Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean

Asaf Weinstein, Zhuang Ma, Lawrence D. Brown & Cun-Hui Zhang


To link to this article: https://doi.org/10.1080/01621459.2017.1280406
Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean

Asaf Weinstein\textsuperscript{a}, Zhuang Ma\textsuperscript{b}, Lawrence D. Brown\textsuperscript{c}, and Cun-Hui Zhang\textsuperscript{d}

\textsuperscript{a}Department of Statistics, Stanford University, Stanford, CA; \textsuperscript{b}Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA; \textsuperscript{c}Miers Busch Professor and Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA; \textsuperscript{d}Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ

1. Introduction

Let \( X = (X_1, \ldots, X_n)^T \), \( \theta = (\theta_1, \ldots, \theta_n)^T \) and \( V = (V_1, \ldots, V_n)^T \), and suppose that

\[
X_i|\theta, V_i \sim N(\theta_i, V_i)
\]

independently for \( 1 \leq i \leq n \), where \( \theta \) and \( V \) are deterministic. In the heteroscedastic normal mean problem, the goal is to estimate the vector \( \theta \) based on \( X \) and \( V \) under the (normalized) sum-of-squares loss

\[
L_n(\hat{\theta}, \theta) = n^{-1}\|\hat{\theta} - \theta\|^2 = n^{-1}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2.
\]

Hence, we assume that in addition to the random observations \( X_1, \ldots, X_n \), the variances \( V_1, \ldots, V_n \) are available. Allowing the values of \( V_i \) to be different from each other extends the applicability of the Gaussian mean problem to many realistic situations. A trivial but common example is the design corresponding to a one-way analysis of variance with unequal cell counts; here, \( X_i \) represents the mean of the \( n_i \) iid \( N(\theta_i, \sigma^2) \) observations for the \( i \)th subpopulation, hence \( V_i = \sigma^2/n_i \). More generally, if \( Y \sim N_p(A\theta, \sigma^2I) \) with a known design matrix \( A \), then estimating \( \theta \) under sum-of-squares loss is equivalent to estimating \( \theta \) in (1) where \( n = \text{rank}(A) \) and \( X_i \) and \( V_i/\sigma^2 \) are determined by \( A \) (see, e.g., Johnstone 2011, Section 2.9). In both cases, \( V_i \) are typically known only up to a proportionality constant which can be substituted by a consistent estimator.

The heteroscedastic normal mean problem has been studied extensively for both the special case of equal variances, \( V_i \equiv \sigma^2 \), and the more general case above. Alternative estimators to the usual minimax estimator \( \overline{\theta} = X \) have been suggested that perform better, for fixed \( n \) or only asymptotically (under some conditions), in terms of the risk \( R_n(\theta, \overline{\theta}) = \mathbb{E}[L_n(\theta, \overline{\theta})] \), regardless of \( \theta \). Here and elsewhere, we suppress in notation the dependence of the risk function on \( V \).

In the homoscedastic case, \( V_i \equiv \sigma^2 \), such shrinkage estimators go back, of course, to the James–Stein estimator, \( \overline{\theta} = \sigma^2 \) for some nonnegative \( \nu \) such that

\[
\mathbb{E}[L_n(\theta, \overline{\theta})] = (1 - \nu)\mathbb{E}[L_n(0, \sigma I)]
\]

which, for \( n \geq 3 \), has strictly smaller risk than the usual estimator for any \( \theta \). This estimator can be derived as an empirical Bayes estimator under a model that puts \( \theta \sim N_p(0, \gamma I) \), where \( \gamma \) is unspecified and “estimated” from the data \( X \). Equivalently, as observed by Efron and Morris (1973b), the James–Stein estimator is an empirical version of the linear Bayes rule (that is, the linear estimator with smallest Bayes risk) when \( \theta \) is only assumed to have iid components, not necessarily normally distributed. Therefore, the James–Stein estimator also performs well with respect to the usual estimator in terms of the Bayes risk when \( \theta \) is truly random with iid components. Efron and Morris (1973b, Section 9) analyze and quantify relative savings in Bayes risk when using the James–Stein estimator versus the usual estimator.

In addition to being minimax and exhibiting good Bayes performance, the James–Stein estimator in fact has attractive asymptotic optimality properties uniformly in \( \theta \). Denote \( \overline{\theta} = (1 - b)X \) for some nonnegative \( b \in \mathbb{R} \). Then, it holds that...
for any $\boldsymbol{\theta}$ (with a mild restriction on the sequence $\theta_i$, $i = 1, 2, \ldots, n$),
\[
R_n(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^\text{CS}) = R_n(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^b) + o(1)
\]
where $b^*_n = \arg\min_{b>0} \{R_n(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^b)\}$. The striking fact is that the oracle performance in (3) is achievable without knowing $\boldsymbol{\theta}$, a target of the kind set up and pursued by Herbert Robbins, can be intuitively understood from the connection between the original $n$-dimensional estimation problem with fixed $\boldsymbol{\theta}$ and a one-dimensional Bayesian estimation problem. We say that an estimator $\hat{\boldsymbol{\theta}}$ is separable if $\hat{\theta}_i = t(X_i)$ for some function $t: \mathbb{R} \to \mathbb{R}$. Then, as presented, for example, by Jiang and Zhang (2009), for a separable estimator with $\hat{\theta}_i = t(X_i)$,
\[
R_n(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_i [t(X_i) - \hat{\theta}_i]^2 = \mathbb{E} [t(Y) - \xi]^2
\]
where the expectation in the last term is taken over the pair $(Y, \xi)$ of random variables jointly distributed according to
\[
\xi \sim \frac{1}{n} \sum_{i=1}^n I(\theta_i \leq \xi), \quad Y|\xi \sim N(\xi, \sigma^2).
\]
Hence, the pointwise risk of a separable estimator is precisely the Bayes risk in a single copy of the original compound problem, where the prior is the empirical distribution of the (unknown) $\theta_i$. Since $\hat{\boldsymbol{\theta}}^b = (1 - b)\hat{\boldsymbol{\theta}}$ is a separable rule, it follows that the optimal estimator of this form has $b^*_n$ such that $(1 - b^*_n)\hat{\boldsymbol{\theta}}$ is the linear Bayes rule for predicting $\xi$ from $Y$, namely, $b^*_n = \sigma^2/\mathbb{E}_b (Y^2)$. The constant $b^*_n$ depends on the unknown vector $\theta$, but only through $1/\mathbb{E}_b (Y^2)$, which for large $n$ is well approximated by $(n-2)/|Y|^2$ (this estimator is exactly unbiased for $1/\mathbb{E}_b (Y^2)$ under $\theta = 0$).

In the heteroscedastic case, there is no such agreement as in the homoscedastic case between minimax estimators and existing empirical Bayes estimators regarding how the components of $\boldsymbol{X}$ should be shrunk relative to their individual variances. Existing parametric empirical Bayes estimators, which usually start by putting average on iid normal prior on the elements of $\theta$ and therefore shrink $X_i$ in proportion to $V_i$, are in general not minimax. And vice versa, minimax estimators do not provide substantial reduction in the Bayes risk under such priors, essentially under-shrinking the components with larger variances, and in some constructions (e.g., Berger 1976) even shrink $X_i$ inversely in proportion to $V_i$. Nontrivial spherically symmetric shrinkage estimators that have been suggested, that is, estimators that shrink all components by the same factor regardless of $V_i$, exist only when the $V_i$ satisfy certain conditions that restrict how much they can be spread out. See Tan (2015) for a concise review of some existing estimators and references therein for related literature. Before proceeding, we remark that it is tempting to scale $X_i$ by $1/\sqrt{V_i}$ to make all variances equal; however, after applying this non-orthogonal transformation the loss need be changed accordingly (to a weighted loss) to maintain equivalence between the problems. Hence the heteroscedastic problem cannot be exactly reduced to the equal variances case: the potential gains from shrinking differently on coordinates with different $V_i$ remain after normalization.

There have been attempts to moderate the respective disadvantages of estimators resulting from either of the two approaches mentioned above. For example, Xie, Kou, and Brown (2012) considered the family of Bayes estimators arising from the usual hierarchical model
\[
\hat{\theta}_i \overset{\text{ind}}{\sim} N(\mu, \gamma) \quad X_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, V_i) \quad 1 \leq i \leq n
\]
and indexed by $\mu$ and $\gamma$. They suggested to plug into the Bayes rule,
\[
\hat{\theta}_i^{\mu, \gamma} = \mathbb{E}_{\mu, \gamma}(\theta_i|X_i) = X_i - \frac{V_i}{V_i + \gamma}(X_i - \mu),
\]
values $(\hat{\mu}, \hat{\gamma}) = \arg\min_{\mu, \gamma} \mathcal{R}(\mu, \gamma; \boldsymbol{X})$, where $\mathcal{R}(\mu, \gamma; \boldsymbol{X})$ is an unbiased estimator of the risk of $\hat{\theta}_i^{\mu, \gamma}$. This reduces the sensitivity of the estimator to how appropriate model (5) is, as compared to the usual empirical Bayes estimators, that use maximum likelihood or method-of-moments estimates of $\mu$, $\gamma$ under (5). On the other hand, Berger (1982) suggested a modification of his own minimax estimator (Berger 1976), inspired by an approximate robust Bayes estimator (Berger 1980), that improves Bayesian performance while retaining minimaxity. Tan (2015) recently suggested a minimax estimator with similar properties that has a simpler form.

While empirical Bayes estimators based on (5) can be constructed so they asymptotically dominate the usual estimator (Xie, Kou, and Brown 2012), the modeling of $\theta_i$ as identically distributed random variables is not as well motivated as in the homoscedastic case. The assumption that $\theta_i$ are iid reflects, as commented by Efron and Morris (1973b, Section 8), a “Bayesian statement of belief that the $\theta_i$ are of comparable magnitude.” But this assumption is not always appropriate. For example, in a one-way ANOVA there are situations where the cell counts $n_i$, and hence the variances $V_i = \sigma^2/n_i$, are clearly associated with the effect size. There are other examples where an association between the $V_i$ and the $\theta_i$ is expected: in Section 5, we consider batting records for Major League baseball players, where better performing players tend to also have larger numbers of at-bats (afflicting the sampling variances of the observations). In situations where the true means and the $V_i$ are associated, modeling the $\theta_i$ as iid is not adequate. Also from a non-Bayesian perspective, note that while (4) justifies modeling $\theta_i$ as exchangeable in the homoscedastic case, the same calculation will not go through when $V_i$ are unequal (in that case $X_i - \theta_i$ do not have the same distribution). Nevertheless, we show that symmetry can be restored in the heteroscedastic case to produce a counterpart of (4), which, in turn, gives rise to a useful (oracle) benchmark for the performance of rules of the form $\hat{\theta}_i = t(X_i, V_i)$ where $t$ is linear in the first component. This observation leads us to develop a block-linear empirical Bayes estimator that groups together observations with similar variances and applies a spherically symmetric minimax estimator to each group separately. Importantly, for $n \geq 4$ the risk of our estimator never exceeds $\frac{n-1}{n} V_i$, hence from a minimax point of view there is no cost to using it as compared to the usual estimator.

The rest of the article is organized as follows. Section 2 presents the estimation of a heteroscedastic mean as a compound decision problem. This motivates the construction of a group-linear empirical Bayes estimator in Section 3; we discuss the properties of the proposed estimator and prove two oracle inequalities, which establish a sense of asymptotic optimality with respect to the class of estimators that are “conditionally”
linear. Simulation results are reported in Section 4. In Section 5, we apply our estimator to the baseball data by Brown (2008) and compare it to some of the best-performing estimators that have been tested on this dataset. Proofs appear in the Appendix.

2. A Compound Decision Problem for the Heteroscedastic Case

Let $X$, $\theta$ and $V$ be as in (1). In the homoscedastic case, a separable rule uses only $X_i$ to estimate $\theta_i$ in the heteroscedastic case it is natural to allow $\theta_i$ for a separable rule to also depend on $V_i$. Hence, in the following we refer to a rule $\theta$ as separable if $\theta_i(X, V) = t(X_i, V_i)$ for some function $t : \mathbb{R}^2 \to \mathbb{R}$. Denote by $D_S$ the set of all separable rules. If $\theta \in D_S$ with $\theta_i(X, V) = t(X_i, V_i)$, then

$$R_n(\theta, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} E\{t(X_i, V_i) - \theta_i\}^2 = E[t(Y, A) - \xi]^2. \quad (7)$$

where the expectation in the last term is taken over the random vector $(Y, \xi, A, I)^T$ distributed according to $P(I = i) = 1/n, (Y, \xi, A)(I = i) \sim (X_i, \theta_i, V_i) \quad 1 \leq i \leq n. \quad (8)$

Above, the symbol "$\sim$" stands for "equal in distribution." In words, (8) says that $(\xi, A)$ have the empirical joint distribution of the pairs $(\theta_i, V_i)$; and $Y(\xi, A) \sim N(\xi, A)$. Throughout the article, when we refer to the random triple $(Y, \xi, A)$, its relation to $(X_i, \theta_i, V_i), \quad 1 \leq i \leq n,$ is given by (8). The identity (7) is easily verified by calculating the expectation on the right-hand side by first conditioning on $I$, and says that for a separable estimator, the risk is again equivalent to the Bayes risk in a one-dimensional estimation problem. Note that (7) can be interpreted as an application of (4) to a compound decision problem as originally intended by Robbins—consisting of $n$ symmetric copies of a univariate decision problem—except that the data associated with the unknown parameter $\theta_i$ is now the pair $(X_i, V_i)$.

Now consider $\widehat{\theta} \in D_S$ with $t$ linear (affine, in point of fact, but with a slight abuse of terminology we use the former for convenience) in its first argument, that is,

$$\widehat{\theta}_i^{a,b}(X, V) = X_i - b(V_i)[X_i - a(V_i)] \quad 1 \leq i \leq n \quad (9)$$

for some functions $a, b$. The corresponding Bayes risk in the last expression of (7) is

$$r_n(a, b) := E\{|Y - b(A)|Y - a(A)\} - \xi\}^2. \quad (10)$$

Since

$$Y|\xi, A \sim N(\xi, A), \quad (11)$$

the minimizers of

$$r_n(a, b|\nu) := E\{(Y - b(A)|Y - a(A)| - \nu\}^2|A = \nu. \quad (12)$$

and hence also of (10), are

$$a^*_n(\nu) = E(Y|A = \nu), \quad b^*_n(\nu) = \frac{\nu}{\text{Var}(Y|A = \nu)} \quad (13)$$

and the minimum Bayes risk is

$$R_n(\theta, \hat{\theta}^{a^*_n, b^*_n}) = r_n(a^*_n, b^*_n) = E[A|\{1 - b^*_n(A)\}]. \quad (14)$$

Therefore, (14) is a lower bound on the risk achievable by any estimator of the form (9), and $\hat{\theta}^{a^*_n, b^*_n}$ is the optimal such decision rule. Note that any estimator of the form (6) is also of the form (9), hence the risk of the best (oracle) rule of the form (9) is no greater than the risk of the best rule of the form (6). If $\xi$ and $A$ are independent, $\hat{\theta}_n(\nu) = E(Y|A = \nu) = E(\xi|A = \nu) = E(\xi), b^*_n(\nu) = \nu/\text{Var}(\xi)$, and the oracles of the forms (6) and (9) coincide.

Finally, we note that existing nonparametric empirical Bayes estimators, such as the semiparametric estimator by Xie, Kou, and Brown (2012) and the nonparametric method by Jiang and Zhang (2010), target the best predictor $g(Y, A)$ of $\xi$ where $g$ is restricted to some nonparametric class of functions. While the optimal $g$ may indeed be a nonlinear function of $Y$, these methods implicitly assume independence between $\xi$ and $A$, and might still suffer from the gap between the optimal predictor $g(Y, A)$ assuming independence, and the true Bayes rule, namely, $E(\xi|Y, A)$. Therefore, in some cases the oracle rule of the form (9) might still have smaller risk than the oracle choice of $g$ computed assuming independence between $\xi$ and $A$.

3. Group-linear Shrinkage Methods

Let $X$, $\theta$ and $V$ be as in (1). The spherically symmetric estimator in the following lemma will serve as a building block for our group-linear estimator. We remark that a version of the estimator in the lemma below that shrinks toward a known mean, and sufficient conditions for its minimaxity, appear, in a slightly more general form, in Lehmann and Casella (1998, Theorem 5.7; although there are some typos) and reviewed by Tan (2015). Bock (1975) and Brown (1975, Theorem 3) independently obtained these conditions earlier as sufficient for the existence of a minimax estimator.

Lemma 1. Let $\widehat{\theta}^{\nu}$ be an estimator given by $\widehat{\theta}^{\nu}_i = X_i$ if $n = 1$, and otherwise

$$\widehat{\theta}^{\nu}_i = X_i - b(X_i - \bar{X}), \quad \widehat{\nu} = \min\{1, c_n\nu/\bar{\nu}^2\}, \quad (15)$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n, \quad \bar{\nu} = \sum_{i=1}^{n} V_i/n, \quad \nu^2 = \sum_{i=1}^{n} X_i - \bar{X}\nu^2/(n - 1)$ and $c_n$ is a positive constant. Let $V_{\max} = \max_{i \leq n} V_i$ and

$$c_n^* = \min\{(n - 3) - 2 (V_{\max}/\bar{\nu} - 1)/(n - 1)\}, \quad = (1 - 2 (V_{\max}/\bar{\nu} - 1))/(n - 1)\}. \quad(16)$$

Then for $0 \leq c_n \leq 2c_n^*$

$$\frac{1}{n} \sum_{i=1}^{n} E(\widehat{\theta}_i - \theta_i)^2 \leq \bar{\nu}[1 - (1 - 1/n)$$

$$\times E\{(2c_n^* - c_n)\widehat{\nu} + (2 - 2c_n^* + c_n - s_n^2/\bar{\nu}^2)\widehat{\theta}_i|\nu \leq c_n\}] \leq \bar{\nu}. \quad (16)$$
Remarks:

1. The main reason for using $X$ is analytical simplicity. When $\theta_i$ are all equal, the MLE of the common mean is the weighted least-squares estimate $(\sum_{i=1}^{n} X_i/V_i)/(\sum_{i=1}^{n} 1/V_i)$.

2. In (16) note that when $s^2_n/\sqrt{V} \geq c_n$, $(2c_n^2 - c_n)\bar{b} = (2c_n^2 - c_n)c_n/\sqrt{s^2_n}$ attains maximum at $c_n = c_n^*$. In the homoscedastic case $V_{\text{max}} = \sqrt{V}$ and $c_n^* = (n - 3)/(n - 1)$ is the usual constant for the James–Stein estimator that shrinks toward the sample mean. In the heteroscedastic case, for a version of the estimator above that shrinks toward zero, a sufficient condition for minimaxity appears in Tan (2015) as $0 \leq c_n \leq 2(1 - 2(V_{\text{max}}/\sqrt{V})/n)$. This is consistent with Lemma 1.

3. For one-way unbalanced ANOVA, $V_i = \sigma^2/n_i$ where $\sigma^2$ is the error variance and $n_i$ is the number of observations for the $i$th subpopulation. Suppose that $\sigma^2$ is unknown and that we have an unbiased estimator $\hat{\sigma}^2 = S_k/k$ of $\sigma^2$ independent of the observations, where $S_k/\sigma^2 \sim \chi^2_k$. Then if we replace the $V_i$ in the lemma with the corresponding estimates $\hat{V}_i = \hat{\sigma}^2/n_i$, the same conclusion still holds with $0 \leq c_n(1 + 2/k) \leq 2c_n^*$.

We are now ready to introduce an empirical Bayes estimator, which employs the spherically symmetric estimator of Lemma 1 to mimic the oracle rule $\hat{\theta}^{*,*}$. When the number of distinct values $V_i$ is very small compared to $n_i$, a natural competitor of $\hat{\theta}^{*,*}$ is obtained by applying a James–Stein estimator separately to each group of homoscedastic observations. Under appropriate conditions, this estimator asymptotically approaches the oracle risk (14). The situation in the general heteroscedastic problem, when the number of distinct values $V_i$ is not very small compared to $n_i$, is not as obvious; still, the expression for the optimal function $a^*$ and $b^*$ in (13) suggests grouping together observations with similar variances $V_i$, and then applying a spherically symmetric estimator separately to each group.

Block-linear shrinkage has been suggested before for the homoscedastic case by Cai (1999) in the context of wavelet estimation. However, the estimator by Cai (1999) is motivated from an entirely different perspective, and addresses a very different oracle rule (itself a blockwise rule) from the oracle associated with our procedure. Still for homoscedastic observations, Ma, Foster, and Stine (2015) proposed a block-linear empirical Bayes estimator with shrinkage factors that are increasing in magnitude, when the order of the variances of $\theta_i$ is assumed to be known. For the heteroscedastic case, Tan (2014) commented briefly that block shrinkage methods building on a minimax estimator can be considered to allow different shrinkage patterns for observations with different sampling variances; this is very much in line with the approach pursued in the current article.

Definition 1 (Group-linear Empirical Bayes Estimator for a Heteroscedastic Mean). Let $I_1, \ldots, I_m$ be disjoint intervals. For $k = 1, \ldots, m$ denote

$$I_k = \{ i : V_i \in J_k \}, \quad n_k = |I_k|, \quad V_k = \sum_{i \in I_k} V_i / n_k,$$

$$X_k = \sum_{i \in I_k} X_i / n_k, \quad S^2_k = \sum_{i \in I_k} (X_i - X_k)^2 / n_k \vee 2 - 1.$$

Define a corresponding group-linear estimator $\hat{\theta}^{GL}$ component-wise by

$$\hat{\theta}^{GL}_k = \begin{cases} X_i - \min (1, c_k V_k / S_k^2)(X_i - X_k), & i \in I_k \\ X_i, & \text{otherwise} \end{cases} \quad (17)$$

and note that $\hat{\theta}_i = X_i$ when $V_i \notin \cup_{k=1}^{m} I_k$ or $V_i \in I_k$ for some $k$ with $c_k = 0$.

Theorem 1. For $\hat{\theta} = \hat{\theta}^{GL}$ in Definition 1 with $c_k = (1 - 2(\max_{i \in I_k} V_i / V_k))/(n_k - 1)$, the following holds:

1. Under the Gaussian model (1) with deterministic $(\theta_i, V_i)$, $i = 1, \ldots, n$, the risk of $\hat{\theta}$ is no greater than that of the naive estimator $X$ and therefore $\theta$ is minimax

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}((\hat{\theta}_i - \theta_i)^2) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i - \theta_i)^2 = \frac{1}{n} \sum_{i=1}^{n} V_i = \sqrt{V}.$$

(18)

2. Let $(X_i, \theta_i, V_i)$, $i = 1, \ldots, n$, be iid vectors from any fixed (with respect to $n$) population satisfying (1). Let $(Y, \xi, A)$ be defined by (8); $r(a, b)$ as defined in (10); and $a^*$ and $b^*$ as defined in (13). Then

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2 | V_i] \leq \frac{1}{n} \sum_{i=1}^{n} r(a^*, b^* | V_i) + o(1)$$

(19)

with $V = (V_1, \ldots, V_n)$ and for any sequence $V_1, V_2, \ldots$ such that the following holds:

With $|J|$ being the length of interval $J$,

$$\max_{1 \leq k \leq m} |J_k| \to 0, \quad \min_{1 \leq k \leq m} n_k \to \infty,$$

$a^*(u)$, $b^*(u)$ are uniformly continuous,

$$\limsup_{n \to \infty} \frac{\sum_{i=1}^{n} V_i}{n} < \infty, \quad \limsup_{n \to \infty} \frac{\sum_{i=1}^{n} V_i (\cup_{1 \leq i \leq n})}{n} = 0$$

(20)

Remarks on the second part of the theorem:

1. Note that when $(X_i, \theta_i, V_i)$ are iid, then each triple is distributed as $(Y, \xi, A)$. We assumed that the "population" distribution $(Y, \xi, A)$ itself does not depend on $n$ (in which case $r(a, b)$ and $a^*$, $b^*$ indeed do not depend on $n$). A similar statement would still hold when the distribution of $(Y, \xi, A)$ depends on $n$, under the conditions that $[a^*_n]$, $[b^*_n]$ are equicontinuous and $[a^*_n]$ is uniformly bounded for any given finite interval. Although not considered here, an analog of the second part of the theorem could be stated for the nonrandom situation, $X_i|\{\theta_i, V_i\} \sim N(\theta_i, V_i)$, $1 \leq i \leq n$ with deterministic $\theta_i$ and $V_i$. In this case, suppose that the empirical joint distribution $G_n$ of $\{\theta_i, V_i\}$, $1 \leq i \leq n$ has a limiting distribution $G$. Then, if we define the risk for candidates $a_n$, $b_n$ to be computed with respect to $G$, our estimator enjoys $r(\hat{a}_n, \hat{b}_n) \to r(a^*, b^*)$ under appropriate conditions on $a^*$, $b^*$. 
2. The continuity of shrinkage factor and location $b^*(v)$, $a^*(v)$ allows to borrow strength from neighboring observations with similar variances. To asymptotically mimic the performance of the oracle rule, $\max_{1 \leq k \leq m} |\hat{J}_k| \to 0$, $\min_{1 \leq k \leq m} n_k \to \infty$ are necessary wherever shrinkage is needed. The only intrinsic assumption is $\limsup_{n \to \infty} \sum_{i=1}^n V_i/n < \infty$, essentially "equivalent" to bounded expectation of $A$. It ensures that $\max_{1 \leq k \leq m} |\hat{J}_k| \to 0$, $\min_{1 \leq k \leq m} n_k \to \infty$ are satisfied when $\cup_{k=1}^m \hat{J}_k$ are chosen to cover most of the observations, and at the same time $\limsup_{n \to \infty} \sum_{i=1}^n |V_i|/n = 0$, which takes care of the remaining observations (large or isolated $V_i$), and guarantees that their contribution to the risk is negligible.

3. A statement on Bayes risk, when expectation is taken over $V$ in (19), can be obtained in a similar way by replacing the conditions on $V$ with bounded expectation of the random variable $A$. We skip this for simplicity.

For the iid situation of the second part of Theorem 1, the case $r(a^*, b^*) = 0$ corresponds to $\xi = a^*(A)$, a deterministic function of $A$ (equivalently, $b^*(A) \equiv 1$). In this case, the precision in estimating the function $a^*$ is crucial, and calls for a sharper result than (19) regarding the rate of convergence of the excess risk. Noting that, trivially, $\xi = a^*(A)$ implies that $E(\xi|A = \nu) = a^*(\nu)$,

$$X_i|V_i \sim N(a^*(V_i), V_i)$$

is a nonparametric regression model, that is, $\theta_i$ is a deterministic measurable function of $V_i$. In this case, the rate of convergence in (19) depends primarily on the smoothness of the function $a^*(\nu)$. In the homoscedastic case, the smoothing feature of the James–Stein estimator was studied by Li and Hwang (1984). The following theorem states that the group-linear estimator attains the optimal convergence rate under a Lipschitz condition, at least when $A$ is bounded.

**Theorem 2.** Let $(X_i, \theta_i, V_i)$, $i = 1, \ldots, n$, be iid vectors from a population satisfying (1). If $r(a^*, b^*) = 0$ and $a^*(\cdot)$ is $L$-Lipschitz continuous, then the group linear estimator in Definition 1 with equal block size $|J_k| = |J| = (10V_2/n)^{\frac{1}{3}}$ and $c_n = c_n^*$ attains the optimal nonparametric rate of convergence

$$\frac{1}{n} \sum_{i=1}^n E[(\hat{\theta}_i - \theta_i)^2|V_i] \leq 2 \left( \frac{10V_2^{\frac{1}{2}}\sqrt{L}}{n} \right)^{\frac{1}{3}}$$

for any deterministic sequence $V = (V_1, \ldots, V_n)$.

For the asymptotic results in Theorems 1 and 2 to hold, it is enough to choose bins $J_k$ of equal length $|J| = (10V_2/n)^{\frac{1}{3}}$. However, in realistic situations, where $n$ is some fixed number, other strategies for binning observations according to the $V_i$ might be more sensible. For example, by Lemma 1 and the first remark that follows it, bins that keep $(\max|V_i|: i \in J_k)/\sqrt{k}$ (rather than $\max|V_i|: i \in J_k| - \min|V_i|: i \in J_k|$) approximately fixed may be more appropriate. Hence, we propose to bin observations to windows of equal lengths in $\log(V_i)$ instead of $V_i$. Furthermore, instead of the constant multiplying $n^{-\frac{1}{3}}$ in $|J|$, which may be suitable when the $V_i \in (0, 1]$, we suggest in general to fix the number of bins to $n^{1/3}$, that is, divide $\log(V_i)$ to bins of equal length $[\max(\log(V_i)| - \min(\log(V_i)|)/n^{1/3}$. On a finer scale, for a given choice of $|J_k|$, there is also the question whether any two groups should be combined together, and the shrinkage factors adjusted accordingly; this issue arises even in the homoscedastic case (Efron and Morris 1973a). Note that, trivially, minimaxity is preserved when the values of $V_i$, but not $X_i$, are used to choose the bins $J_k$.

As for performance of the group-linear estimator for fixed $n$, some situations are certainly harder than others. In the best scenario where the variances are clustered at a fixed finite set of possible values, the method is expected to work very well with fast convergence in (19). Otherwise, the method is expected to work reasonably well in the sense of (19) when $\max V_i/ \min V_i$ is not too large, whether the distribution of $V_i$ is continuous or not, because the large clusters will benefit from shrinkage and small clusters will have small total contribution to the risk due to minimaxity within each group. Still, the difference between the two cases could be nontrivial in finite samples. In the third and worst-case scenario, the sequence of variances is rapidly increasing so that the benefit of grouping is small for a large fraction of relatively large variances. This could also happen when the variances are small, as the risk ratio between the group and naive estimators depends only on the ratio $V_i/V_{\max}$.

**4. Simulation Study**

In this section, we carry out a simulation study using the examples by Xie, Kou, and Brown (2012), and compare the performance of our group-linear estimator to the methods proposed in their work. In each example, we draw $n$ iid triples $(X_i, \theta_i, V_i) \sim (Y, \xi, A)$, $i = 1, \ldots, n$, from a probability distribution satisfying (1). If $r(a^*, b^*) = 0$ and $a^*(\cdot)$ is $L$-Lipschitz continuous, then the group linear estimator in Definition 1 with equal block size $|J_k| = |J| = (10V_2/n)^{\frac{1}{3}}$ and $c_n = c_n^*$ attains the optimal nonparametric rate of convergence

$$\frac{1}{n} \sum_{i=1}^n E[(\hat{\theta}_i - \theta_i)^2|V_i] \leq 2 \left( \frac{10V_2^{\frac{1}{2}}\sqrt{L}}{n} \right)^{\frac{1}{3}}$$

for any deterministic sequence $V = (V_1, \ldots, V_n)$.

For the asymptotic results in Theorems 1 and 2 to hold, it is enough to choose bins $J_k$ of equal length $|J| = (10V_2/n)^{\frac{1}{3}}$. However, in realistic situations, where $n$ is some fixed number, other strategies for binning observations according to the $V_i$ might be more sensible. For example, by Lemma 1 and the first remark that follows it, bins that keep $(\max|V_i|: i \in J_k)/\sqrt{k}$ (rather than $\max|V_i|: i \in J_k| - \min|V_i|: i \in J_k|$) approximately fixed may be more appropriate. Hence, we propose to bin observations to windows of equal lengths in $\log(V_i)$ instead of $V_i$. Furthermore, instead of the constant multiplying $n^{-\frac{1}{3}}$ in $|J|$, which may be suitable when the $V_i \in (0, 1]$, we suggest in general to fix the number of bins to $n^{1/3}$, that is, divide $\log(V_i)$ to bins of equal length $[\max(\log(V_i)| - \min(\log(V_i)|)/n^{1/3}$. On a finer scale, for a given choice of $|J_k|$, there is also the question whether any two groups should be combined together, and the shrinkage factors adjusted accordingly; this issue arises even in the homoscedastic case (Efron and Morris 1973a). Note that, trivially, minimaxity is preserved when the values of $V_i$, but not $X_i$, are used to choose the bins $J_k$.

As for performance of the group-linear estimator for fixed $n$, some situations are certainly harder than others. In the best scenario where the variances are clustered at a fixed finite set of possible values, the method is expected to work very well with fast convergence in (19). Otherwise, the method is expected to work reasonably well in the sense of (19) when $\max V_i/ \min V_i$ is not too large, whether the distribution of $V_i$ is continuous or not, because the large clusters will benefit from shrinkage and small clusters will have small total contribution to the risk due to minimaxity within each group. Still, the difference between the two cases could be nontrivial in finite samples. In the third and worst-case scenario, the sequence of variances is rapidly increasing so that the benefit of grouping is small for a large fraction of relatively large variances. This could also happen when the variances are small, as the risk ratio between the group and naive estimators depends only on the ratio $V_i/V_{\max}$.
Figure 1. Estimated risk for various estimators vs. number of observations.

\[
r(\mu^*, \gamma^*) = \min_{\mu, \gamma \in \mathbb{R} : \gamma \geq 0} \mathbb{E} \left\{ \left[ \frac{Y}{A + \gamma} (Y - \mu) - \xi \right]^2 \right\}
\]

(23)

and

\[
r(a^*, b^*) = \min_{a(\cdot), b(\cdot) : a(v) \geq 0 \ \forall v} \mathbb{E} \left\{ \left[ Y - b(A) (Y - a(A)) - \xi \right]^2 \right\}
\]

(24)

corresponding to the optimal rule in the parametric family of estimators considered by Xie, Kou, and Brown (2012, labeled “XKB oracle” in the legend of Figure 1), and to the optimal linear-in-\(x\) rule of Section 2, respectively. Note that \(\mu^*\) and \(\gamma^*\) are numbers, whereas \(a^*\) and \(b^*\) are functions. Table 1 displays the oracle shrinkage locations and shrinkage factors corresponding to (23) and (24); note that \(v/(v + \gamma^*)\) is strictly increasing in \(v\), while \(b^*(v)\) is not necessarily.

Figure 1 shows the average loss across the \(N = 10,000\) repetitions for the parametric SURE, semiparametric SURE and the group-linear estimators, plotted against the different values of \(n\). The horizontal line corresponds to \(r(\mu^*, \gamma^*)\). The general picture arising from the simulation examples is consistent with our expectation that the limiting risk of the group-linear estimator is smaller than that of both the parametric SURE estimator, as \(r(a^*, b^*) \leq r(\mu^*, \gamma^*)\), and the semiparametric SURE estimator, as \(r(a^*, b^*) \leq \inf_{b(\cdot)} \{r(a, b) : b(v) \text{ monotone increasing in } v\}\). For moderate \(n\), whenever \(\xi\) and \(A\) are independent, the SURE estimators are appropriate and achieve smaller risk. By contrast, the situations where \(\xi\) and \(A\) are dependent are handled best by the group-linear estimator, which indeed achieves significantly smaller risk than both SURE estimators.

In example (a) (7.1 of Xie, Kou, and Brown 2012) \(A \sim \text{Unif}(0.1, 1)\) and \(\xi \sim N(0, 1)\), independently. In this case, the linear Bayes rule is of the form (6) and, in particular, the functions \(a^*\) and \(b^*\) are constant in \(v\). The parametric SURE...
Table 1. Oracle shrinkage locations and shrinkage factors, $\mu^*, \nu/(\nu + \gamma^*)$ and $a^*(\nu), b^*(\nu)$, corresponding to the family of estimators by Xie, Kou, and Brown (equation (23)) and to the family of estimators that are linear in $Y$ (equation (24)). Columns correspond to simulation examples (a)–(f). Values of $\mu^*, \gamma^*$ for each example are from Xie, Kou, and Brown (2012).

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
<th>(f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu^<em>, \frac{b}{\nu + \gamma^</em>}$</td>
<td>0, $\frac{b}{\nu + 1}$</td>
<td>0.5, $\frac{b}{\nu + 0.083}$</td>
<td>0.6, $\frac{b}{\nu + 0.078}$</td>
<td>0.13, $\frac{b}{\nu + 0.0032}$</td>
<td>0.15, $\frac{b}{\nu + 0.84}$</td>
<td>0.6, $\frac{b}{\nu + 0.078}$</td>
</tr>
<tr>
<td>$a^<em>(\nu), b^</em>(\nu)$</td>
<td>0, $\frac{b}{\nu + 1}$</td>
<td>0, $\frac{b}{\nu + 1}$</td>
<td>$\nu, 0$</td>
<td>$\nu, 0$</td>
<td>$2\delta_{(\nu=0.1)}(\nu)$, 0.5</td>
<td>$\nu, 0$</td>
</tr>
</tbody>
</table>

The estimator is therefore appropriate, and it performs best, requiring estimation of only two hyperparameters. The group-linear estimator and the semiparametric SURE perform comparably across values of $n$. Here, $r(\mu^*, \gamma^*)$, $r(a^*, b^*)$ and the limiting risks of the parametric SURE and the group-linear estimator, are all equal ($\approx 0.3357$). In example (b), (7.2 of Xie, Kou, and Brown 2012), $A \sim \text{Unif}(0.1, 1)$ and $\xi \sim N(0, 1)$, independently. This situation is not very different from the first example when it comes to comparing the SURE estimators to the group-linear, since the functions $a^*$ and $b^*$ are constant in $\nu$ as long as $\xi$ and $A$ are independent. The risk of the group-linear approaches the oracle risk ($\approx 0.0697$), but here the semiparametric SURE estimator seems to do a little better, perhaps in part because it (correctly) shrinks all data points toward exactly the same location.

The third example (c) (7.3 of Xie, Kou, and Brown 2012) takes $A \sim \text{Unif}(0.1, 1)$, $\xi = A$. Here, $\xi$ and $A$ are strongly dependent, and indeed the gap between the two oracle risks, $r(\mu^*, \gamma^*) \approx 0.0540$ and $r(a^*, b^*) = 0$, is material. The advantage of the group-linear estimator over the SURE estimators is seen already for moderate values of $n$. Although it is hard to tell from the figure, the limiting risk of the semiparametric SURE is slightly smaller than that of the parametric SURE, because of the improved capability of the semiparametric oracle to accommodate the dependence between $\xi$ and $A$. In the fourth case (d) (7.3 of Xie, Kou, and Brown 2012), we take $A \sim \text{Inv-Chi}_0^*, \xi = A$. $\xi$ is still a deterministic function of $A$, but it takes larger values of $n$ for the group-linear estimator to outperform the SURE estimators. This is not seen before $n = 500$, which seems to be a consequence of the non-uniform distribution of the $V_i$, and only partially mitigated by binning according to $\log(V_i)$.

Example (e) (7.5 of Xie, Kou, and Brown 2012) reflects grouping: $A$ equals 0.1 or 0.5 with equal probability; $\xi(A = 0.1) \sim N(2, 0.1)$ and $\xi(A = 0.5) \sim N(0, 0.5)$. In each of the two variance groups, the group-linear estimator reduces to a (positive-part) James–Stein estimator, and performs significantly better than the SURE estimators. While not plotted in the figure, the other semiparametric SURE estimator by Xie, Kou, and Brown (2012), which uses a SURE criterion to choose also the shrinkage location, achieves significantly smaller risk than the SURE estimators considered here; still, its limiting risk is 16% higher than that of the group-linear.

Finally, in (f) (7.6 of Xie, Kou, and Brown 2012) $A \sim \text{Unif}(0.1, 1)$, $\xi = A$ and $Y|A \sim \text{Unif}(\xi - \sqrt{3A}, \xi + \sqrt{3A})$, violating the normality assumption for the data. The group-linear estimator is again seen to outperform the SURE estimators starting at relatively small values of $n$, and its risk still tends to the oracle risk $r(a^*, b^*) = 0$. By contrast, the risk of the parametric SURE estimator approaches $r(\mu^*, \gamma^*) = 0.054$. The semiparametric SURE estimator does just a little better, its risk approaching $\approx 0.0423$.

5. Real-Data Example

We now turn to a real-data example to test our group-linear methods. We use the popular baseball dataset from Brown (2008), which contains batting records for all Major League baseball players in the 2005 season. As in Brown (2008), the entire season is split into two periods, and the task is to predict the batting averages of individual players in the second half-season based on records from the first half-season. Denoting by $H_{ji}$ the number of hits and by $N_{ji}$ the number of at-bats for player $i$ in period $j$ of the season, it is assumed that

$$H_{ji} \sim \text{Bin}(N_{ji}, p_i), \quad j = 1, 2, \ i = 1, \ldots, P_j. \quad (25)$$

As suggested in Brown (2008), a variance-stabilizing transformation is first applied, $X_{ji} = \arcsin((H_{ji} + 1/4)/(N_{ji} + 1/2))^{1/2}$, resulting in

$$X_{ji} \sim N(\theta_i, 1/(4N_{ji})) \quad \theta_i = \arcsin(p_i)$$

and $\{X_{1i, 1i, \ldots, P_i} : i = 1, \ldots, P_j\}$ are then used to estimate the means $\theta_i$. We should remark that there is no reason for using this transformation, and for focusing on estimating $\theta_i$ instead of $p_i$, other than making the data (approximately) fit the heteroscedastic normal model (note that the variance of the obvious statistic $H_{ji}/N_{ji}$ depends explicitly on $p_i$, and therefore is not suitable).

Indeed, one might object to analyzing the baseball data using a normal model instead of using the binomial model (25) directly

![Figure 2. Shrinkage factors vs. at-bats (all players). Vertical axis shows 1 – shrinkage factor. For the parametric SURE estimator $\hat{\theta}$ this is $\hat{\gamma}/(1/(4N_{ji}) + \hat{\gamma});$ for semiparametric SURE $\hat{\theta}^G$ and for group-linear, this is $1 - \hat{\delta}(1/(4N_{ji}))$ with $\hat{\delta}$ of (22) or (77), respectively. The curves for both SURE estimators (dotted and broken lines) are non-decreasing in $N_{ji}$ by construction. Shrinkage factors are not constrained to be monotone for the group-linear.](image-url)
Table 2. Prediction errors of transformed batting averages. For the five estimators at the bottom of the table, numbers in parentheses are estimated TSE for permuted data.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Pitchers</th>
<th>Non-pitchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Grand mean</td>
<td>0.852</td>
<td>0.127</td>
<td>0.378</td>
</tr>
<tr>
<td>Nonparametric EB</td>
<td>0.508</td>
<td>0.212</td>
<td>0.372</td>
</tr>
<tr>
<td>Binomial mixture</td>
<td>0.588</td>
<td>0.156</td>
<td>0.314</td>
</tr>
<tr>
<td>Weighted Least Squares</td>
<td>1.07</td>
<td>0.127</td>
<td>0.468</td>
</tr>
<tr>
<td>Weighted nonparametric</td>
<td>0.306</td>
<td>0.173</td>
<td>0.326</td>
</tr>
<tr>
<td>MLE</td>
<td>0.337</td>
<td>0.087</td>
<td>0.290</td>
</tr>
<tr>
<td>Weighted nonparametric (AB)</td>
<td>0.301</td>
<td>0.141</td>
<td>0.261</td>
</tr>
<tr>
<td>James–Stein</td>
<td>0.535 (0.543)</td>
<td>0.165 (0.239)</td>
<td>0.348 (0.234)</td>
</tr>
<tr>
<td>SURE (\hat{\theta}^M)</td>
<td>0.421 (0.484)</td>
<td>0.123 (0.211)</td>
<td>0.289 (0.265)</td>
</tr>
<tr>
<td>SURE (\hat{\theta}^{SG})</td>
<td>0.408 (0.468)</td>
<td>0.091 (0.169)</td>
<td>0.261 (0.219)</td>
</tr>
<tr>
<td>Group-linear (\hat{\theta}^{GL})</td>
<td>0.302 (0.280)</td>
<td>0.178 (0.244)</td>
<td>0.325 (0.175)</td>
</tr>
<tr>
<td>Group-linear (dynamic)</td>
<td>0.288 (0.276)</td>
<td>0.168 (0.283)</td>
<td>0.349 (0.175)</td>
</tr>
</tbody>
</table>

For the group-linear estimator, in addition to the plain estimator \(\hat{\theta}^{GL}\) that uses \(k = \lfloor n^{1/3} \rfloor\) equal length bins on \(\log(1/4N)\) (as in the simulation study), we considered a data-dependent strategy for binning. The estimator labeled “dynamic” in Table 2 chooses, among all partitions of the data into contiguous bins containing no more than \(\lfloor n^{2/3} \rfloor\) observations each, the partition which minimizes an unbiased estimate of the risk of the corresponding group-linear estimator. This can be viewed as an extension of the plain version, which for uniformly spaced data would put \(\sim n^{2/3}\) observations in each of \(\lfloor n^{1/3} \rfloor\) bins. Our implementation uses dynamic programming (code available online at https://github.com/Ma2Zhuang/grouplinear). We remark that using the observed data in forming the bins may lead to loss of minimaxity of the group-linear estimator. Nevertheless, we find it appropriate to explore such strategies when applying the estimator to real data.

Both versions of the group-linear estimator perform well in predicting batting averages for all players relative to the other estimators. As discussed by Brown (2008), nonconformity to the hierarchical normal–normal model on which most parametric empirical Bayes estimators are based, is evident in the data: first of all, non-pitchers tend to have better batting averages than pitchers, hence, it is more plausible that the \(\theta_i\) come from a mixture of two distributions. Second, players with higher batting averages tend to play more, suggesting that there is statistical dependence between the true means, \(\theta_i\), and the sampling variances of \(X_{ji}(\alpha 1/N_{ji})\). While the nonparametric MLE method handles well nonnormality in the “prior” distribution of the \(\theta_i\), its derivation still assumes statistical independence between the true means and the sampling variances. The group-linear estimator achieves good performance in this situation because it is able to accommodate this dependence between the true means and the sampling variances.

When analyzing pitchers and non-pitchers separately on the original data, the SURE methods achieve dramatic improvement and outperform the group-linear estimators by a significant amount. However, the results are quite different for shuffled data. The difference is seen most prominently for non-pitchers: when actual second half records are used, the group-linear incurs higher prediction error as compared to the semiparametric SURE estimator (0.325 vs. 0.261); but the opposite emerges for shuffled data (0.175 vs. 0.219). For pitchers only, the estimators by Xie, Kou, and Brown (2012) outperformed the group-linear in both the standard analysis and the permutation analysis. This is reasonable as the association between the number of at-bats and the true ability is expected to be weaker than within non-pitchers.

Figure 2 displays shrinkage factors (in fact, \(1 - \text{shrinkage factor}\)) versus number of at-bats (all players) for using the SURE criterion to choose both the shrinkage and the location parameter; and \(\hat{\theta}^{SC}\) is the semiparametric SURE estimator by Xie, Kou, and Brown (2012) that shrinks toward the grand mean. Also included in the table are the nonparametric shrinkage methods of Brown and Greenshtein (2009), the weighted least-squares estimator; the nonparametric maximum likelihood estimators of Jiang and Zhang (2009, 2010) (with and without number of at-bats as covariate) and the binomial mixture estimator of Muralidharan (2010).

As in Muralidharan (2010). Our only response is that the purpose of our analysis is primarily to illustrate the possible advantages of the group-linear estimator—and more generally, of methods that can accommodate statistical dependence between the means and the known variances—in the heteroscedastic normal problem.

To measure the performance of an estimator \(\hat{\theta}\), we use the total squared error,

\[ \text{TSE}(\hat{\theta}) = \sum_i \left[ \frac{(X_{2i} - \hat{\theta}_i)^2}{4N_{2i}} \right], \]

proposed by Brown (2008) as an (approximately) unbiased estimator of the risk of \(\hat{\theta}\). Following Brown (2008), only players with at least 11 at-bats in the first half-season are considered in the estimation process, and only players with at least 11 at-bats in both half-seasons are considered in the validation process, namely, when evaluating the TSE. To support our comparison, in addition to the analysis for the original data, we present an analysis under a permutation of the order in which successful hits appear throughout the entire season: for each player we draw the number of hits in the \(N_{ji}\) at-bats of the first period from a hypergeometric distribution, \(\mathcal{HG}(N_{1i} + N_{2i}, H_{1i} + H_{2i}, N_{ji})\). Under the binomial model (25), this amounts to resampling \((H_{1i}, H_{2i})\) conditional on the sufficient statistic \(H_{1i} + H_{2i}\). In the permutation analysis, we concentrate on the two SURE methods of Xie, Kou, and Brown (2012), which we consider as the main competitors of our method; the extended James–Stein estimator; and the group-linear estimators.

Table 2 shows TSE for various estimators reported in Table 2 of Xie, Kou, and Brown (2012), when applied separately to all players, pitchers only, and non-pitchers only. The values displayed are fractions of the TSE of the naive estimator, which, in each of the cases (i)–(iii), simply predicts \(X_{ji}\) by \(X_{1i}\). Numbers in parentheses correspond to permuted data, and were computed as the average of the relative TSE over 1000 rounds of shuffling as described above. In the table, the Grand mean estimator uses the simple average of all \(X_{ji}\); the extended positive-part James–Stein estimator is given by \(\hat{\theta}^{PPJS} = \mu_{JS^+} + (1 - \sum_{k < \mu_{JS^+}}) (X_k - \mu_{JS^+})\) where \(\mu_{JS^+} = (\sum j X_j / V_j) / (\sum j / V_j)\); \(\hat{\theta}^M\) is the parametric empirical Bayes estimator by Xie, Kou, and Brown (2012).
the group-linear estimator and the two SURE estimators of Xie, Kou, and Brown (2012), in some sense the two immediate competitors to the group-linear method.

6. Conclusion and Directions for Further Investigation

For a heteroscedastic normal vector, empirical Bayes estimators that have been suggested, both parametric and nonparametric, usually rely on a hierarchical model in which the parameter \( \theta_i \) has a prior distribution unrelated to the observed sampling variance \( V_i = \operatorname{var}(X_i|\theta_i) \). If separable estimators are considered, representing the heteroscedastic normal mean estimation problem as a compound decision problem, reveals that this model is generally inadequate to achieve risk reduction as compared to the naive estimator. Group-linear methods, on the other hand, are capable of capturing dependency between \( \theta_i \) and \( V_i \) and therefore are more appropriate for problems where it exists.

There is certainly room for further research. We point out a few possible directions for extending Theorems 1 and 2, that are outside the scope of the current work:

1. In the iid case, the distribution of the population \((Y, \xi, A)\) may be allowed to depend on \( n \) in such a way that \( r_n(a_n^*, b_n^*) \to 0 \) as \( n \to \infty \). In this case, the criterion (19) should be strengthened to the asymptotic ratio optimality criterion

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\hat{\theta}_i - \theta_i)^2 \leq (1 + o(1)) r_n(a_n^*, b_n^*) \tag{26}
\]

as \( n \to \infty \). As (26) does not hold uniformly for all joint distributions of \((Y, \xi, A)\), a reasonable target would be to prove (26) when \( r_n(a_n^*, b_n^*) \geq \eta_n \) for small \( \eta_n \) under suitable side conditions on the joint distribution of \((Y, \xi, A)\). This theory should include (19) as a special case and still maintain the property (18).

2. When \( a^*(n) \) satisfies an order \( \alpha \) smoothness condition with \( \alpha > 1 \), a higher-order estimate of \( a^*(V_i) \) needs to be used to achieve the optimal rate \( n^{-\alpha/(2n+1)} \) in the nonparametric regression case, \( r(a^*, b^*) = 0 \), for example, \( \hat{a}(V_i) \) with an estimated polynomial \( \hat{a}(\cdot) \) for each \( V_i \). We speculate that such a group-polynomial estimator might still always outperform the naive estimator \( \hat{\theta}_i = X_i \) under a somewhat stronger minimum sample size requirement.

Appendix: Proofs

Proof of Lemma 1. It suffices to consider \( 0 < c_n \leq 2c_n^* \). Let \( b(x) = \min(1, c_n \sqrt{x/n}) \) such that \( \overline{b} = b(s_n^*) \). Noting that \( (\partial / \partial X_i) s_n^* = 2(X_i - \overline{X})/(n - 1) \), Stein's lemma yields

\[
\mathbb{E}(X_i - \theta_i)(X_i - \overline{X})\hat{b} = V_i \mathbb{E}\left\{ (1 - 1/n)b(\hat{s}_n^*) + 2(X_i - \overline{X})\hat{b}(\hat{s}_n^*)/(n - 1) \right\}.
\]

By definition, \( 2V_i/(n - 1) \leq \overline{V}(1 - c_n^*) \) and \( x \hat{b}'(x) = -b(x)1_{\{b(x) < 1\}} \),

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i - \theta_i)(X_i - \overline{X})\hat{b} = V_i \mathbb{E}\left\{ (1 - 1/n)b(\hat{s}_n^*) + 2(X_i - \overline{X})\hat{b}(\hat{s}_n^*)/(n - 1) \right\}.
\]

The rest of this section, we define \( e_{ij} = \max_{j \neq i} \mathbb{E}(V_i, V_j, |a^*(V_i) - a^*(V_j)|, |b^*(V_i) - b^*(V_j)|, |g(\cdot)|, \operatorname{Var}(\xi|A = n) \) and \( h(n) = \mathbb{E}(\xi^2|A = n) \). Unless otherwise stated, all the expectations and variances in this section are conditional on \( V \).

Lemma 2 (Analysis within each block). Let \((X_i, \theta_i, V_i)_{i=1}^n\) be iid vectors drawn from some population \((Y, \xi, A)\) satisfying (11). If \( V_1, \ldots, V_n \in I \) for some interval \( J \) and \( \min_{1 \leq i \leq n} b^*(V_i) \geq \varepsilon, b^*(\overline{V}) \geq \varepsilon \) for some \( \varepsilon > 0 \). Then, the spherically symmetric shrinkage estimator defined in (17) with \( c_n = c_n^* \) satisfies

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\hat{\theta}_i - \theta_i)^2|V| \leq \frac{1}{n} \sum_{i=1}^{n} (a^*, b^*)|V_i| + \frac{7V_{\max}}{n \sqrt{V}} + \frac{\sqrt{\mathbb{E}[V]} + |J|}{\varepsilon^2} + \frac{1}{\varepsilon^2} + e_{ij} \tag{27}
\]

where \( V_{\max} = \max(V_1, \ldots, V_n) \) and \( \overline{V} = \sum_{i=1}^{n} V_i/n \).

Proof of Lemma 2. As in the proof of Lemma 1 with \( c_n = c_n^* \),

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\hat{\theta}_i - \theta_i)^2|V| \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i - \theta_i)(X_i - \overline{X})\hat{b} - \theta_i|V|^2 \]

\[
\leq \overline{V} + \frac{1}{n} \mathbb{E}(\overline{V}b(\hat{s}_n^*)) \times \left\{ \min(\hat{s}_n^*/\overline{V}, c_n^*)^2 - 2 + 2(1 - c_n^*)I_{\{\hat{s}_n^*/\overline{V} \leq c_n^*\}} \right\}.
\]
By definition, \( r(a^*, b^*)|V_i| = V_i(1 - b^*(V_i)) \) and \( \min(s^2_i/V, c^*_n) \leq c^*_n \leq 1. \) Then,
\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\widehat{V}_i - \theta_j]^2 |V| \leq \frac{1}{n} \sum_{i=1}^{n} r(a^*, b^*)|V_i| + \frac{1}{n} \sum_{i=1}^{n} b^*(V_i)V_i
\leq \left(1 - \frac{1}{n}\right) \nabla \mathbb{E}(\widehat{b}) + 2\nabla(1 - c^*_n)
\]

Observing that \( 0 \leq \widehat{b} \leq 1 \) and \( \nabla(1 - c^*_n) \leq 2\nabla_{\text{max}}/(n \vee 2 - 1), \)
\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\widehat{V}_i - \theta_j]^2 |V| \leq \frac{1}{n} \sum_{i=1}^{n} r(a^*, b^*)|V_i| + \frac{4\nabla_{\text{max}}}{n \vee 2 - 1} + \nabla/n
\]

where the last inequality is due to the uniformly continuity of \( b^*(\cdot) \).

Next we will bound \( \nabla(b^*(\nabla) - \nabla\widehat{b}) \). Following the definition of \( b^*(\cdot) \) and \( \widehat{b}, \)
\[
\nabla(b^*(\nabla) - \nabla\widehat{b}) = \nabla\mathbb{E}\{\nabla/\text{var}(Y|A = \nabla) - \text{var}(1, c^*_n/\nabla_s^*)\}
\]

and because \( \nabla/\text{var}(Y|A = \nabla) = \nabla/(\nabla + \text{var}(\xi|A = \nabla)) \leq 1, \)
\[
\nabla(b^*(\nabla) - \nabla\widehat{b}) \leq \nabla\mathbb{E}\{\nabla/\text{var}(Y|A = \nabla) - c^*_n/\nabla_s^*\}|_{\xi | \mathcal{V} \in \mathcal{V}}
\]
Finally, we will control \( \mathbb{E}[\text{var}(s_\theta^2)] \). Again, \( \mathbb{X}[V, \theta \sim N(\sum_{i=1}^n \theta_i/n, \sum_{i=1}^n V_i/n^2) \), hence

\[
\mathbb{E}\{\text{var}(s_\theta^2)\} = \frac{1}{(n \vee 2 - 1)^2} \mathbb{E}\left\{\text{var}\left(\sum_{i=1}^n X_i^2 - n \mathbb{X}^2(\theta)\right)\right\}
\leq \frac{2}{(n \vee 2 - 1)^2} \mathbb{E}\left\{\sum_{i=1}^n X_i^2(\theta) + \text{var}(n \mathbb{X}^2(\theta))\right\} = \frac{2}{(n \vee 2 - 1)^2} \sum_{i=1}^n \left(2V_i^2 + 4\theta_i^2 V_i + n^2(2\mathbb{V}^2/n^2 + 4\theta_i^2 \mathbb{V}/n)\right)
\]

By definition, \( h(\nu) = \mathbb{E}(\xi^2|A = \nu) \), and, noting that \( \nu \mathbb{X}^2 \leq \sum_{i=1}^n \theta_i^2 \),

\[
\mathbb{E}\{\text{var}(s_\theta^2)\} \leq \frac{4}{(n \vee 2 - 1)^2} \left(\sum_{i=1}^n V_i^2 + 2 \sum_{i=1}^n V_i h(V_i) + 2 \mathbb{V} \sum_{i=1}^n h(V_i)\right) \leq \frac{4}{(n \vee 2 - 1)^2} \left(\sum_{i=1}^n V_i^2 + 2 \sum_{i=1}^n (V_i + \mathbb{V}) h(V_i) + \mathbb{V}^2\right)
\]

(30)

Combining (29), (30), we have

\[
\mathbb{V}(b^*(\mathbb{V}) - \mathbb{V}) \leq \frac{2V_{\text{max}}}{n \vee 2 - 1} + \frac{\mathbb{V} \epsilon_j |J|}{\epsilon^2} + |J| + \epsilon_j^2 + \frac{2}{n \vee 2 - 1} \left(\sum_{i=1}^n V_i^2 + 2 \sum_{i=1}^n (V_i + \mathbb{V}) h(V_i) + \mathbb{V}^2\right)^{1/2}
\]

and therefore,

\[
\frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\hat{\theta}_i - \theta_i)^2] \mathbb{V} \leq \frac{1}{n} \sum_{i=1}^n r(a^*, b^*) |V_i| + \frac{7V_{\text{max}}}{n \vee 2 - 1} + \frac{\mathbb{V} \epsilon_j |J|}{\epsilon^2} + \epsilon_j^2 + \frac{2}{n \vee 2 - 1} \left(\sum_{i=1}^n V_i^2 + 2 \sum_{i=1}^n (V_i + \mathbb{V}) h(V_i) + \mathbb{V}^2\right)^{1/2}
\]

Proof of Theorem 1. The first part of Theorem 1 is trivial from Lemma 1. For the second part, it suffices to prove that for all \( \epsilon > 0 \), the excess risk is \( O(\epsilon) \) for large enough \( n \). Because the contribution to the normalized risk for observation outside \( U_{k,j}^{\text{m}} \) is \( \sum_{i=1}^n V_i |V_i|/n = o(1) \), we only need to consider the case where \( V_1 \leq i \leq n, V_i \in \bigcup_{k,j} U_{k,j}^{\text{m}} \). Without loss of generality, we can assume \( V_1 \leq k \leq m \), either \( J_k \subset [0, \epsilon) \) or \( J_k \subset (\epsilon, +\infty) \) because we can always reduce \( \epsilon \) such that this happens. Due to the assumption that \( \lim sup_{n \to \infty} \sum_{i=1}^n V_i/n < \infty \), we can also choose \( M_\epsilon \) large enough such that \( \sum_{i=1}^n V_i |V_i|/n \leq \epsilon \) and \( V_i \) with \( J_k \subset (\epsilon, +\infty) \), either \( J_k \subset (\epsilon, M_\epsilon) \) or \( J_k \subset (M_\epsilon, +\infty) \).

For the rest of the proof, we divide all the observations into four disjoint groups and handle them separately. Let \( V_k^* = \sum_{j \in J_k} V_j/n_k \) and define \( S_1 = \{k|1 \leq k \leq n, J_k \subset (0, \epsilon)\}, S_2 = \{k|1 \leq k \leq n, J_k \subset (\epsilon, M_\epsilon)\}, \min_{j \in J_k} b^*(V_j) \geq \epsilon, b^*(\mathbb{V}) \geq \epsilon\}, S_3 = \{k|1 \leq k \leq n, J_k \subset (\epsilon, M_\epsilon)\}, \min_{j \in J_k} b^*(V_j) < \epsilon \) or \( b^*(\mathbb{V}) \leq \epsilon\}, S_4 = \{k|1 \leq k \leq n, J_k \subset (M_\epsilon, +\infty)\} \).\)

Case i. For the small variance part, \( V_i \in (0, \epsilon) \), the contribution to the risk is negligible. Because the group linear shrinkage estimator dominate the MLE in each interval, then

\[
\frac{1}{n} \sum_{k \in S_1} \sum_{i \in J_k} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2 |V] \leq \sum_{k \in S_1} V_i/n \leq \sum_{k \in S_1} \epsilon/n \leq \epsilon
\]

Case ii. For moderate variance with large shrinkage factor, \( V_i \in (\epsilon, M_\epsilon) \) and \( b^*(V_i) \geq \epsilon \), shrinkage is necessary to mimic the performance of the oracle rule. Applying Lemma 2 to each interval \( J_k \) such that \( k \in S_2 \),

\[
\frac{1}{n} \sum_{k \in S_2} \sum_{i \in J_k} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2 |V] \leq \frac{1}{n} \sum_{k \in S_2} \sum_{i \in J_k} \epsilon_{\text{max}} \left|14(V_i^k + J_i)^{1/2} + 14V_i^k + \epsilon_i^2\right| \leq \frac{1}{n} \sum_{k \in S_2} \sum_{i \in J_k} \epsilon_{\text{max}} \left|14(V_i^k + J_i)^{1/2} + 14V_i^k + \epsilon_i^2\right| \leq \frac{1}{n} \sum_{k \in S_2} \sum_{i \in J_k} \epsilon_{\text{max}} \left|14(V_i^k + J_i)^{1/2} + 14V_i^k + \epsilon_i^2\right|
\]

Let \( |J|_{\text{max}} = \max_{1 \leq k \leq m} |J_k|, \epsilon_{\text{max}} = \max_{1 \leq k \leq m} \epsilon_{J_k} \). Using the fact that \( \max_{1 \leq k \leq m} \frac{n_k}{n \vee 2 - 1} \leq 2 \),

\[
\frac{1}{n} \sum_{k \in S_2} \sum_{i \in J_k} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2 |V] \leq \frac{1}{n} \sum_{k \in S_2} \sum_{i \in J_k} \epsilon_{\text{max}} \left|14(V_i^k + J_i)^{1/2} + 14V_i^k + \epsilon_i^2\right| \leq \frac{1}{n} \sum_{k \in S_2} \sum_{i \in J_k} \epsilon_{\text{max}} \left|14(V_i^k + J_i)^{1/2} + 14V_i^k + \epsilon_i^2\right|
\]

\[\forall k \in S_2, i \in J_k, V_i \leq M_\epsilon. \text{ Because } a^*(\nu) \text{ is uniformly continuous on } [0, M_\epsilon], \text{ there exists a constant } C_\epsilon \text{ depending only on } \epsilon \text{ such that } a^*(V_i) \leq C_\epsilon. \text{ Then,}
\]

\[h(V_i) = \mathbb{V}^2(V_i) + (\mathbb{V}^2(V_i) + (\mathbb{V}^2(V_i))^2) \leq \frac{V_i}{b^*(V_i)} - V_i + (a^*(V_i))^2 \leq \frac{M_\epsilon}{\epsilon} + C_\epsilon^2
\]

Therefore,

\[
\frac{1}{n} \sum_{k \in S_2} \sum_{i \in J_k} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2 |V]
\]
Since the proposed group linear shrinkage estimator dominates a
\[ \sum_{k \in S_2} \frac{n_k}{n} \leq \sqrt{|S_2|} \sum_{k \in S_2} n_k \leq \sqrt{|S_2|} n, \]
Further observe that \(|S_2| \leq m \leq \frac{n}{\min_{1 \leq k \leq m} n_k},\) then
\[ \frac{1}{n} \sum_{k \in S_1} \sum_{i \in I_k} E[(\hat{\theta}_i - \theta_i)^2 | V] \leq \frac{1}{n} \sum_{k \in S_1} \sum_{i \in I_k} r(a^*, b^* | V_i) + o(\varepsilon) \]

**Case ii.** For moderate variance with negligible shrinkage factor, \(V_i \in (\varepsilon, M_i)\) and \(n_{k \in I_k} b^*(V_i) \) or \(b^*(V) < \varepsilon.\) The uniform continuity of \(b^*(\cdot)\) implies that \(V_i \in I_k, \) \(b^*(V_i) \leq \varepsilon + \varepsilon_{\max}.\) By definition \(r(a^*, b^* | V_i) = V_i (1 - b^*(V_i)),\) then
\[ \frac{1}{n} \sum_{k \in S_1} \sum_{i \in I_k} E[(\hat{\theta}_i - \theta_i)^2 | V] \leq \frac{1}{n} \sum_{k \in S_1} \sum_{i \in I_k} r(a^*, b^* | V_i) + \nabla(\varepsilon + \varepsilon_{\max}) \]

Since the proposed group linear shrinkage estimator dominates MLE in each block,
\[ \frac{1}{n} \sum_{k \in S_1} \sum_{i \in I_k} E[(\hat{\theta}_i - \theta_i)^2 | V] \leq \frac{1}{n} \sum_{k \in S_1} \sum_{i \in I_k} r(a^*, b^* | V_i) + \nabla(\varepsilon + \varepsilon_{\max}) \]

**Case iv.** For the large variance part, \(V_i \in (M_i, +\infty),\) the contribution to the risk is also negligible. By definition of \(M_i,\)
\[ \frac{1}{n} \sum_{k \in S_1} \sum_{i \in I_k} E[(\hat{\theta}_i - \theta_i)^2 | V] \leq \sum_{k \in S_1} \sum_{i \in I_k} V_i/n = \sum_{i = 1}^n V_i | V_i \geq M_1 | n \leq \varepsilon. \]

Summing up the inequalities of all four cases
\[ \frac{1}{n} \sum_{i = 1}^n E[(\hat{\theta}_i - \theta_i)^2 | V] \leq \frac{1}{n} \sum_{i = 1}^n r(a^*, b^* | V_i) + (\nabla + 2) \varepsilon + o(\varepsilon) \],

which completes the proof by the assumption that \(\lim_{n \to \infty} \sum_{i = 1}^n V_i/n \leq 0 \)
Proof of Theorem 2. Applying Lemma 3 to each interval and using
\[ \frac{n_k}{n_{k}\sqrt{2-1}} \leq 2, \]
\[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2|V] \]
\leq \frac{1}{n} \sum_{k=1}^{m} \left( n_k |J_k| + 2V_k + 4\max \frac{n_k}{n_k \vee 2 - 1} \right)
\leq L|J|^2 + 10m\max V/n = L|J|^2 + 10V^2/(m|J|)
\]
Letting \( |J| = (\frac{10V^2}{nL})^\frac{1}{4} \), we have that \( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2|V] \leq 2(\frac{10V^2}{nL})^\frac{1}{4}. \)

\[ \square \]

Funding

National Science Foundation [DMS-1512084].

References


—— (2015), "Improved Minimax Estimation of a Multivariate Normal Mean Under Heteroscedasticity," Bernoulli, 21, 574–603. [2,3]