, \theta)\pi_{\gamma,\beta}(\theta)\,\mathrm{d}\theta$, and the Bayes rule is given componentwise by

$$(\delta^{\mathrm{B}}_{\gamma,\beta})_j = Y_j - \frac{d_j}{d_j + \gamma}(Y_j - x_j^{\mathrm{T}}\beta).$$

In contrast with Berger's (1976) minimax estimation, the greater d_j is, the more Y_j is shrunk to $x_j^{\mathrm{T}}\beta$. Informally speaking, the Bayes rule can achieve a much smaller pointwise risk than $\sum_{j=1}^n d_j$, where the true values $(\theta_1, \ldots, \theta_n)$ can be seen as a typical sample from the prior distribution. But, for any fixed choices of (γ, β), the Bayes rule is non-minimax and hence can have a pointwise risk exceeding $\sum_{j=1}^n d_j$, when the true values $(\theta_1, \ldots, \theta_n)$ are incompatible with the prior distribution.

To mitigate the difficulty of prior specification, an empirical Bayes approach based on (1.1) and (1.2) is to estimate (γ, β) as unknown parameters in the marginal distribution, $Y_j \sim \mathrm{N}(x_j^{\mathrm{T}}\beta, d_j + \gamma)$, and then substitute the estimates of (β, γ) into the Bayes rule for estimating θ (Efron and Morris (1973); Morris (1983)). This approach is known to be equivalent to empirical best unbiased linear prediction in a two-level normal model (e.g., Datta and Ghosh (2012); Morris and Lysy (2012)):

$$\left.\begin{array}{l}\text{First-level:} \quad Y_j = \theta_j + \varepsilon_j \\ \text{Second-level: } \theta_j = x_j^{\mathrm{T}}\beta + u_j\end{array}\right\} \quad j = 1, \ldots, n, \tag{1.3}$$

where β and γ are unknown parameters, $(u_1, \ldots, u_n)$ are independent with $u_j \sim \mathrm{N}(0, \gamma)$ and, independently, $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent with $\varepsilon_j \sim \mathrm{N}(0, d_j)$. For an important application, model (1.3) is also called Fay and Herriot's (1979) model in small-area survey estimation, where Y_j represents a direct survey estimate and d_j is an estimated variance but treated as the true variance. See, for example, Rao (2003) and Pfeffermann (2013) for further discussion on small-area estimation.

The empirical Bayes approach relaxes the Bayes approach in depending on the appropriateness of a class of priors instead of a single prior. But similarly as discussed for the Bayes approach, the performance of empirical Bayes estimation of θ is still affected by how well the true values $(\theta_1, \ldots, \theta_n)$ are captured by any

prior in the class (1.2) or, equivalently, by the second-level model, $\theta_j = x_j^{\mathrm{T}}\beta + u_j$, in (1.3). To address this issue, Jiang, Nguyen, and Rao (2011) and Xie, Kou, and Brown (2012) proposed two closely related methods for estimating θ, although Xie, Kou, and Brown (2012) considered only the case where $x_j \equiv 1$ and β is one-dimensional and then allowed more flexible shrinkage factors than in the form $d_j/(d_j+\gamma)$. With such differences ignored, their estimators are in fact equivalent to each other, both in the form of the Bayes rule $\delta^{\mathrm{B}}_{\gamma,\beta}$ similar to model-based empirical Bayes estimators, but with (γ, β) selected by minimizing Stein's (1981) unbiased risk estimate (SURE) of $\delta^{\mathrm{B}}_{\gamma,\beta}$. The resulting estimator of θ is expected to be more robust than model-based empirical Bayes estimators to misspecification of the second-level model (1.2), but the protection is, in general, limited because the estimator is still constructed from the Bayes rule based on (1.2).

We develop a new approach, called Steinized empirical Bayes estimation, for using the second-level model (1.2) but without requiring (1.2) to be correctly specified. As described above, both the model-based and the SURE-based empirical Bayes approaches involve finding the Bayes rule $\delta^{\mathrm{B}}_{\gamma,\beta}$ under a prior (1.2) with fixed (γ, β) and then selecting (γ, β) according to some empirical criterion. By comparison, there are two corresponding steps in the proposed approach.

The first step is to develop a minimax Bayes estimator that is not only minimax over $\theta \in \mathbb{R}^n$ but also achieves close to the minimum Bayes risk under a prior (1.2) with fixed (γ, β). The James–Stein estimator has these two properties in the homoscedastic case (i.e., $d_1 = \cdots = d_n$) by Efron and Morris (1973). In the general heteroscedastic case, the estimator of Berger (1982) achieves the two desired properties, but seems complicated and difficult to interpret. Alternatively, Tan (2015) developed a shrinkage estimator, which is not only simple and easily interpretable, but also is minimax over $\theta \in \mathbb{R}^n$ and achieves close to the minimum Bayes risk over a scale class of normal priors including the specified prior. In fact, the estimator of Tan (2015) leads to a shrinkage pattern such that one group of observations are shrunk in the direction of Berger's (1976) minimax estimator and the remaining observations are shrunk in the direction of the Bayes rule $\delta^{\mathrm{B}}_{\gamma,\beta}$. Moreover, the observations are shrunk in these directions by a scalar factor similar to that in Stein's (1981) and Li's (1985) linear shrinkage estimators. Therefore, the minimax Bayes estimator of Tan (2015) can be interpreted, by its form, as a Steinized Bayes estimator.

The second step of our approach, following the general idea of empirical Bayes, is to choose (γ, β) in a data-dependent manner for the minimax Bayes estimator of Tan (2015). In principle, both model-based and SURE-based strategies can be used. On one hand, the SURE of the estimator of Tan (2015) is, in general, a non-smooth, multi-modal function of (γ, β), which makes it computationally difficult to find a global minimum of the SURE. From pilot simulation studies,

we found that the values of (γ, β), even if identified, might appear unnatural in showing how the second-level model (1.2) could be fitted to the true θ-values. For these reasons, the SURE-based strategy seems unappealing for empirically selecting (γ, β) for the estimator of Tan (2015). On the other hand, among several model-based estimators for (γ, β) (e.g., Datta, Rao, and Smith (2005)), the moment method in Fay and Herriot (1979) seems particularly attractive. For this method, not only the observations and residuals are weighted in a balanced manner as discussed in Fay and Herriot (1979), but also the estimating function for γ is always monotonic, which facilitates numerical solution of the estimating equation and development of asymptotic theory.

We propose two Steinized empirical Bayes methods based on the foregoing ideas. The first method involves substituting the Fay–Herriot estimates $(\hat{\gamma}, \hat{\beta})$ for (γ, β) in the minimax Bayes estimator of Tan (2015). The second method, developed from the perspective of subspace shrinkage (e.g., Sclove, Morris, and Radhakrishnan (1972); Oman (1982)), involves using a particular estimator $\tilde{\beta}$ of β such that the fitted location $X^{\mathrm{T}}\tilde{\beta}$ and the residual $Y - X^{\mathrm{T}}\tilde{\beta}$ are uncorrelated, and then applying to the residual the first Steinized empirical Bayes estimator, modified for shrinkage toward 0.

The Steinized empirical Bayes methods can be seen as a new approach for improving the robustness to misspecification of second-level model (1.2), while preserving the effectiveness in risk reduction when second-level model (1.2) is appropriate in capturing the unknown means. We provide asymptotic results on the convergence of the pointwise risks of the proposed estimators to that of the minimax Bayes estimator of Tan (2015) with (γ, β) replaced by the limit values of the estimates used, when $n \to \infty$ and q is fixed. Particularly, we show that the proposed estimators have no greater asymptotic risks than existing estimators using the same estimates of (γ, β). These results make clear that the model-based estimates of (γ, β) are mainly used to capture the center and spread of the true θ-values, so that effective risk reduction can be achieved due to near-Bayes optimality of the estimator of Tan (2015). We present both an application to the baseball data in Brown (2008) and two simulation studies to demonstrate the practical advantage of the proposed methods.

2. Minimax Bayes Estimation

We describe the minimax Bayes method in Tan (2015) for estimating normal means under heteroscedasticity. This method is developed by combining the two ideas of minimax and Bayes estimation in a principled manner.

First, a class of shrinkage estimators,$\delta_{A,\lambda}$, is constructed with the jth component

$$(\delta_{A,\lambda})_j = \left\{1 - \frac{\lambda c(D, A)}{Y^{\mathrm{T}} A^{\mathrm{T}} A Y} a_j\right\}_+ Y_j,$$

where $A = \mathrm{diag}(a_1, \ldots, a_n)$ with elements $a_j \geq 0$ indicating the direction of shrinkage, $\lambda \geq 0$ is a scalar indicating the magnitude of shrinkage, and $c(D, A) = (\sum_{j=1}^n d_j a_j) - 2\max_{j=1,\ldots,n}(d_j a_j)$. Throughout, $u_+ = \max(0, u)$ for any scalar u. The estimator $\delta_{A,\lambda}$ is invariant to a scale transformation $A \mapsto uA$for a scalar $u > 0$. If $c(D, A) \geq 0$, then an upper bound on the risk function of $\delta_{A,\lambda}$ is

$$R(\delta_{A,\lambda}, \theta) \leq \sum_{j=1}^n d_j - E_\theta \left\{ \frac{\lambda(2-\lambda)c^2(D,A)}{Y^{\mathrm{T}} A^{\mathrm{T}} A Y} \right\}. \tag{2.1}$$

Therefore, $\delta_{A,\lambda}$ is minimax over $\theta \in \mathbb{R}^n$ provided $c(D, A) \geq 0$ and $0 \leq \lambda \leq 2$.

Second, to measure average risk reduction in an elliptical region, consider a normal prior π_Γ, $\theta \sim \mathrm{N}(0, \Gamma)$, where $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_n)$ is a diagonal variance matrix (e.g., Berger (1982)). That is, $(\theta_1, \ldots, \theta_n)$ are independent and

$$\theta_j \sim \mathrm{N}(0, \gamma_j), \quad j = 1, \ldots, n.$$

By (2.1) and Jensen's inequality, if $c(D, A) \geq 0$, then an upper bound on the Bayes risk of $\delta_{A,\lambda}$ is $R(\delta_{A,\lambda}, \pi_\Gamma) \leq \sum_{j=1}^n d_j - \lambda(2-\lambda)c^2(D,A)/E^m(Y^{\mathrm{T}} A^{\mathrm{T}} A Y)$, where E^m denotes the expectation in the Bayes model, $Y_j \sim \mathrm{N}(0, d_j + \gamma_j)$. The shrinkage direction A for $\delta_{A,\lambda}$ is then chosen to minimize this upper bound on the Bayes risk subject to $c(D, A) \geq 0$ or, equivalently, to minimize $\sum_{j=1}^n (d_j + \gamma_j) a_j^2$ with $c(D, A) \geq 0$ being fixed at a particular value. This optimization problem turns out to admit a non-iterative solution, $A^\dagger$, as follows by (Tan (2015, Thm. 2)).

Assume that $n \geq 3$ and the indices are sorted such that $d_1^* \geq d_2^* \geq \cdots \geq d_n^*$ with $d_j^* = d_j^2/(d_j + \gamma_j)$. Then the solution $A^\dagger = \mathrm{diag}(a_1^\dagger, \ldots, a_n^\dagger)$ is determined, uniquely up to proportionality, by

$$a_j^\dagger = \left(\sum_{k=1}^\nu \frac{d_k + \gamma_k}{d_k^2} \right)^{-1} \frac{\nu - 2}{d_j} \quad (j = 1, \ldots, \nu), \tag{2.2}$$

$$a_j^\dagger = \frac{d_j}{d_j + \gamma_j} \quad (j = \nu + 1, \ldots, n), \tag{2.3}$$

where ν is the smallest index k such that $3 \leq k \leq n-1$ and $(k-2)/\{\sum_{j=1}^k (d_j + \gamma_j)/d_j^2\} > d_{k+1}^2/(d_{k+1} + \gamma_{k+1})$, or $\nu = n$ if there exists no such k. By (2.2) and (2.3), ν is the largest index k such that $d_1 a_1^\dagger = \cdots = d_k a_k^\dagger > d_j a_j^\dagger$ for $j \geq k+1$. Moreover, $C(D, A^\dagger) = \sum_{j=1}^n (d_j + \gamma_j) a_j^{\dagger 2}$. The resulting estimator of θ is

$$(\delta_{A^\dagger,\lambda})_j = \left\{ 1 - \frac{\lambda \sum_{k=1}^n a_j^{\dagger 2}(d_k + \gamma_k)}{\sum_{k=1}^n a_k^{\dagger 2} Y_k^2} a_j^\dagger \right\}_+ Y_j,$$

and is minimax over $\theta \in \mathbb{R}^n$ provided $0 \leq \lambda \leq 2$.

The estimator $\delta_{A^\dagger,\lambda}$ has an interesting simple form, related to both the Bayes rule, $\delta_j^{\mathrm{B}} = \{1 - d_j/(d_j + \gamma_j)\}Y_j$, and Berger's (1976) minimax estimator, $\delta_j^{\mathrm{Ber}} = \{1 - (n-2)d_j^{-1}/(Y^{\mathrm{T}}D^{-2}Y)\}_+ Y_j$. By (2.2) and (2.3), there is a dichotomous segmentation in the shrinkage direction of the observations Y_j 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 (1982). The observations Y_j with high d_j^* are shrunk inversely in proportion to their variances d_j as in Berger's (1976) estimator δ^{Ber}, whereas the observations Y_j with low d_j^* are shrunk in the direction of the Bayes rule. Therefore, $\delta_{A^\dagger,\lambda}$ mimics the Bayes rule to reduce the Bayes risk, except that $\delta_{A^\dagger,\lambda}$ mimics δ^{Ber} for some observations of highest Bayes "importance" in order to achieve minimaxity.

The shrinkage factor, $\lambda \sum_{k=1}^n a_k^{\dagger 2}(d_k + \gamma_k)/\{\sum_{k=1}^n a_k^{\dagger 2} Y_k^2\}$, in $\delta_{A^\dagger,\lambda}$ is similar to that in Stein's (1981) and Li's (1985) linear shrinkage estimators. Then $\delta_{A^\dagger,\lambda}$ can be interpreted as a Steinized Bayes estimator, except for the shrinkage of some observations Y_j of highest "Bayes" importance in the direction proportional to d_j^{-1}. In fact, if $a_j^\dagger$ were reset to $d_j/(d_j + \gamma_j)$ for all $j = 1, \ldots, n$, then $\delta_{A^\dagger,\lambda}$ would differ from the Bayes rule, $(1 - a_j^\dagger)Y_j$, only by the Stein-type shrinkage factor.

In addition to simplicity and minimaxity, the risk properties of $\delta_{A^\dagger,\lambda=1}$, with $\lambda = 1$ used, are further studied in Tan (2015). Write $\delta_{A^\dagger(\Gamma)} = \delta_{A^\dagger,\lambda=1}$ to make explicit the dependency of $A^\dagger$ on Γ. Let $\Gamma_\alpha = \alpha(D + \Gamma) - D$ and $\alpha_0 = \max_{j=1,\ldots,n}\{d_j/(d_j + \gamma_j)\}\,(\le 1)$. The Bayes risk of $\delta_{A^\dagger(\Gamma)}$ satisfies, for each $\alpha \ge \alpha_0$,

$$R\{\delta_{A^\dagger(\Gamma)}, \pi_{\Gamma_\alpha}\} \le R(\delta_{\Gamma_\alpha}^{\mathrm{B}}, \pi_{\Gamma_\alpha}) + \alpha^{-1}(d_1^* + d_2^* + d_3^* + d_4^*), \tag{2.4}$$

where $R(\delta_{\Gamma_\alpha}^{\mathrm{B}}, \pi_{\Gamma_\alpha}) = \sum_{j=1}^n d_j - \alpha^{-1}\sum_{j=1}^n d_j^*$, the Bayes risk of the Bayes rule with the prior $\mathrm{N}(0, \Gamma_\alpha)$. Therefore, $\delta_{A^\dagger(\Gamma)}$ achieves close to the minimum Bayes risk, with the difference no greater than the sum of the four highest Bayes "importance" of the obsevations, simultaneously over a scale class of normal priors, $\{\mathrm{N}(0, \Gamma_\alpha) : \alpha \ge \alpha_0\}$, including the specified prior $\mathrm{N}(0, \Gamma)$. This extends the previous result that in the homoscedastic case ($d_1 = \cdots = d_n$), the James–Stein estimator achieves the minimum Bayes risk up to the sum of two (equal-valued) Bayes "importance" over the scale class of homoscedastic normal priors (Efron and Morris (1973)).

The minimax Bayes method can be directly extended to accommodate a normal prior with a non-zero mean vector. For the normal prior (1.2), applying $\delta_{A^\dagger,\lambda}$ with $\gamma_j = \gamma$ and Y_j replaced by $Y_j - x_j^{\mathrm{T}}\beta$ leads to the following estimator of θ:

$$\delta_{\lambda,\gamma,\beta} = x_j^{\mathrm{T}}\beta + \left\{1 - \frac{\lambda \sum_{k=1}^n a_k^2(\gamma)(d_k + \gamma)}{\sum_{k=1}^n a_k^2(\gamma)(Y_k - x_k^{\mathrm{T}}\beta)^2} a_j(\gamma)\right\}_+ (Y_j - x_j^{\mathrm{T}}\beta),$$

where the indices are sorted such that $d_1 \geq d_2 \geq \ldots \geq d_n$ and

$$a_j(\gamma) = \left(\sum_{k=1}^{\nu} \frac{d_k + \gamma}{d_k^2} \right)^{-1} \frac{\nu - 2}{d_j} \quad (j = 1, \ldots, \nu), \tag{2.5}$$

$$a_j(\gamma) = \frac{d_j}{d_j + \gamma} \quad (j = \nu + 1, \ldots, n), \tag{2.6}$$

with ν being the smallest index k such that $3 \leq k \leq n-1$ and $(k-2)/\{\sum_{j=1}^{k}(d_j + \gamma)/d_j^2\} > d_{k+1}^2/(d_{k+1} + \gamma_{k+1})$, or $\nu = n$ if there exists no such k. The relative magnitudes of shrinkage in $\delta_{\lambda,\gamma,\beta}$ can be bounded as follows.

Lemma 1. *For any $\gamma \geq 0$ and $1 \leq j < k \leq n$ (hence $d_j \geq d_k$ by the assumption $d_1 \geq d_2 \geq \cdots \geq d_n$), we have*

$$\frac{d_k}{d_j} \leq \frac{a_j(\gamma)}{a_k(\gamma)} \leq \frac{d_j}{d_k}.$$

By minimaxity of $\delta_{A^\dagger,\lambda}$ (Tan (2015, Thm. 2)), the estimator $\delta_{\lambda,\gamma,\beta}$ is minimax over $\theta \in \mathbb{R}^n$ for any data-independent choices of $0 \leq \lambda \leq 2$, $\gamma \geq 0$, and $\beta \in \mathbb{R}^q$. Moreover, application of (2.4) with $\Gamma_\alpha = \alpha(D + \gamma I) - D$ and $\alpha_0 = \max_{j=1,\ldots,n}\{d_j/(d_j + \gamma)\}$ shows that the Bayes risk of $\delta_{\lambda=1,\gamma,\beta}$ satisfies, for each $\alpha \geq \alpha_0$,

$$R\{\delta_{\lambda=1,\gamma,\beta}, \pi_{\Gamma_\alpha}\} \leq R(\delta^{\mathrm{B}}_{\alpha,\gamma,\beta}, \pi_{\alpha,\gamma,\beta}) + \alpha^{-1}(d_1^* + d_2^* + d_3^* + d_4^*), \tag{2.7}$$

where $\pi_{\alpha,\gamma,\beta}$ denotes the prior $\theta \sim \mathrm{N}(X^{\mathrm{T}}\beta, \Gamma_\alpha)$, $R(\delta^{\mathrm{B}}_{\alpha,\gamma,\beta}, \pi_{\alpha,\gamma,\beta})$ is the Bayes risk of the corresponding Bayes rule, and $d_j^* = d_j^2/(d_j + \gamma)$ for $j = 1, \ldots, n$.

3. Steinized Empirical Bayes Estimation

For any fixed choices $0 \leq \lambda \leq 2$, $\gamma \geq 0$, and $\beta \in \mathbb{R}^q$, the estimator $\delta_{\lambda,\gamma,\beta}$ is minimax over $\theta \in \mathbb{R}^n$ and $\delta_{\lambda=1,\gamma,\beta}$ is scale-adaptive in achieving close to the minimum Bayes risk over a scale class of normal priors (including the specified prior). Nevertheless, the pointwise risk $R(\delta_{\lambda,\gamma,\beta}, \theta)$ for a particular unknown θ is still affected by the choices of (λ, γ, β), which can all be regarded as tuning parameters. To address this issue, we develop two empirical Bayes methods, extending the minimax Bayes method of Tan (2015). These methods are called Steinized empirical Bayes estimation, in parallel to the interpretation of $\delta_{\lambda,\gamma,\beta}$ as a Steinized Bayes estimator.

The first empirical Bayes method is to directly apply the estimator $\delta_{\lambda,\gamma,\delta}$, but with $(\lambda, \delta, \gamma)$ chosen in a data-dependent manner, using a combination of Stein's (1981) unbiased risk estimation and Fay and Herriot's (1979) model-based estimation. The proposed method consists of the following steps:

- Estimate (γ, β) as $(\hat{\gamma}, \hat{\beta})$ by applying the method of Fay and Herriot (1979) for the random-effect model (1.3), represented by (1.1) and (1.2);
- Estimate λ as $\hat{\lambda} = \hat{\lambda}(\hat{\gamma}, \hat{\beta})$ by minimizing Stein's (1981) unbiased risk estimate of $\delta_{\lambda,\hat{\gamma},\hat{\beta}}$ but with $(\hat{\gamma}, \hat{\beta})$ treated as non-random;
- Apply the estimator $\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}$.

If shrinkage is desired particularly toward 0, then the corresponding estimator of θ is defined as $\delta_{\hat{\lambda}_0,\hat{\gamma}_0} = \delta_{\hat{\lambda}_0,\hat{\gamma}_0,\beta=0}$, where $(\hat{\lambda}_0, \hat{\gamma}_0)$ are obtained similarly as $(\hat{\lambda}, \hat{\gamma}, \hat{\beta})$ above but with $\beta = 0$ fixed and hence $\hat{\beta}$ no longer needed.

The second method is developed by extending the minimax Bayes method of Tan (2015) to handle shrinkage toward a linear subspace (e.g., Sclove, Morris, and Radhakrishnan (1972); Oman (1982)). This method consists of the following steps:

- Estimate β by $\tilde{\beta} = (X^{\mathrm{T}} D^{-1} X)^{-1} (X^{\mathrm{T}} D^{-1} Y)$;
- Apply the estimator $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}$ after linearly transforming the residual $Y - X\tilde{\beta}$ into a canonical form with a diagonal variance matrix.

The second empirical Bayes method differs from the first one mainly in how the regression coefficient vector β is estimated to determine the center of shrinkage and in how the residual$Y - X\hat{\beta}$ or $Y - X\tilde{\beta}$ is shrunk.

3.1. SURE tuning

To empirically choose (λ, γ, β) for $\delta_{\lambda,\gamma,\beta}$, a possible strategy is to use Stein's (1981) unbiased risk estimate (SURE). Rewrite $\delta_{\lambda,\gamma,\beta}$ componentwise as

$$(\delta_{\lambda,\gamma,\beta})_j = x_j^{\mathrm{T}}\beta + \{1 - \lambda b_j(\gamma, \beta)\}_+ (Y_j - x_j^{\mathrm{T}}\beta),$$

where $b_j(\gamma, \beta) = c(\gamma) a_j(\gamma) / \{\sum_{k=1}^n a_k^2(\gamma)(Y_k - x_k^{\mathrm{T}}\beta)^2\}$ and $c(\gamma) = \sum_{k=1}^n a_k^2(\gamma)(d_k + \gamma)$. By the SURE formula, define

$$\begin{aligned}\mathrm{SURE}(\delta_{\lambda,\gamma,\beta}) = \sum_{j=1}^n &+ \sum_{j \notin J_{\lambda,\gamma,\beta}} \{(Y_j - x_j^{\mathrm{T}}\beta)^2 - 2d_j\} \\ &+ \sum_{j \in J_{\lambda,\gamma,\beta}} \Big\{\lambda^2 b_j^2(\gamma, \beta)(Y_j - x_j^{\mathrm{T}}\beta)^2 - 2\lambda d_j b_j(\gamma, \beta) \\ &+ \frac{4\lambda d_j b_j(\gamma, \beta) a_j^2(\gamma)(Y_j - x_j^{\mathrm{T}}\beta)^2}{\sum_{k=1}^n a_k^2(\gamma)(Y_k - x_k^{\mathrm{T}}\beta)^2}\Big\},\end{aligned}$$

where $J_{\lambda,\gamma,\beta} = \{1 \le j \le n : \lambda b_j(\gamma, \beta) < 1\}$. Then $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})$ is an unbiased estimator of the risk of $\delta_{\lambda,\gamma,\beta}$, that is, $E_\theta\{\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})\} = R(\delta_{\lambda,\gamma,\beta}, \theta)$ for any $\theta \in \mathbb{R}^n$. In principle, (λ, γ, β) can be selected by directly minimizing $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})$

over $0 \le \lambda \le 2$, $\gamma \ge 0$, and $\beta \in \mathbb{R}^q$. But there seems to be several difficulties in this direct approach, as we realized from pilot simulation studies.

First, $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})$ is, in general, a non-smooth, multi-modal function of (λ, γ, β), and β can be multi-dimensional. No reliable algorithm seems available for global optimization of such a complicated function. Second, even when β is one-dimensional and $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})$ is minimized in a nested manner, for example, by solving $\min_{\gamma\ge 0}[\min_{\beta\in\mathbb{R}}\{\min_{0\le\lambda\le 2}\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})\}]$, the computational challenge remains, because most one-dimensional optimization algorithms are still designed to find a local, not global, minimizer. Finally, from simulation examples, we found that the values of (γ, β) minimizing $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})$, if identified, might not reflect how the second-level model (1.2) could be properly fitted to the true values θ_j. This phenomenon seems to also occur for other SURE-based methods (e.g., Jiang, Nguyen, and Rao (2011); Xie, Kou, and Brown (2012)) where SURE is minimized to empirically select both the center and magnitude of shrinkage. See the Supplementary Material for further discussion.

In view of the foregoing issues, we use a SURE-based strategy only for choosing λ and, to be discussed in Section 3.2, adopt a model-based approach for choosing (γ, β). For any choices of (γ, β), take

$$\hat{\lambda}(\gamma, \beta) = \operatorname*{argmin}_{0\le\lambda\le 2} \mathrm{SURE}(\delta_{\lambda,\gamma,\beta}).$$

The optimization problem can be easily solved in a non-iterative manner as follows. Sort the values $\sum_{k=1}^n a_k^2(\gamma)(Y_k - x_k^{\mathrm{T}}\beta)^2/\{c(\gamma)a_j(\gamma)\}$ for $j = 1, \ldots, n$, and partition the interval [0, 2] at those values. Then $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})$ is a quadratic function of λ in each subinterval. Therefore, we determine $\hat{\lambda}(\gamma, \beta)$ as a minimizer of $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})$ in a subinterval such that the corresponding minimum value is the smallest among all the minimum values from different subintervals.

3.2. Model-based residual shrinkage

As explained in Section 3.1, we choose (γ, β) for $\delta_{\lambda,\gamma,\beta}$ using a model-based approach as in conventional empirical Bayes estimation. Several methods have been proposed for estimating (γ, β) in the random-effect model (1.3) in the context of small-area estimation, including the maximum likelihood method, the restricted maximum likelihood method, Fay and Herriot's (1979) method based on weighted residuals, and Prasad and Rao's (1990) method based on unweighted residuals. For various reasons, we adopt Fay and Herriot's (1979) estimators of (γ, β), to be used in $\delta_{\lambda,\gamma,\beta}$.

For any fixed $\gamma \ge 0$, the generalized least squares estimator of β under (1.3) is

$$\hat{\beta}(\gamma) = (X^{\mathrm{T}} D_\gamma^{-1} X)^{-1}(X^{\mathrm{T}} D_\gamma^{-1} Y), \tag{3.1}$$

where $D_\gamma = D + \gamma I$. Then Fay and Herriot's (1979) estimators of γ and β are respectively $\hat{\gamma}_{\mathrm{FH}}$ and $\hat{\beta}_{\mathrm{FH}} = \hat{\beta}(\hat{\gamma}_{\mathrm{FH}})$, where the moment equation for $\hat{\gamma}_{\mathrm{FH}}$ is

$$\sum_{j=1}^{n} \frac{\{Y_j - x_j^{\mathrm{T}} \hat{\beta}(\gamma)\}^2}{d_j + \gamma} = n - q. \tag{3.2}$$

As noted by Datta and Ghosh (2012), the left-hand side of (3.2) is $\min_{\beta \in \mathbb{R}^q} \{\sum_{j=1}^{n} (Y_j - x_j^{\mathrm{T}} \beta)^2 / (d_j + \gamma)\}$ and hence non-increasing in $\gamma \geq 0$. If the left-hand side at $\gamma = 0$ is smaller than $n - q$, then there is no solution to (3.2) and $\hat{\gamma}_{\mathrm{FH}}$ is defined as 0. Otherwise, $\hat{\gamma}_{\mathrm{FH}}$ is a unique solution to (3.2) and can be computed without ambiguity by a root-finding algorithm such as `uniroot()` in R (R Core Team (2012)). In addition to computational simplicity, the monotonicity associated with (3.2) is also technically useful to asymptotic theory for the resulting estimator of θ.

By comparison, the maximum likelihood estimators of γ and β under (1.3) are, respectively, $\hat{\gamma}_{\mathrm{ML}}$ and $\hat{\beta}_{\mathrm{ML}} = \hat{\beta}(\hat{\gamma}_{\mathrm{ML}})$, where the score equation for $\hat{\gamma}_{\mathrm{ML}}$ is

$$\sum_{j=1}^{n} \frac{\{Y_j - x_j^{\mathrm{T}} \hat{\beta}(\gamma)\}^2}{(d_j + \gamma)^2} = \sum_{j=1}^{n} \frac{1}{d_j + \gamma}. \tag{3.3}$$

In contrast with equation (3.2), the difference and the ratio of the two sides of (3.3) may not be monotonic in $\gamma \geq 0$. Moreover, it is possible that the left-hand side of (3.3) is strictly less than the right-hand side at $\gamma = 0$, but there exist multiple solutions to (3.3). See the Supplementary Material for a numerical example.

Fay and Herriot (1979) provided a thoughtful discussion, modified in our notation as follows, on the comparison between (3.2) and (3.3). Equations (3.2) and (3.3) weight the significance of the deviations $Y_j - x_j^{\mathrm{T}} \hat{\beta}(\gamma)$ differently: (3.3) places relatively more importance on the observations with small d_j than does (3.2). The maximum likelihood method improves the efficiency of the estimation of γ in the random-effect model (1.3). But it was preferred to balance out the estimation of γ over all observations rather than of just those with small d_j. From this discussion, the choice of (3.2) over (3.3) for estimating γ would be even more justified in our approach than in Fay and Herriot's, because we explicitly allow the second-level model (1.2) to be misspecified and substitute the estimators of (γ, β) in $\delta_{\lambda,\gamma,\beta}$ instead of the Bayes rule $\delta^{\mathrm{B}}_{\gamma,\beta}$ based on (1.2). For our approach, $(\hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}})$ are used to approximately capture the center and spread of the true values $(\theta_1, \ldots, \theta_n)$, so that effective risk reduction can be achieved by the minimax Bayes estimator $\delta_{\lambda,\gamma,\beta}$.

By substituting $(\hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}})$ for (γ, β) and setting $\hat{\lambda}_{\mathrm{FH}} = \hat{\lambda}(\hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}})$ for λ in $\delta_{\lambda,\gamma,\beta}$, we define a model-based residual shrinkage estimator of θ as

$$\delta_{\mathrm{Res}} = \delta_{\hat{\lambda}_{\mathrm{FH}},\hat{\gamma}_{\mathrm{FH}},\hat{\beta}_{\mathrm{FH}}}.$$

Effectively, δ_{Res} is obtained by taking $(\hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}})$ from Fay and Herriot's (1979) estimation of random-effect model (1.3) and then feeding the residual $Y_j - x_j^{\mathrm{T}}\hat{\beta}_{\mathrm{FH}}$ to Tan's (2015) shrinkage estimator toward 0 with the prior variance $\gamma_j \equiv \hat{\gamma}_{\mathrm{FH}}$. By comparison, Fay and Herriot's (1979) empirical Bayes estimator of θ is

$$\delta_{\mathrm{FH}} = \delta^{\mathrm{B}}_{\hat{\gamma}_{\mathrm{FH}},\hat{\beta}_{\mathrm{FH}}},$$

defined by substituting $(\hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}})$ for (γ, β) in the Bayes rule $\delta^{\mathrm{B}}_{\lambda,\beta}$. As discussed in Section 1, δ_{FH} is fully based on random-effect model (1.3), and its performance may substantially deteriorate when second-level model (1.3) is misspecified.

It is helpful to compare both δ_{FH} and δ_{Res} with the following estimator of θ in Jiang, Nguyen, and Rao (2011) and Xie, Kou, and Brown (2012):

$$\delta_{\mathrm{JX}} = \delta^{\mathrm{B}}_{\hat{\gamma}_{\mathrm{JX}},\hat{\beta}_{\mathrm{JX}}},$$

where $(\hat{\gamma}_{\mathrm{JX}}, \hat{\beta}_{\mathrm{JX}}) = \mathrm{argmin}_{\gamma\geq 0,\beta\in\mathbb{R}^q}\mathrm{SURE}(\delta^{\mathrm{B}}_{\gamma,\beta})$ and $\mathrm{SURE}(\delta^{\mathrm{B}}_{\gamma,\beta})$ is Stein's unbiased risk estimate of $\delta^{\mathrm{B}}_{\gamma,\beta}$ defined by

$$\mathrm{SURE}(\delta^{\mathrm{B}}_{\gamma,\beta}) = \sum_{j=1}^{n}\Big\{\frac{d_j^2(Y_j - x_j^{\mathrm{T}}\beta)^2}{(d_j+\gamma)^2} + \frac{2\gamma d_j}{d_j+\gamma} - d_j\Big\}.$$

Setting the gradient to 0 shows that $\hat{\beta}_{\mathrm{JX}} = \bar{\beta}(\hat{\gamma}_{\mathrm{JX}})$ with

$$\bar{\beta}(\gamma) = \Big\{\sum_{j=1}^{n}\frac{d_j^2 x_j x_j^{\mathrm{T}}}{(d_j+\gamma)^2}\Big\}^{-1}\Big\{\sum_{j=1}^{n}\frac{d_j^2 x_j Y_j}{(d_j+\gamma)^2}\Big\}, \tag{3.4}$$

and $\hat{\gamma}_{\mathrm{JX}}$, if nonzero, satisfies

$$\sum_{j=1}^{n}\frac{d_j^2\{Y_j - x_j^{\mathrm{T}}\hat{\beta}(\gamma)\}^2}{(d_j+\gamma)^3} = \sum_{j=1}^{n}\frac{d_j^2}{(d_j+\gamma)^2}. \tag{3.5}$$

In contrast with (3.2), there is, in general, no monotonicity associated with equation (3.5). In fact, there may exist multiple solutions to (3.5), corresponding to a local minimizer or a local maximizer of $\mathrm{SURE}(\delta^{\mathrm{B}}_{\gamma,\beta})$. In our numerical work, $\hat{\gamma}_{\mathrm{JX}}$ is computed as a local, perhaps non-global, minimizer to $\mathrm{SURE}\{\delta^{\mathrm{B}}_{\gamma,\bar{\beta}(\gamma)}\}$ by the one-dimensional optimization algorithm `optimize()` in R (R Core Team (2012)). See the Supplementary Material for numerical examples and further discussion.

In retrospect, $\bar{\beta}(\gamma)$ in (3.4) can be seen as a weighted least squares estimator of β for a fixed $\gamma \geq 0$, and equation (3.5) can be verified as a moment equation, based on the marginal distribution $Y_j \sim \mathrm{N}(x_j^{\mathrm{T}}\beta, d_j + \gamma)$. However, the observations and residuals are weighted directly in proportion to the variances d_j in (3.4) and (3.5), and are weighted inversely in proportion to the variances d_j in previous equations (3.1), (3.2) and (3.3). Moreover, the observations are shrunk by magnitudes proportional to the variances d_j toward the regression line $x_j^{\mathrm{T}}\hat{\beta}_{\mathrm{JX}}$ in the estimator δ_{JX}. Therefore, δ_{JX} may be overly influenced by the observations with large variances. Essentially, these patterns of weighting and shrinkage in δ_{JX} are tied to the fact that δ_{JX} is constructed in the form of the Bayes rule $\delta^{\mathrm{B}}_{\gamma,\beta}$ based on (1.2). The performance of δ_{JX} may still deteriorate when second-level model (1.2) is misspecified.

3.3. Subspace shrinkage

The estimator δ_{Res} is constructed to shrink the observation vector Y toward a data-dependent location $X\hat{\gamma}_{\mathrm{FH}}$, so that substantial risk reduction can be achieved whenever the true mean vector θ lies near the linear subspace

$$\mathbb{S} = \{\theta \in \mathbb{R}^n : \theta = X\beta \text{ for } \beta \in \mathbb{R}^2\}.$$

In this section, we develop an alternative method from the perspective of subspace shrinkage (e.g., Sclove, Morris, and Radhakrishnan (1972); Oman (1982)). This method involves the same idea of shrinking Y toward a data-dependent location $X\tilde{\beta}$, but deals with the data-dependency of $\tilde{\beta}$ explicitly in two ways. First, $\tilde{\beta}$ is defined such that $X\tilde{\beta}$ and $Y - X\tilde{\beta}$ are uncorrelated. Second, the residual $Y - X\tilde{\beta}$ is shrunk toward 0 by taking account of the variance matrix of $Y - X\tilde{\beta}$, which is no longer D or even a diagonal matrix. In contrast, $\hat{\beta}_{\mathrm{FH}}$ is derived together with $\hat{\gamma}_{\mathrm{FH}}$ from the random-effect model (1.3), and then $Y - X\hat{\beta}_{\mathrm{FH}}$ is shrunk toward 0 in δ_{Res} as if $\hat{\beta}_{\mathrm{FH}}$ were data-independent and the variance matrix of $Y - X\hat{\beta}_{\mathrm{FH}}$ were still D.

Shrinkage toward a linear subspace has been of interest since the early research on shrinkage estimation (e.g., Lindley (1962)). In the homoscedastic case ($D = \sigma^2 I$), a known extension of the James–Stein estimator is

$$\delta_{\mathrm{JS,Sub}} = HY + \left\{1 - \frac{(n-q-2)\sigma^2}{Y^{\mathrm{T}}(I-H)Y}\right\}_+ (I-H)Y,$$

where $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. The estimator $\delta_{\mathrm{JS,Sub}}$ is minimax if $n \geq q+3$ (Sclove, Morris, and Radhakrishnan (1972)). Taking $X = (1, \ldots, 1)^{\mathrm{T}}$ leads to the shrinkage estimator toward the mean $\bar{Y} = n^{-1}\sum_{j=1}^n Y_j$ (Efron and Morris (1973)). However, there seems to be limited results on shrinkage estimation

toward a subspace in the general heteroscedastic case. See Oman (1982) on an extension of Bock's (1975) estimator and Judge and Mittelhammer (2004) on a shrinkage estimator that combines two competing estimators $\hat{\theta}$ and $\hat{\psi}$, which can be viewed as shrinkage toward the subspace $\{(\theta, \psi) : \theta = \psi\}$. These estimators are only shown to be minimax under some restrictive conditions.

We extend the minimax Bayes method of Tan (2015) to handle shrinkage toward a linear subspace. By generalized least squares, let $\tilde{\beta} = (X^{\mathrm{T}} D^{-1} X)^{-1} (X^{\mathrm{T}} D^{-1} Y)$. Then $X\tilde{\beta}$ and $Y - X\tilde{\beta}$ are uncorrelated, and the variance matrix of $Y - X\tilde{\beta}$ is $D - X(X^{\mathrm{T}} D^{-1} X)^{-1} X^{\mathrm{T}}$, with rank $n - q$, under the basic model (1.1). The estimate $\tilde{\beta}$ is formally identical to $\hat{\beta}(0)$ with $\gamma = 0$ by (3.1). But, in contrast with the method in Section 3.2, we apply shrinkage estimation to the residual $Y - X\tilde{\beta}$ by taking account of the variability of $\tilde{\beta}$ under model (1.1) for achieving minimaxity.

Decompose the observation vector Y and mean vector θ as $Y = X\tilde{\beta} + (Y - X\tilde{\beta}) = H_D Y + (I - H_D)Y$ and $\theta = H_D\theta + (I - H_D)\theta$, where $H_D = X(X^{\mathrm{T}} D^{-1} X)^{-1} X^{\mathrm{T}} D^{-1}$ and hence $H_D\theta \in \mathbb{S}$. If θ lies near the subspace S, then $H_D\theta \approx \theta$ and $(I - H_D)\theta \approx 0$. To achieve risk reduction, we estimate $H_D\theta$ directly by $X\tilde{\beta} = H_D Y$, and then shrink the residual $Y - X\tilde{\beta}$ toward 0 for estimating $(I - H_D)\theta$ as follows. We linearly transform the residual $Y - X\tilde{\beta}$ into a canonical form, $L_2^{\mathrm{T}}(Y - X\tilde{\beta})$, with a diagonal variance matrix V_2, by the spectral decomposition:

$$D - X(X^{\mathrm{T}} D^{-1} X)^{-1} X^{\mathrm{T}} = (L_1, L_2) \begin{pmatrix} 0 & 0 \\ 0 & V_2 \end{pmatrix} \begin{pmatrix} L_1^{\mathrm{T}} \\ L_2^{\mathrm{T}} \end{pmatrix} = L_2 V_2 L_2^{\mathrm{T}},$$

where V_2 is a positive definite, diagonal $(n - q) \times (n - q)$ matrix and (L_1, L_2) is an orthogonal $n \times n$ matrix with L_2 an $n \times (n - q)$ matrix. Applying the method of Tan (2015) leads to the following estimator of θ:

$$\delta^{\mathrm{S}}_{\lambda,\gamma} = X\tilde{\beta} + L_2 \delta_{\lambda,\gamma,\beta=0}\{L_2^{\mathrm{T}}(Y - X\tilde{\beta})\},$$

where $\delta_{\lambda,\gamma,\beta=0}(\eta_2)$ denotes $\delta_{\lambda,\gamma,\beta=0}$ from Section 2 applied with Y replaced by $\eta_2 = L_2^{\mathrm{T}}(Y - X\tilde{\beta})$ and D replaced by V_2 for estimating $\psi_2 = L_2^{\mathrm{T}}(I - H_D)\theta$, of dimension $n - q$. If $D = \sigma^2 I$, then $D - X(X^{\mathrm{T}} D^{-1} X)^{-1} X^{\mathrm{T}}$ is an idempotent matrix, $I - X(X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}}$, multiplied by σ^2 and hence V_2 is proportional to the $(n - q) \times (n - q)$ identity matrix. In this special case, $\delta^{\mathrm{S}}_{\lambda=1,\gamma}$ reduces to the subspace shrinkage estimator $\delta_{\mathrm{JX,Sub}}$, because, as noted in Tan (2015), $\delta_{\lambda=1,\lambda,\beta=0}$ with any $\gamma \geq 0$ reduces to the James–Stein positive-part estimator, $\{1 - (n - 2)\sigma^2 / \sum_{k=1}^n Y_k^2\}_+ Y_j$, in the homoscedastic case. The following result is proved in the Supplementary Material.

Theorem 1. *The pointwise risk of $\delta^S_{\lambda,\gamma}$ for estimating θ is related to that of $\delta_{\lambda,\gamma,\beta=0}(\eta_2)$ for estimating ψ_2 as follows:*

$$R(\delta^S_{\lambda,\gamma},\theta) = \mathrm{tr}\{X(X^{\mathrm{T}}D^{-1}X)^{-1}X^{\mathrm{T}}\} + R\{\delta_{\lambda,\gamma,\beta=0}(\eta_2),\psi_2\},$$

where $\eta_2 = L_2^{\mathrm{T}}(Y - X\tilde{\beta})$, $\psi_2 = L_2^{\mathrm{T}}(I - H_D)\theta)$, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. For $n \geq q+3$, the estimator $\delta^S_{\lambda,\gamma}$ is minimax over $\theta \in \mathbb{R}^n$ for any fixed $0 \leq \lambda \leq 2$ and $\gamma \geq 0$. Moreover, a similar result to (2.7) *holds for the Bayes risk of $\delta^S_{\lambda=1,\gamma}$.*

The estimator $\delta^{\mathrm{S}}_{\lambda,\gamma}$ incorporates a data-dependent choice, $X\tilde{\beta}$, for the center of shrinkage. But there remain two tuning parameters, λ and γ, inherited from $\delta_{\lambda,\gamma,\beta=0}$ for the shrinkage of the residual, $Y - X\tilde{\beta}$. To empirically choose (λ,γ) for $\delta_{\lambda,\gamma,\beta=0}$, we adopt the following strategy in parallel to that of selecting (λ,γ,β) for $\delta_{\lambda,\gamma,\beta}$ in Sections 3.1 and 3.2, except with $\beta = 0$ fixed. We define an empirical Bayes extension of $\delta_{\lambda,\gamma,\beta=0}$ for shrinkage toward 0 with the observation vector Y as $\delta_{\hat{\lambda}_0,\hat{\gamma}_0} = \delta_{\hat{\lambda}_0,\hat{\gamma}_0,\beta=0}$ where $\hat{\lambda}_0 = \hat{\lambda}(\hat{\gamma}_0)$ with $\hat{\lambda}(\gamma) = \mathrm{argmin}_{0\leq\lambda\leq 2}\mathrm{SURE}(\delta_{\lambda,\gamma,\beta=0})$ and $\hat{\gamma}_0$ is a solution to

$$\sum_{j=1}^{n} \frac{Y_j^2}{(d_j+\gamma)} = n, \tag{3.6}$$

if $\sum_{j=1}^n Y_j^2/d_j \geq n$ and $\hat{\gamma}_0 = 0$ otherwise. The estimator $\hat{\gamma}_0$ is directly modiifed from Fay and Herriot's estimator $\hat{\gamma}_{\mathrm{FH}}$ to handle a fixed $\beta = 0$. Then we define a subspace shrinkage estimator of θ as

$$\delta_{\mathrm{Sub}} = X\tilde{\beta} + L_2\delta_{\hat{\lambda}_0,\hat{\gamma}_0}\{L_2^{\mathrm{T}}(Y - X\tilde{\beta})\},$$

where $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}(\eta_2)$ denotes $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}$ applied with Y replaced by $\eta_2 = L_2^{\mathrm{T}}(Y - X\tilde{\beta})$ and D replaced by V_2 for estimating $\psi_2 = L_2^{\mathrm{T}}(I - H_D)\theta$. Effectively, δ_{Sub} is obtained by feeding the linearly transformed residual, $L_2^{\mathrm{T}}(Y - X\tilde{\beta})$, to Tan's (2015) shrinkage estimator $\delta_{\lambda,\gamma,\beta=0}$ toward 0, with data-dependent choices of γ and λ by, respectively, Fay and Herriot's (1979) method and a SURE strategy.

The estimators $(\hat{\lambda}_0,\hat{\gamma}_0)$ depend on Y only through the residual $Y - X\tilde{\beta}$ and hence are independent of $X\tilde{\beta}$. Then the proof of Theorem 1 also shows that

$$R(\delta_{\mathrm{Sub}},\theta) = \mathrm{tr}\{X(X^{\mathrm{T}}D^{-1}X)^{-1}X^{\mathrm{T}}\} + R\{\delta_{\hat{\lambda}_0,\hat{\gamma}_0}(\eta_2),\psi_2\}. \tag{3.7}$$

Therefore, the risk difference between δ_{Sub} and Y for estimating θ is the same as that between $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}(\eta_2)$ and η_2 for estimating ψ_2.

4. Asymptotic Theory

In this section, we study pointwise risk properties of the proposed estimators in an asymptotic framework where $n \to \infty$ and q is fixed, under the basic model

(1.1), without requiring the second-level model (1.2) to be correct. The results obtained are distinct from existing ones where random-effect model (1.3) (i.e., both (1.1) and (1.2)) is assumed to be correct (e.g., Prasad and Rao (1990)). Moreover, our asymptotic analysis allows that the maximum ratio between $(d_1, \ldots, d_n)$ be unbounded as $n \to \infty$.

For the model-based residual shrinkage estimator δ_{Res} in Section 3.2, we will show that uniformly in θ over any bounded subset of $\mathbb{R}^n$,

$$n^{-1}R(\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}, \theta) \le \inf_{0\le\lambda\le 2} n^{-1}R(\delta_{\lambda,\gamma^*,\beta^*}, \theta) + o(1), \tag{4.1}$$

where $(\hat{\lambda}, \hat{\gamma}, \hat{\beta}) = (\hat{\lambda}_{\mathrm{FH}}, \hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}})$ and (γ^*, β^*) are the probability limits of $(\hat{\gamma}, \hat{\beta})$ as $n \to \infty$, unless otherwise stated. That is, the normalized risk of δ_{Res} is asymptotically no greater than $\delta_{\lambda,\gamma^*,\beta^*}$ for any fixed choice $0 \le \lambda \le 2$ and for the limit values (γ^*, β^*) of Fay and Herriot's (1979) estimators. Therefore, asymptotic optimality is achieved over possible choices of $0 \le \lambda \le 2$, but not over (γ, β). This difference is caused by the fact that a SURE-based strategy is used to define $\hat{\lambda}$, but a model-based strategy is used to construct $(\hat{\gamma}, \hat{\beta})$, due to various difficulties that would arise were a SURE-based strategy used for selecting $(\hat{\gamma}, \hat{\beta})$, as discussed in Section 3.1. Moreover, the lack of theoretical optimality over (γ, β) should be of limited concern for the following reasons. For any fixed choices of $0 \le \lambda \le 2$, $\gamma \ge 0$, and $\beta \in \mathbb{R}^q$, the estimator $\delta_{\lambda,\gamma,\beta}$ is already shown to be minimax over $\theta \in \mathbb{R}^n$ and, if $\lambda = 1$, achieves close to the minimum Bayes risk over a scale class of normal priors, depending on (γ, β). The choices of $(\hat{\gamma}, \hat{\beta})$ are mainly used to determine a particular prior (1.2) capturing the center and spread of the true values $(\theta_1, \ldots, \theta_n)$, so that effective risk reduction can be achieved due to near-Bayes optimality of $\delta_{\lambda,\gamma,\beta}$ under the prior (1.2).

As a consequence of (4.1), Theorem 2 provides a definite comparison of asymptotic risks between $\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}$ and δ_{FH}. Up to negligible terms, the asymptotic risk of $\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}$ is no greater than $\sum_{j=1}^n d_j$, the risk of $\delta_0 = Y$, and than that of δ_{FH}.

Theorem 2. *Assume* $(\max_j d_j)/(\min_j d_j) = o(n^{1/2})$. *For any sequence of* $\theta \in \mathbb{R}^n$ *such that* (4.1) *holds and* $n^{-1}R(\delta_{FH}, \theta) \le n^{-1}R(\delta^B_{\gamma^*,\beta^*}, \theta) + o(1)$ *as* $n \to \infty$, *we have*

$$n^{-1}R(\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}, \theta) \le n^{-1}R(\delta_{FH}, \theta) - \frac{(S_1 - S_2)^2}{nS_2} + \Big(1 + n^{-1}\sum_{j=1}^n d_j\Big)o(1),$$

where $S_1 = \sum_{j=1}^n d_j a_j^*$, $S_2 = \sum_{j=1}^n \{d_j + (\theta_j - x_j^{\mathrm{T}}\beta^*)^2\}a_j^{*2}$, *and* $a_j^* = d_j/(d_j + \gamma^*)$ *for* $j = 1, \ldots, n$.

Comparison of asymptotic risks between $\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}$ and δ_{JX} seems complicated, mainly because (γ,β) are estimated differently by $(\hat{\gamma}_{\mathrm{FH}},\hat{\beta}_{\mathrm{FH}})$ and $(\hat{\gamma}_{\mathrm{JX}},\hat{\beta}_{\mathrm{JX}})$. Nevertheless, if $(\hat{\gamma}_{\mathrm{JX}},\hat{\beta}_{\mathrm{JX}})$ are used as $(\hat{\gamma},\hat{\beta})$ in $\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}$, then, by the proof of Theorem 2, it can also be shown that $n^{-1}R(\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}},\theta)\le n^{-1}R(\delta_{\mathrm{JX}},\theta)+o(1)(1+n^{-1}\sum_{j=1}^n d_j)$.

The subspace shrinkage estimator δ_{Sub} in Section 3.3 involves the use of $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}=\delta_{\hat{\lambda}_0,\hat{\gamma}_0,\beta=0}$ for shrinkage toward 0. To assess pointwise risk properties of δ_{Sub}, it suffices, by the relationship (3.7), to study those of $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}$ in the general setup with the observation vector Y. In parallel to the result (4.1), it is of interest to show that uniformly in θ over any bounded subset of $\mathbb{R}^n$,

$$n^{-1}R(\delta_{\hat{\lambda}_0,\hat{\gamma}_0},\theta)\le \inf_{0\le\lambda\le 2} n^{-1}R(\delta_{\lambda,\delta_0^*,\beta=0},\theta)+o(1), \tag{4.2}$$

where γ_0^* is the probability limit of $\hat{\gamma}_0$ as $n\to\infty$. The implications of (4.2) can be seen similarly as in the preceding discussion on (4.1).

4.1. Convergence results

We divide the results into three parts, dealing with three distinct situations: shrinkage estimation with a data-independent prior, shrinkage estimation toward 0 (or a data-independent location) with a data-dependent prior variance, and shrinkage estimation with a data-dependent prior mean and variance. Moreover, we remove the normality assumption in (1.1), i.e., only assume that $E_{\theta_j}(Y_j)=\theta_j$ for $j=1,\ldots,n$, in all our asymptotic results throughout this section.

First, we examine the situation where λ is estimated by a SURE-based strategy, but for the estimator $\delta_{A,\lambda}$ with a general, data-independent choice of $A=\mathrm{diag}(a_1,\ldots,a_n)$. For the choice $(a_1^\dagger,\ldots,a_n^\dagger)$ in (2.2) and (2.3), the prior variances $(\gamma_1,\ldots,\gamma_n)$ need not be a constant γ as in the estimator $\delta_{\lambda,\gamma,\beta}$. More generally, $(a_1,\ldots,a_n)$ are allowed to differ from the particular choice $(a_1^\dagger,\ldots,a_n^\dagger)$. The following theorem shows that the SURE of $\delta_{A,\lambda}$ is, uniformly over $0\le\lambda\le 2$, close to the actual loss in probability and in mean, uniformly in $\theta\in\mathbb{R}^n$. See Li (1985), Donoho and Johnstone (1995), and Cai and Zhou (2009) for related results of this nature. By comparison, Xie, Kou, and Brown (2012) showed that $\mathrm{SURE}(\delta^{\mathrm{B}}_{\gamma,\beta=0})$ is, uniformly over $\gamma\ge 0$, close to the actual loss of $\delta^{\mathrm{B}}_{\gamma,\beta=0}$, provided that both $n^{-1}\sum_{j=1}^n d_j^2$ and $n^{-1}\sum_{j=1}^n d_j\theta_j^2$ are bounded for all $n\ge 1$. Such a convergence is not uniform in $\theta\in\mathbb{R}^n$.

Theorem 3. *Assume that there exist some constants K_1, K_2, K_3, K_4, and $0\le\eta\le 1/4$, such that for all $n\ge 1$,*

(A1) $E(\varepsilon_j^4)\le K_1 d_j^2$, *with* $\varepsilon_j=Y_j-\theta_j$, *for* $j=1,\ldots,n$,

(A2) $n^{-1}\sum_{j=1}^n d_j\le K_2$,

(A3) $(\max_j d_j)/(\min_j d_j) \le K_3 n^{\eta}$, *and*

(A4) $(\max_j a_j)/(\min_j a_j) \le K_4(\max_j d_j)/(\min_j d_j)$.

Let $\zeta_n(\lambda) = n^{-1}\{SURE(\delta_{A,\lambda}) - L(\delta_{A,\lambda}, \theta)\}$, *where* $SURE(\delta_{A,\lambda})$ *is defined as* $SURE(\delta_{\lambda,\gamma,\beta})$ *with* $\beta = 0$ *and* $a_j(\gamma)$ *replaced by* a_j. *Then the following results hold.*

(i) $\sup_{0\le\lambda\le 2} |\zeta_n(\lambda)| = O_p\{n^{-(1-4\eta)/2}\}$, *uniformly in* $\theta \in \mathbb{R}^n$. *That is, for any* $\tau_1 > 0$, $\sup_{\theta\in\mathbb{R}^n} P\{\sup_{0\le\lambda\le 2} |\zeta_n(\lambda)| \ge \tau_2 n^{-(1-4\eta)/2}\} \le \tau_1$ *for all sufficiently large* τ_2 *and* n.

(ii) *If further* $0 \le \eta < 1/5$, *then* $\sup_{\theta\in\mathbb{R}^n} E\{\sup_{0\le\lambda\le 2} |\zeta_n(\lambda)|\} = O\{n^{-(1-5\eta)/2}\}$.

Assumption (A1) requires fourth moments to be bounded for all ε_j, which are directly satisfied when the disturbances ε_j are normal. Assumptions (A2) and (A3) place restrictions on both the absolute and relative magnitudes of the variances $(d_1, \ldots, d_n)$, which are needed mainly to ensure the uniform convergence in $0 \le \lambda \le 2$. To understand the implication of (A2) and (A3), suppose, for the moment, that $(d_1, \ldots, d_n)$ are independent realizations from a certain distribution with density $\varphi(\cdot)$ such that $\varphi(d) = O\{d^{-(1+k_1)}\}$ as $d \to \infty$ and $O\{d^{-(1-k_0)}\}$ as $d \to 0$ for $k_1 > 0$ and $k_0 > 0$. By extreme value theory (e.g., Ferguson (1996)), it is easy to show that $\max_j d_j = O_p(n^{1/k_1})$ and $\min_j d_j = O_p(n^{-1/k_0})$ as $n \to \infty$. If $1/k_1 + 1/k_0 < 1/4$, then Assumptions (A2) and (A3) would be satisfied in probability as $n \to \infty$. For example, when $(d_1, \ldots, d_n)$ are inverse chi-squared distributed with degrees of freedom k, then k_1 can be set to $k/2$ but k_0 can be made arbitrarily large. Therefore, if $k > 8$, then Assumptions (A2) and (A3) would hold with probability tending to 1. Finally, Assumption (A4) places an upper bound on the relative magnitudes of shrinkage, a_j/a_k, in terms of the maximum variance ratio $(\max_j d_j)/(\min_j d_j)$. If $(a_1, \ldots, a_n)$ are defined by (2.5) and (2.6) with a homoscedastic prior variance $\Gamma = \gamma I$ as in the estimator $\delta_{\lambda,\gamma,\beta}$, then Assumption (A4) is satisfied with $K_4 = 1$ for any $\gamma \ge 0$ by Lemma 1.

Second, we deal with the shrinkage estimator toward 0, $\delta_{\hat\lambda_0,\hat\gamma_0} = \delta_{\hat\lambda_0,\hat\lambda_0,\beta=0}$, that is, $\delta_{\lambda,\gamma,\beta=0}$ with a Fay–Herriot type estimator $\hat\gamma_0$ for γ, leading to data-dependent $a_j(\hat\gamma_0)$, and a SURE-based choice $\hat\lambda_0$ for λ. By moment equation (3.6) for $\hat\gamma_0$, let $\gamma_0^* \ge 0$ be a solution to $n = \sum_{j=1}^n (d_j + \theta_j^2)/(d_j + \gamma)$. Such a solution γ_0^* always exists and is unique. The following proposition shows that $\hat\gamma_0$ converges to γ_0^* in probability and in mean as $n \to \infty$. The regularity conditions involved are simple, due to the monotonicity of the estimating function in γ, i.e., the right-hand side of (3.6).

Proposition 1. *Let* $\Theta_n = \{\theta \in \mathbb{R}^n : n^{-1}\sum_j \theta_j^2/d_j \le M\}$ *for a constant* M *free of* n. *Under Assumptions* (A1) *and* (A3) *with* $0 \le \eta < 1/2$, *the following results hold.*

(i) $\hat{\gamma}_0 - \gamma_0^* = O_p\{n^{-(1-2\eta)/2}\}$ *uniformly in* $\theta \in \Theta_n$. *That is, for any* $\tau_1 > 0$, $\sup_{\theta\in\Theta_n} P\{|\hat{\gamma}_0 - \gamma_0^*| \geq \tau_2 n^{-(1-2\eta)/2}\} \leq \tau_1$ *for all sufficiently large* n *and* τ_2.

(ii) $\sup_{\theta\in\Theta_n} E\{|\hat{\gamma}_0 - \gamma_0^*|^2\} = O\{n^{-(1-2\eta)/2}\}$.

We then study the accuracy of the plug-in risk estimator, $\mathrm{SURE}(\delta_{\lambda,\hat{\gamma}_0,\beta=0})$, defined as $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta=0})$ with γ replaced by $\hat{\gamma}_0$. Although $\mathrm{SURE}(\delta_{\lambda,\gamma,\beta})$ is an unbiased estimator of the risk of $\delta_{\lambda,\gamma,\beta}$ for fixed (λ,γ,β), the estimator $\mathrm{SURE}(\delta_{\lambda,\hat{\gamma}_0,\beta=0})$ is no longer unbiased for the risk of $\delta_{\lambda,\hat{\gamma}_0,\beta=0}$ because of the data-dependency of $\hat{\gamma}_0$. Nevertheless, we show that $\mathrm{SURE}(\delta_{\lambda,\hat{\gamma}_0,\beta=0})$ is, uniformly over $0 \leq \lambda \leq 2$, close to the actual loss of $\delta_{\lambda,\hat{\gamma}_0,\beta=0}$ in probability and in mean as $n \to \infty$. The rates of convergence are the same as in Theorem 3, but uniform only in $\theta \in \Theta_n$, with Θ_n defined in Proposition 1 to obtain uniform convergence of $\hat{\gamma}_0$ to γ_0^*.

Theorem 4. *Let* $\zeta_n(\lambda,\gamma) = n^{-1}\{SURE(\delta_{\lambda,\gamma,\beta=0}) - L(\delta_{\lambda,\gamma,\beta=0},\theta)\}$. *Under Assumptions* (A1)−(A3) *with* $0 \leq \eta < 1/4$, *the following results hold.*

(i) $\sup_{0\leq\lambda\leq 2} |\zeta_n(\lambda,\hat{\gamma}_0)| = O_p\{n^{-(1-4\eta)/2}\}$, *uniformly in* $\theta \in \Theta_n$. *That is, for any* $\tau_1 > 0$, $\sup_{\theta\in\Theta_n} P\{\sup_{0\leq\lambda\leq 2} |\zeta_n(\lambda,\hat{\gamma}_0)| \geq \tau_2 n^{-(1-4\eta)/2}\} \leq \tau_1$ *for all sufficiently large* τ_2 *and* n.

(ii) *If further* $0 \leq \eta < 1/5$, *then* $\sup_{\theta\in\Theta_n} E\{\sup_{0\leq\lambda\leq 2} |\zeta_n(\lambda,\hat{\gamma}_0)|\} = O\{n^{-(1-5\eta)/2}\}$.

As a consequence of Theorem 4, we see that, uniformly in $\theta \in \Theta_n$,

$$L(\delta_{\hat{\lambda}_0,\hat{\gamma}_0},\theta) \leq \inf_{0\leq\lambda\leq 2} L(\delta_{\lambda,\hat{\gamma}_0,\beta=0},\theta) + O_p\{n^{-(1-4\eta)/2}\}, \tag{4.3}$$

$$R(\delta_{\hat{\lambda}_0,\hat{\gamma}_0},\theta) \leq \inf_{0\leq\lambda\leq 2} R(\delta_{\lambda,\hat{\gamma}_0,\beta=0},\theta) + O\{n^{-(1-5\eta)/2}\}. \tag{4.4}$$

Of course, (4.3) and (4.4) are then valid pointwise for any $\theta \in \mathbb{R}^n$. Similarly as discussed in Xie, Kou, and Brown (2012), the actual loss of $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}$ is, by (4.3), asymptotically as small as that of the oracle loss estimator $\delta_{\lambda,\hat{\gamma}_0,\beta=0}$ with $0 \leq \lambda \leq 2$ selected to minimize $L(\delta_{\lambda,\hat{\gamma}_0,\beta=0},\theta)$. Moreover, the risk (i.e., expected loss) of $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}$, by (4.3), is asymptotically no greater than that of $\delta_{\lambda,\hat{\gamma}_0,\beta=0}$ for any fixed choice $0 \leq \lambda \leq 2$.

To obtain the earlier statement (4.2), the risk of $\delta_{\hat{\lambda}_0,\hat{\gamma}_0}$ can be directly related to that of $\delta_{\lambda,\gamma_0^*,\beta=0}$ which is not only minimax over $\theta \in \mathbb{R}^n$ but also, if $\lambda = 1$, near-Bayes optimal under a class of normal priors including $\mathrm{N}(0,\gamma_0^* I)$.

Corollary 1. *If Assumptions* (A1)−(A3) *hold with* $0 \leq \eta < 1/5$, *then, uniformly in* $\theta \in \Theta_n$,

$$R(\delta_{\hat{\lambda}_0,\hat{\gamma}_0},\theta) \leq \inf_{0\leq\lambda\leq 2} R(\delta_{\lambda,\gamma_0^*,\beta=0},\theta) + O(n^{-(1-5\eta)/2}\}.$$

Third, we study risk properties of $\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}$, that is, $\delta_{\lambda,\gamma,\beta}$ with the data-dependent choices $(\hat{\lambda},\hat{\gamma},\hat{\beta}) = (\hat{\lambda}_{\mathrm{FH}},\hat{\gamma}_{\mathrm{FH}},\hat{\beta}_{\mathrm{FH}})$ for (λ,γ,β). By estimating equations (3.1) and (3.2) for $(\hat{\gamma}_{\mathrm{FH}},\hat{\beta}_{\mathrm{FH}})$, we define, for any $\gamma \geq 0$,

$$\beta^*(\gamma) = (X^{\mathrm{T}} D_\gamma^{-1} X)^{-1} (X^{\mathrm{T}} D_\gamma^{-1} \theta),$$

and define $\beta^* = \beta^*(\gamma^*)$ and γ^* as a solution to the equation

$$\sum_{j=1}^{n} \frac{d_j + \{\theta_j - x_j^{\mathrm{T}} \beta^*(\gamma)\}^2}{d_j + \gamma} = n - q$$

if the right-hand side of the equation at $\gamma = 0$ is at least $n - q$ or let $\gamma^* = 0$ otherwise. The following proposition, similar to Proposition 1, concerns the convergence of $\hat{\gamma}$ to γ^* in probability and in mean as $n \to \infty$. In the special case of $\eta = 0$, Assumption (A3) reduces to saying that $(d_1, \ldots, d_n)$ are bounded from below and above by some positive constants, and Assumption (A5) reduces to saying that $\max_j \{x_j^{\mathrm{T}} (\sum_k x_k^{\mathrm{T}} x_k)^{-1} x_j\}$ is bounded from above. These simple conditions are commonly assumed in existing asymptotic theory for small-area estimation using the Fay–Herriot model (1.3) (e.g., Prasad and Rao (1990, Thm. A2)).

Proposition 2. *Assume that* (A1) *and* (A3) *hold with* $0 \leq \eta < 1/2$, *and there exists a constant* K_5 *such that for all* $n \geq 1$,
(A5) $\max_j \{x_j^{\mathrm{T}} d_j^{-1/2} (\sum_k x_k^{\mathrm{T}} x_k / d_k)^{-1} d_j^{-1/2} x_j\} \leq K_5 n^{-(1-2\eta)}$.
Then the following results hold.

(i) $\hat{\gamma} - \gamma^* = O_p\{n^{-(1-2\eta)/2}\}$ *uniformly in* $\theta \in \Theta_n$. *That is, for any* $\tau_1 > 0$, $\sup_{\theta \in \Theta_n} P\{|\hat{\gamma} - \gamma^*| \geq \tau_2 n^{-(1-2\eta)/2}\} \leq \tau_1$ *for all sufficiently large* n *and* τ_2.
(ii) $\sup_{\theta \in \Theta_n} E|\hat{\gamma} - \gamma^*| = O\{n^{-(1-2\eta)/2}\}$.
(iii) *If, in addition,* $0 \leq \eta < 1/4$, *then* $\sup_{\theta \in \Theta_n} E\{|\hat{\gamma} - \gamma^*|^2\} = O\{n^{-(1-4\eta)/2}\}$.

We then have the following theorem on the SURE approximation of the loss of $\delta_{\lambda,\hat{\gamma},\beta_{\hat{\gamma}}}$. The rate of convergence in probability is the same as in Theorem 3, but the rate of convergence in mean is slower than that in Theorem 3.

Theorem 5. *Let* $\zeta_n(\lambda,\gamma,\beta) = n^{-1}\{SURE(\delta_{\lambda,\gamma,\beta}) - L(\delta_{\lambda,\gamma,\beta},\theta)\}$. *Under Assumptions* (A1)–(A3) *and* (A5) *with* $0 \leq \eta < 1/4$, *the following results hold.*

(i) $\sup_{0 \leq \lambda \leq 2} |\zeta_n(\lambda,\hat{\gamma},\hat{\beta})| = O_p\{n^{-(1-4\eta)/2}\}$, *uniformly in* $\theta \in \Theta_n$. *That is, for any* $\tau_1 > 0$, $\sup_{\theta \in \Theta_n} P\{\sup_{0 \leq \lambda \leq 2} |\zeta_n(\lambda,\hat{\gamma},\hat{\beta})| \geq \tau_2 n^{-(1-4\eta)/2}\} \leq \tau_1$ *for all sufficiently large* τ_2 *and* n.
(ii) *If further* $0 \leq \eta < 1/6$, *then* $\sup_{\theta \in \Theta_n} E\{\sup_{0 \leq \lambda \leq 2} |\zeta_n(\lambda,\hat{\gamma},\hat{\beta})|\} = O\{n^{-(1-6\eta)/2}\}$.

In parallel to (4.3) and (4.4), Theorem 5 implies that, uniformly in $\theta \in \Theta_n$,

$$L(\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}},\theta) \le \inf_{0\le\lambda\le 2} L(\delta_{\lambda,\hat{\gamma},\hat{\beta}},\theta) + O_p\{n^{-(1-4\eta)/2}\}, \tag{4.5}$$

$$R(\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}},\theta) \le \inf_{0\le\lambda\le 2} R(\delta_{\lambda,\hat{\gamma},\hat{\beta}},\theta) + O\{n^{-(1-6\eta)/2}\}. \tag{4.6}$$

These results, (4.5) and (4.6), are then valid pointwise for $\theta \in \mathbb{R}^n$. Moreover, the risk of $\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}}$ can be directly related to that of $\delta_{\lambda,\gamma^*,\beta^*}$ as follows, thereby justifying the previous statement (4.1) with a specific rate of convergence.

Corollary 2. *If Assumptions* (A1)−(A3) *and* (A5) *hold with* $0 \le \eta < 1/6$*, then, uniformly in* $\theta \in \Theta_n$*,*

$$R(\delta_{\hat{\lambda},\hat{\gamma},\hat{\beta}},\theta) \le \inf_{0\le\lambda\le 2} R(\delta_{\lambda,\gamma^*,\beta^*},\theta) + O(n^{-(1-6\eta)/2}\}.$$

5. Application to Baseball Data

For the Major League Baseball season of 2005, Brown (2008) studied the problem of using the batting data from all the players with at least 11 at-bats in the first half season to estimate their latent batting probabilities or, as a validation, to predict their batting averages in the second half season. This problem has since been used to test and compare various shrinkage estimators in, for example, Jiang and Zhang (2010), Xie, Kou, and Brown (2012), and Koenker and Mizera (2014). We adopt the same setup to evaluate the performance of the proposed estimators.

For the jth player, let N_{ji} and H_{ji} be the number of at-bats and number of hits in the first ($i = 1$) or second ($i = 2$) half season. Assume that $H_{ji} \sim \text{Binomial}(N_{ji}, p_j)$, where p_j is the batting probability of the jth player for both half seasons. Brown (2008) suggested the following variance-stabilizing transformation

$$y_{ji} = \arcsin\sqrt{\frac{H_{ji}+1/4}{N_{ji}+1/2}},$$

such that y_{ji} is approximately distributed as $\mathrm{N}\{\theta_j, (4N_{ji})^{-1}\}$ with $\theta_j = \arcsin\sqrt{p_j}$. The problem of interest is then using $\{y_{j1} : j \in S_1\}$ to estimate $\{\theta_j : j \in S_1 \cap S_2\}$ or to predict $\{y_{j2} : j \in S_1 \cap S_2\}$, where $S_i = \{j : S_{ji} \ge 11\}$ $(i = 1, 2)$.

As noted in Brown (2008), the estimation and prediction problems are directly related to each other. For an estimator $\delta = \{\delta_j : j \in S_1\}$ of $\theta = \{\theta_j : j \in S_1\}$, define the sum of squared estimation error as $\text{SSEE} = \sum_{j\in S_1\cap S_2}(\delta_j - \theta_j)^2$, and the sum of squared prediction error as $\text{SSPE} = \sum_{j\in S_1\cap S_2}(y_{j2} - \delta_j)^2$. By the independence of the half seasons, $E_\theta(\text{SSPE}) = E_\theta(\text{SSEE}) + \sum_{j\in S_1\cap S_2}(4N_{j2})^{-1}$.

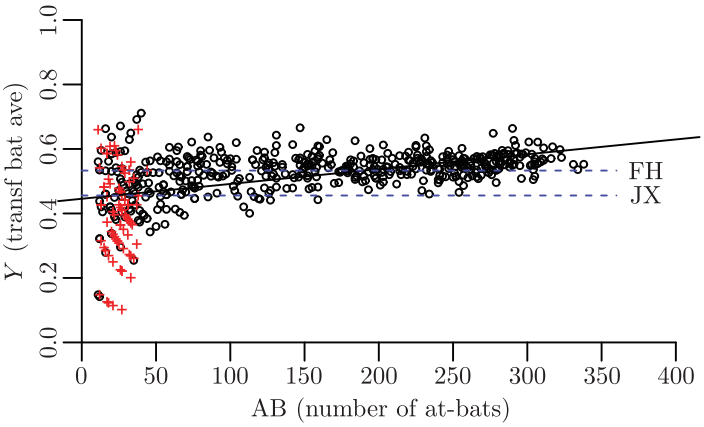

Figure 1. Scatterplot of transformed batting averages y_{j1} versus numbers of at-bats N_{j1} for nonpitchers ($\circ$) and pitchers ($+$), with the unweighted least squares regression line. Two horizontal lines are placed at the estimates $\hat{\beta}_{\mathrm{FH}}$ and $\hat{\beta}_{\mathrm{JX}}$ with 1 as the only covariate.

An unbiased estimator of $E_\theta(\mathrm{SSEE})$, called total squared error (TSE) in Brown (2008), is

$$\widehat{\mathrm{TSE}}(\delta) = \sum_{j \in S_1 \cap S_2} (y_{j2} - \delta_j)^2 - \sum_{j \in S_1 \cap S_2} \frac{1}{4N_{2j}}.$$

This error measure will be used in the subsequent comparison of estimators.

Figure 1 shows the batting data on $|S_1| = 567$ player from the first half season. There are considerable differences in the numbers of at-bats, ranging from 11 to 338, in contrast with the previous data used in Efron and Morris (1975) where all the players were selected to have the same number of at-bats by a certain date. Moreover, as observed in Brown (2008), the batting averages appear to be dispersed about different means for the two groups of pitchers and nonpitchers, and are positively correlated with the numbers of at-bats, especially within the nonpitchers.

Table 1 summarizes the performance of various estimators, depending on what covariates are used, for predicting the batting averages of $|S_1 \cap S_2| = 499$ players in the second half season. As known from Brown (2008), the naive estimator $\delta_0 = Y$ performs very poorly, even compared with the grand mean estimator, which ignores differences between individual players. The proposed estimators, δ_{Res} and δ_{Sub}, yield estimation errors comparable to or smaller than those of the competing estimators δ_{FH} and δ_{JX}, in all situations except when 1 is used as the only covariate, i.e., all observations are shrunk to a constant location. There are a number of other competing estimators studied for this problem by Brown (2008), Jiang and Zhang (2010), Xie, Kou, and Brown (2012),

Table 1. Relative values of $\widehat{\mathrm{TSE}}$.

Covariates	Naive	Grand mean	FH	JX	Residual	Subspace
1	1.000	0.853	0.702	0.421	0.524	0.551
1 + AB	1.000	0.853	0.444	0.398	0.359	0.418
1 + pitcher	1.000	0.853	0.249	0.213	0.241	0.250
1 + AB + pitcher	1.000	0.853	0.193	0.215	0.180	0.184
1 + AB ∗ pitcher	1.000	0.853	0.180	0.215	0.169	0.169

Note: The relative values shown are $\widehat{\mathrm{TSE}}$ of all estimators divided by those of the naive estimator $\delta_0 = Y$; the grand mean is defined as $n^{-1}\sum_{j=1}^{n} Y_j$; AB = the number of at-bats in the first half season; AB ∗ pitchers = AB + pitcher + interaction.

and Koenker and Mizera (2014). Examination of their results still shows that the performance of δ_{Res} and δ_{Sub} compares favorably with the best of these other estimators, in all situations except with 1 as the only covariate.

For the situation where 1 is used as the only covariate, δ_{Res} and δ_{Sub} yield estimation errors smaller than that of δ_{FH}, but greater than that of δ_{JX}. By construction, δ_{FH} and δ_{Res} involve using the same estimates $(\hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}})$. Therefore, the outperformance of δ_{Res} over δ_{FH} demonstrates the benefit of Steinization in δ_{Res}, inherited from the minimax Bayes estimator $\delta_{\lambda,\gamma,\beta}$. On the other hand, there are substantial differences between the estimates of (γ, β) used in δ_{Res} and δ_{JX}, $(\hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}}) = (0.00188, 0.533)$ and $(\hat{\gamma}_{\mathrm{JX}}, \hat{\beta}_{\mathrm{JX}}) = (0.00540, 0.456)$. As shown in Figure 1, the estimate $\hat{\beta}_{\mathrm{FH}}$ seems to represent the center of the true θ-values better than $\hat{\beta}_{\mathrm{JX}}$, which is much smaller than $\hat{\beta}_{\mathrm{FH}}$. To compensate for underestimation in $\hat{\beta}_{\mathrm{JX}}$, the estimate $\hat{\gamma}_{\mathrm{JX}}$ is noticeably higher than $\hat{\gamma}_{\mathrm{FH}}$ for capturing the spread of the true θ-values. Therefore, δ_{JX} performs better than δ_{FH} and δ_{Res}, but has a nonintuitive shrinkage pattern determined by $(\hat{\gamma}_{\mathrm{JX}}, \hat{\beta}_{\mathrm{JX}})$. However, the comparison between δ_{JX} and δ_{Res} is situation-dependent, even when the second-level model (1.3) is misspecified. In fact, δ_{Res} performs better than δ_{JX} when the covariates used are 1 + AB, with the pitcher effect ignored. See Section 6 and Supplementary Material for simulation results and further discussion, where δ_{Res} performs similarly to or better than δ_{JX} even when using only the covariate 1.

To illustrate the advantage of δ_{Res}, Figure 2 shows how the observations are shrunk under δ_{Res}, δ_{FH}, and δ_{JX} when the covariates used are 1 + AB + pitcher. For all three estimators, the observations from pitchers are approximately linearly shrunk, because their numbers of at-bats fall in a narrow range and hence their variances are relatively homogeneous. For the nonpitchers, the observations with large variances are also approximately linearly shrunk, whereas those with small variances are shrunk less substantially, by varying magnitudes, than those with large variances. The associated range of magnitudes of shrinkage for δ_{Res} appears to be narrower than for δ_{FH} and δ_{JX}. Overall, the shrinkage pattern in δ_{Res} seems to be better aligned than δ_{FH} and δ_{JX} with the linear predictor that would be

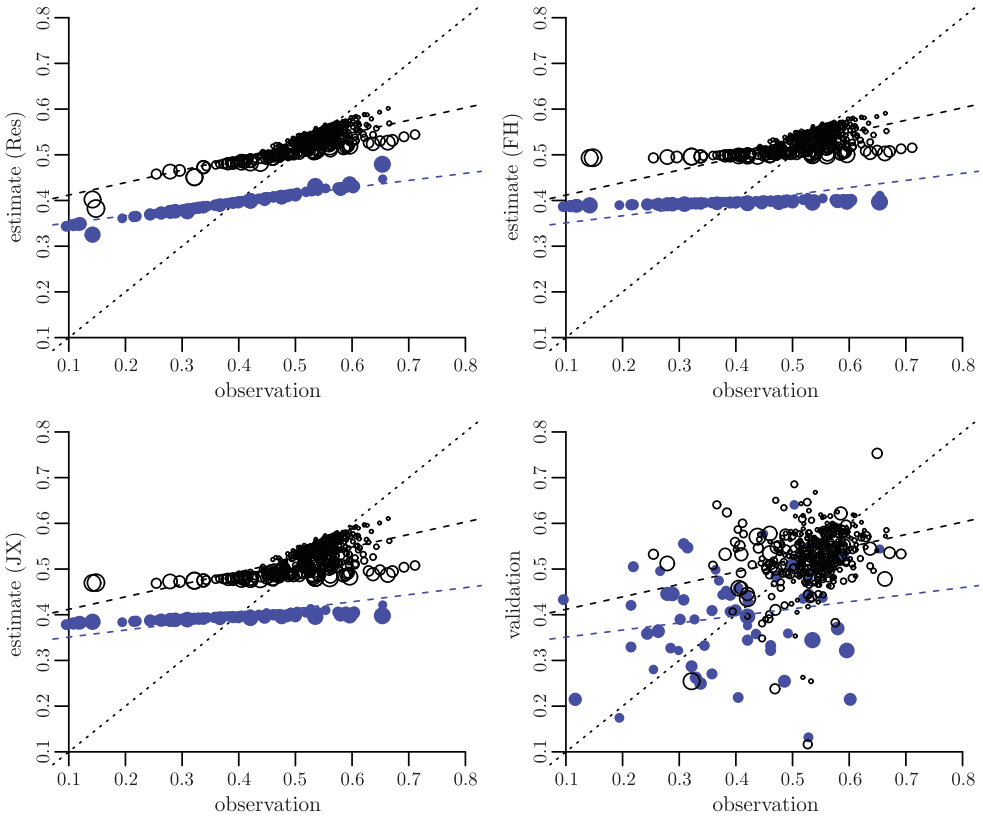

Figure 2. Scatterplots of estimates δ_j versus observations y_{j1} for three estimators using the covariates 1+AB+pitcher, and the scatterplot (lower right) of validation data y_{j2} from the second half season versus training data y_{j1} from the first half season. All circles, empty (nonpitcher) and filled (pitcher), have areas proportional to the variances d_j. Two regression lines are drawn by unweighted least squares for y_{j2} versus y_{j1} in nonpitchers and pitchers respectively, in the lower-right plot and superimposed in all other plots.

obtained within pitchers and, separately, nonpitchers if validation data were used. Although none of the estimators δ_{Res}, δ_{FH}, and δ_{JX} are strictly linear predictors within the pitchers or the nonpitchers, the closer alignment of δ_{Res} with the oracle linear predictor serves to explain the outperformance of δ_{Res} over δ_{FH} and δ_{JX} in the present situation.

6. Simulation Study

Simulation studies were conducted to further compare the three estimators δ_{FH}, δ_{JX}, and δ_{Res}. We present in the Supplementary Material a simulation study where the variances d_j were randomly generated as in Xie, Kou, and Brown

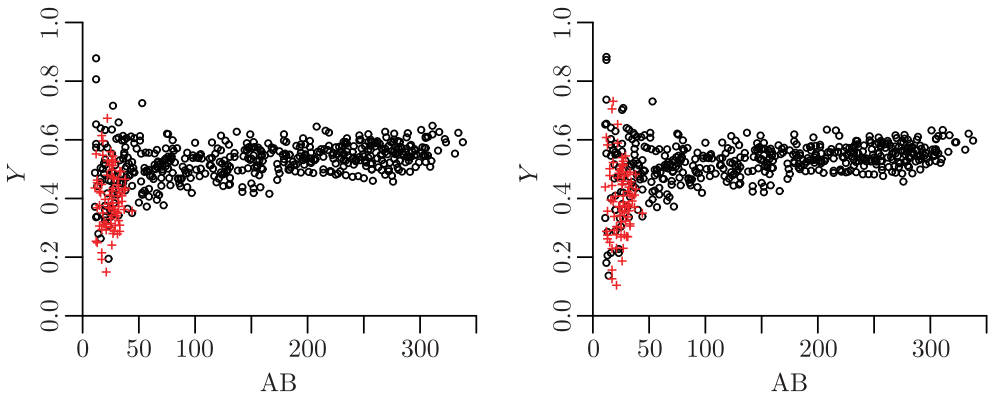

Figure 3. Scatterplots of simulated observations versus numbers of at-bats for nonpitchers (∘) and pitchers (+), based on the homoscedastic prior (1.2) with $\sqrt{\gamma} = 0.02$ (left) and the heteroscedastic prior (6.1) with $\sqrt{\alpha} = 0.2$ (right).

(2012) and report here a simulation study mimicking the setting of baseball data in Brown (2008). The observations Y_j were simulated from the basic model (1.1), with θ_j generated from the prior (1.2), $\theta_j \sim \mathrm{N}(x_j^{\mathrm{T}}\beta, \gamma)$, or alternatively from

$$\theta_j \sim \mathrm{N}\{x_j^{\mathrm{T}}\beta, \alpha(\mathrm{AB}_j - 10)^{-1}\}, \quad j = 1, \ldots, n, \tag{6.1}$$

where $d_j = (4\,\mathrm{AB}_j)^{-1}$ and $x_j = (1, \mathrm{AB}_j, \mathrm{pitcher}_j)^{\mathrm{T}}$, with AB_j the number of at-bats in the first half season and $\mathrm{pitcher}_j$ the group indicator for pitchers as in Brown's (2008) baseball data. Then the mean θ_j depends on d_j through x_j. These priors are referred to as the homoscedastic and heteroscedastic data-generating priors. Nevertheless, for all the estimators δ_{FH}, δ_{JX}, and δ_{Res}, the second-level model (1.2) was used but perhaps misspecified, in the mean or the variance or both, as compared with the data-generating prior. There is a mean misspecification in (1.2) when the data-generating prior is homoscedastic but some of the required covariates are dropped. There is a variance misspecification in (1.2) when the data-generating prior is heteroscedastic.

The Fay–Herriot estimates from the real data were $\hat{\beta}_{\mathrm{FH}} = (0.50, 0.00023, -0.11)^{\mathrm{T}}$ and $\hat{\gamma}_{\mathrm{FH}} = (0.018)^2$. To mimic these estimates, the true value of β was set such that $x_j^{\mathrm{T}}\beta = 0.5 + 0.0002(\mathrm{AB}_j) - 0.1(\mathrm{pitcher}_j)$. The possible values of $\sqrt{\gamma}$ were $0.01, 0.02, 0.04$, and 0.08, and those of $\sqrt{\alpha}$ were $0.1, 0.2, 0.4$, and 0.8. For $\sqrt{\gamma} = 0.02$ and $\sqrt{\alpha} = 0.2$, Figure 3 shows two sets of simulated observations based on the homoscedastic and heteroscedastic priors. The data configuration in each plot is superficially similar to that in Figure 1, but a careful examination of the lower extremes of the observations suggests that the simulated data based on the heteroscedastic prior might mimic the real data more closely than when based on the homoscedastic prior.

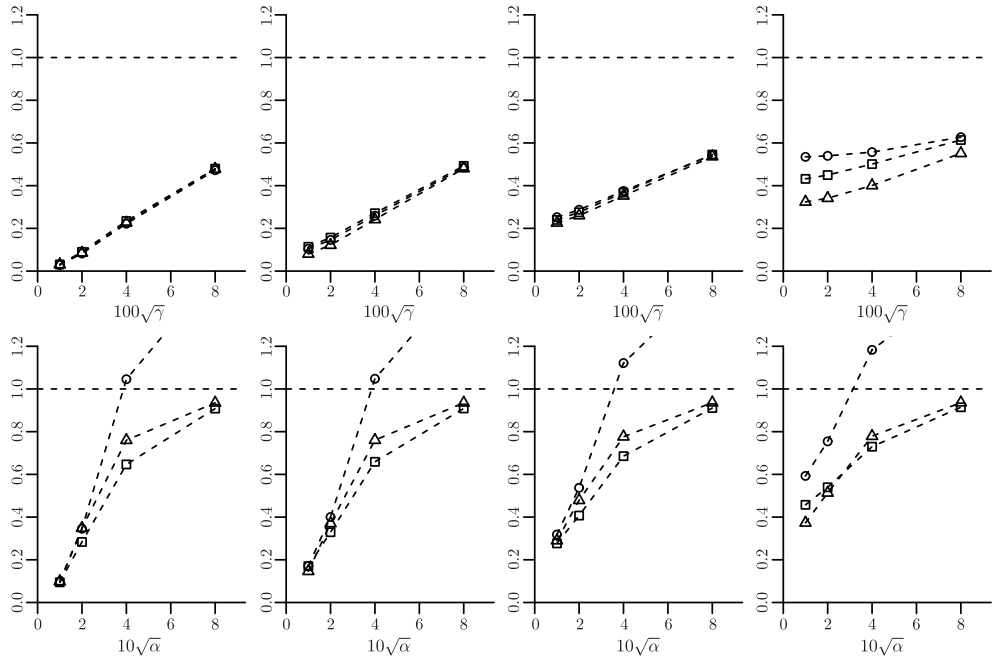

Figure 4. Relative Bayes risks of three estimators δ_{FH} ($\circ$), δ_{JX} ($\triangle$), and δ_{Res} ($\square$) using the covariates $1 + \mathrm{AB} + \mathrm{pitcher}$ (first column), $1 + \mathrm{pitcher}$ (second column), $1 + \mathrm{AB}$ (third column), and 1 (fourth column), based on simulated observations from the homoscedastic prior (1.2) as γ varies (first row) and the heteroscedastic prior (6.1) as α varies (second row).

Figure 4 shows the relative Bayes risks of δ_{FH}, δ_{JX}, and δ_{Res} versus that of the naive estimator, obtained from 10^4 repeated simulations. Similarly as in Section 5, the proposed estimator δ_{Res} yields Bayes risks at least as small as those of δ_{FH} in all situations, and as δ_{JX} in all situations except when 1 is used as the only covariate. In fact, if the data-generating prior is homoscedastic and hence the prior variance in (1.2) is correctly specified, the three estimators perform similarly to each other when the prior mean in (1.2) is either correctly specified or misspecified with only pitcher or AB included as a non-constant covariate. If the data-generating prior is heteroscedastic and hence the prior variance in (1.2) is misspecified, then δ_{Res} performs similarly to or noticeably better than both δ_{FH} and δ_{JX}, regardless of how the prior mean in (1.2) is specified. In this case, δ_{FH} performs poorly, with Bayes risks rising quickly above even the constant risk of the naive estimator as the scale parameter α increases in the heteroscedastic data-generating prior.

For the situation where the data-generating prior is homoscedastic but the prior mean in (1.2) is specified with 1 as the only covariate, δ_{Res} yields Bayes risks smaller than those of δ_{FH}, but larger than those of δ_{JX}. However, as seen

from the results in all other situations, such an outperformance of δ_{JX} over δ_{Res} depends on the particular data configuration and model misspecification. The presence of both a negative pitcher effect and a positive AB effect seems to lead to substantially different estimates $(\hat{\gamma}_{\mathrm{JX}}, \hat{\beta}_{\mathrm{JX}})$ from $(\hat{\gamma}_{\mathrm{FH}}, \hat{\beta}_{\mathrm{FH}})$, such that δ_{JX} achieves smaller Bayes risks than δ_{Res}. To support this explanation, we replicated the simulations in the same setup, except with a negative AB effect, i.e., the true value of β was set such that $x_j^{\mathrm{T}}\beta = 0.5 - 0.0002(\mathrm{AB}_j) - 0.1(\mathrm{pitcher}_j)$. As shown in Figure S5 in the Supplementary Material, the three estimators δ_{FH}, δ_{JX}, and δ_{Res} perform similarly to each other when the data-generating prior is homoscedastic, regardless of how the prior mean in (1.2) is specified. But δ_{Res} still outperforms δ_{FH} and δ_{JX}, sometimes more substantially than in Figure 4, when the data-generating prior is heteroscedastic.

7. Conclusion

To estimate normal means with heteroscedastic observations from the basic model (1.1), conventional empirical Bayes methods (Efron and Morris (1973); Fay and Herriot (1979)) based on the second-level model (1.2) involves employing the Bayes rule under a fixed prior (1.2) with (γ, β) selected by maximum likelihood or moment equations. However, the performance of such methods tends to substantially deteriorate when the second-level model (1.2) is misspecified. To address this issue, Jiang, Nguyen, and Rao (2011) and Xie, Kou, and Brown (2012) independently proposed a SURE-based empirical Bayes method, which retains the Bayes rule under a fixed prior (1.2), but with (γ, β) selected by minimizing the SURE of the Bayes rule. The SURE-based method is often more robust than the model-based methods, but still susceptible to unsatisfactory performance when the second-level model (1.2) is misspecified, particularly in the variance, which directly determines the magnitude of shrinkage.

There is a crucial difference in how the existing and proposed estimators are constructed using the second-level model (1.2). For a fixed prior (1.2), all the existing empirical Bayes estimators would reduce to the Bayes rule, which is sensitive to possible misspecification of (1.2). In contrast, the proposed estimators would reduce to the minimax Bayes estimator of Tan (2015), which, due to Steinization, is not only globally minimax but also achieves close to the minimum Bayes risk over a scale class of normal priors including the fixed prior. This difference helps to explain why the Steinized empirical Bayes methods, even using model-based estimates of (γ, β), could perform better than existing empirical Bayes methods, using either model-based estimates or the SURE-based estimates of (γ, β).

The development in this article can be extended when a more complicated second-level model than a homoscedastic prior (1.2) is considered. For example,

Fay and Herriot (1979) mentioned a second-level model, $\theta_j \sim \mathrm{N}(x_j^{\mathrm{T}}\beta, \gamma d_j^{\alpha})$, with (γ, β, α) as unknown parameters, but suggested that the resulting method would not be preferable unless n is large. Alternatively, following the idea of block shrinkage in the homoscedastic case (e.g., Cai and Zhou (2009)), it is possible to divide the coordinates into blocks and consider a block-homoscedastic prior (i.e., the prior variances are equal within each block) even for heteroscedastic data. Our approach can be extended in this direction and compared with existing block shrinkage methods.

Supplementary Materials

Supplementary materials available at the journal website include (i) additional discussion mentioned in Section 3.1−3.2 and additional simulation results mentioned in Section 6, and (ii) the proofs of theorems.

Acknowledgement

The author thanks Bill Strawderman and Cunhui Zhang for valuable discussions at various stages of the research, and an associate editor and two referees for helpful comments.

References

Ben-Hain, Z. and Eldar, Y. C. (2007). Blind minimax estimation. *IEEE Trans. Inform. Theory* **53**, 3145-3157.

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

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

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

Brown, L. D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters). *J. Amer. Statist. Assoc.* **70**, 417-427.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Statist.* **2**, 113-152.

Cai, T. T. and Zhou, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.* **37**, 569-595.

Datta, D. and Ghosh, M. (2012). Small area shrinkage estimation. *Statist. Sci.* **27**, 95-114.

Datta, G. S., Rao, J. N. K. and Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika* **92**, 183-196.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99**, 277-303.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

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

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**, 311-319.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* **74**, 269-277.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, New York.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 361-379.

Jiang, J., Nguyen, T. and Rao, J. S. (2011). Best predictive small area estimation. *J. Amer. Statist. Assoc.* **106**, 732-745.

Jiang, W. and Zhang, C.-H. (2010). Empirical Bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications: A Festschrift for Lawrence D. Brown*, 263-273, Institute for Mathematical Statistics.

Johnstone, I. M. (2013). *Gaussian Estimation: Sequence and Wavelet Models*. Book draft.

Judge, G. G. and Mittelhammer, R. C. (2004). A semiparametric basis for combining estimation problems under quadratic loss. *J. Amer. Statist. Assoc.* **99**, 479-487.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, com- pound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109**, 674-685.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. 2nd edition. Springer, New York.

Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation. *Ann. Statist.* **13**, 1352-1377.

Lindley, D. V. (1962). Discussion of a paper by C. Stein. *J. Roy. Statist. Soc. Ser. B* **24**, 285-287.

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

Morris, C. N. and Lysy, M. (2012). Shrinkage estimation in multilevel normal models. *Statist. Sci.* **27**, 115-134.

Oman, S. D. (1982). Contracting towards subspaces when estimating the mean of a multivariate normal distribution. *J. Multivariate Anal.* **12**, 270-290.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statist. Sci.* **28**, 40-68.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of mean squared errors of small-area estimators. *J. Amer. Statist. Assoc.* **85**, 163-171.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`.

Sclove, S. L., Morris, C. and Radhakrishnan, R. (1972). Non-optimality of preliminary- test estimators for the mean of a multivariate normal distribution. *Ann. Statist.* **43**, 1481-1490.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multi- variate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**, 197-206.

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

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* (Edited by M.-H. Chen, et al.), 69-82, Springer, New York.

Tan, Z. (2015). Improved minimax estimation of a multivariate normal mean under heteroscedasticity. *Bernoulli* **21**, 574-603.

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

Department of Statistics, Rutgers University. Address: 110 Frelinghuysen Road, Piscataway, NJ 08854, USA.

E-mail: ztan@stat.rutgers.edu

(Received October 2014; accepted October 2015)