Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean

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

Abstract

The problem of estimating the mean of a normal vector with known but unequal variances introduces substantial difficulties that impair the adequacy of traditional empirical Bayes estimators. By taking a different approach, that treats the known variances as part of the random observations, we restore symmetry and thus the effectiveness of such methods. We suggest a group-linear empirical Bayes estimator, which collects observations with similar variances and applies a spherically symmetric estimator to each group separately. The proposed estimator is motivated by a new oracle rule which is stronger than the best linear rule, and thus provides a more ambitious benchmark than that considered in previous literature. Our estimator asymptotically achieves the new oracle risk (under appropriate conditions) and at the same time is minimax. The group-linear estimator is particularly advantageous in situations where the true means and observed variances are empirically dependent. To demonstrate the merits of the proposed methods in real applications, we analyze the baseball data used in Brown (2008), where the group-linear methods achieved the prediction error of the best nonparametric estimates that have been applied to the dataset, and significantly lower error than other parametric and semi-parametric empirical Bayes estimators.

Keywords: empirical Bayes, shrinkage estimator, heteroscedasticity, compound decision, asymptotic optimality

*Asaf Weinstein is at Stanford University, Stanford, CA 94305 (E- mail: asafw@stanford.edu). Zhuang Ma is at The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: zhuangma@wharton.upenn.edu). Lawrence D. Brown is Miers Busch Professor and Professor of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E- mail: lbrown@wharton.upenn.edu). Cun-Hui Zhang is Professor, Rutgers University, Piscataway NJ 08854 (E -mail:cunhui@stat.rutgers.edu)

1 Introduction

Let $\boldsymbol{X} = (X_1, ..., X_n)^T$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T$ and $\boldsymbol{V} = (V_1, ..., V_n)^T$, and suppose that

$$X_i|(\theta_i, V_i) \sim N(\theta_i, V_i) \tag{1}$$

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

$$L_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = n^{-1}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = n^{-1}\sum_{i=1}^{n}(\widehat{\theta}_i - \theta_i)^2. \tag{2}$$

Hence we assume that in addition to the random observations $X_1, ..., X_n$, the variances $V_1, ..., V_n$ are available. Allowing the values of V_i to be different from each other extends the applicability of the homoscedastic Gaussian mean problem to many realistic situations. A simple but common example is the design corresponding to a one-way homoscedastic Analysis of Variance with unequal cell counts; here X_i represents the mean of the n_i i.i.d. $N(\theta_i, \sigma^2)$ observations for the i-th sub-population, hence $V_i = \sigma^2/n_i$. More generally, if $Y \sim N_p(\boldsymbol{A}\boldsymbol{\beta}, \sigma^2 I)$ with a known design matrix $\boldsymbol{A}$, then estimating $\boldsymbol{\beta}$ under sum-of-squares loss is equivalent to estimating $\boldsymbol{\theta}$ in (1) where $n = rank(\boldsymbol{A})$ and X_i and V_i/σ^2 are determined by $\boldsymbol{A}$ (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 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 $\widehat{\boldsymbol{\theta}} = \boldsymbol{X}$ have been suggested that perform better, for fixed n or only asymptotically (under some conditions), in terms of the risk $R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\theta}}[L_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})]$, regardless of $\boldsymbol{\theta}$. Here and elsewhere we suppress in notation the dependence of the risk

function on $\boldsymbol{V}$.

In the heteroscedastic case there is no agreement 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 an i.i.d. normal prior on the elements of $\boldsymbol{\theta}$ 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, are minimax 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 $1/\sqrt{V_i}$ in order to make all variances equal; however, after applying this non-orthogonal transformation the loss needs to be changed accordingly (to a weighted loss) in order to maintain equivalence between the problems.

There have been attempts to moderate the respective disadvantages of estimators resulting from either of the two approaches mentioned above. For example, Xie et al. (2012) consider the family of Bayes estimators arising from the usual hierarchical model

$$\theta_i \overset{\text{iid}}{\sim} N(\mu, \gamma) \qquad X_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, V_i) \qquad 1 \leq i \leq n \tag{3}$$

and indexed by μ and γ. They suggest to plug into the Bayes rule,

$$\widehat{\theta}_i^{\mu,\gamma} = \mathbb{E}_{\mu,\gamma}(\theta_i|X_i) = X_i - \frac{V_i}{V_i + \gamma}(X_i - \mu), \tag{4}$$

values $(\widehat{\mu}, \widehat{\gamma}) = \arg\min_{\mu,\gamma} \mathcal{R}(\mu, \gamma; \boldsymbol{X})$ where $\mathcal{R}(\mu, \gamma; \boldsymbol{X})$ is an unbiased estimator of the

risk of $\widehat{\theta}^{\mu,\gamma}$. This reduces the sensitivity of the estimator to how appropriate model (3) is, as compared to the usual empirical Bayes estimators, that use Maximum Likelihood or Method-of-Moments estimates of μ, γ under (3). On the other hand, Berger (1982) suggested a modification of his own minimax estimator (Berger, 1976), 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 (3) can be constructed so they asymptotically dominate the usual estimator (Xie et al., 2012), the *modeling* of θ_i as identically distributed random variables is often not as well motivated in the heteroscedastic case as it is in the equal variances case. The assumption that θ_i are i.i.d. reflects, as commented by Efron and Morris (1973b), a "Bayesian statement of belief that the θ_i are of comparable magnitude". But this assumption is not always appropriate. There are many examples where an association between the V_i and the θ_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 (affecting the sampling variances of the observations). In situations where the true means and the V_i are associated, modeling the θ_i as i.i.d. is not adequate. Nevertheless, symmetry can be restored in the heteroscedastic case by treating the *pair* (X_i, V_i) as the random data. 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.

The rest of the paper 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

of 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 $\boldsymbol{X}, \boldsymbol{\theta}$ and $\boldsymbol{V}$ be as in (1). It is convenient to think of $\boldsymbol{\theta}$ and $\boldsymbol{V}$ as nonrandom, although the derivations below hold also when $\boldsymbol{\theta}$ or $\boldsymbol{V}$ (or both) are random. In the sequel we refer to a rule $\widehat{\boldsymbol{\theta}}$ as *separable* if $\widehat{\theta}_i(\boldsymbol{X}, \boldsymbol{V}) = t(X_i, V_i)$ for some function $t : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$. Denote by $\mathcal{D}_S$ the set of all separable rules. If $\widehat{\boldsymbol{\theta}} \in \mathcal{D}_S$ with $\widehat{\theta}_i(\boldsymbol{X}, \boldsymbol{V}) = t(X_i, V_i)$, then

$$R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta_i}[t(X_i, V_i) - \theta_i]^2 = \mathbb{E}[t(Y, A) - \xi]^2 \tag{5}$$

where the expectation in the last term is taken over the random vector $(Y, \xi, A, I)^T$ distributed according to

$$\mathbb{P}(I = i) = 1/n, \qquad (Y, \xi, A)|(I = i) \sim (X_i, \theta_i, V_i) \qquad 1 \le i \le n. \tag{6}$$

Above, the symbol " $\sim$ " stands for "equal in distribution". In words, (6) says that (ξ, A) have the empirical joint distribution of the pairs (θ_i, V_i); and $Y|(\xi, A) \sim N(\xi, A)$. Throughout the paper, when we refer to the random triple (Y, ξ, A), its relation to (X_i, θ_i, V_i), $1 \le i \le n$, is given by (6). The identity (5) – a computation à la Robbins – is easily verified by calculating the expectation on the right hand side by first conditioning on I. It says that for a separable estimator, the risk is equivalent to the Bayes risk in a one-dimensional estimation problem.

Now consider $\widehat{\boldsymbol{\theta}} \in \mathcal{D}_S$ with t linear (affine, in point of fact, but with a slight abuse of

terminology we use the former term for convenience) in its first argument, that is,

$$\widehat{\theta}_i^{a,b}(\boldsymbol{X},\boldsymbol{V}) = X_i - b(V_i)[X_i - a(V_i)] \qquad 1 \le i \le n \tag{7}$$

for some functions a, b. The corresponding Bayes risk in the last expression of (5) is

$$r_n(a,b) := \mathbb{E}\Big\{Y - b(A)[Y - a(A)] - \xi\Big\}^2. \tag{8}$$

Since

$$Y|(\xi, A) \sim N(\xi, A), \tag{9}$$

the minimizers of

$$r_n(a,b|v) := \mathbb{E}\Big\{\Big(Y - b(A)[Y - a(A)] - \theta\Big)^2\Big|A = v\Big\}, \tag{10}$$

and hence also of (8), are

$$a_n^*(v) = \mathbb{E}(Y|A = v), \quad b_n^*(v) = \frac{v}{\mathrm{Var}(Y|A = v)} \tag{11}$$

and the minimum Bayes risk is

$$R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{a_n^*,b_n^*}) = r_n(a_n^*, b_n^*) = \mathbb{E}\Big[A\big\{1 - b_n^*(A)\big\}\Big]. \tag{12}$$

Therefore, (12) is a lower bound on the risk achievable by any estimator of the form (7), and $\widehat{\boldsymbol{\theta}}^{a_n^*,b_n^*}$ is the optimal solution within the class. Note that any estimator of the form (4) is also of the form (7), hence the risk of the best (oracle) rule of the form (7) is no greater than the risk of the best rule of the form (4). If ξ and A are independent, $a_n^*(v) = \mathbb{E}(Y|A =$

$v) = \mathbb{E}(\xi|A = v) = \mathbb{E}(\xi)$, $\quad b_n^*(v) = v/(v + \mathrm{Var}(\xi))$, and the oracles of the forms (4) and (7) coincide.

Finally, we note that existing nonparametric empirical Bayes estimators, such as the semi-parametric estimator of Xie et al. (2012) and the nonparametric method of Jiang and Zhang (2010), target the best predictor $g(Y, A)$ of ξ where g is restricted to some nonparametric class of functions. While the optimal g may indeed be a non-linear function of Y, these methods implicitly assume independence between ξ and A, and might still suffer from the gap between the optimal predictor $g(Y, A)$ assuming independence, and the true Bayes rule, namely, $\mathbb{E}(\xi|Y, A)$. Therefore, in some cases the oracle rule of the form (7) might still have smaller risk than the oracle choice of g computed assuming independence between ξ and A.

3 Group-linear Shrinkage Methods

Let $\boldsymbol{X}, \boldsymbol{\theta}$ and $\boldsymbol{V}$ be as in (1). The estimator in the following lemma will serve as a building block for our group-linear estimator. Note in this estimator that $\bar{X}$ is used as an estimate of the overall group mean. In addition, the estimator is spherically symmetric as a function of $\boldsymbol{X} - \bar{X}$. Similar estimators that center on a known mean, and variations, have been discussed in Brown (1975, Theorem 3), Bock (1975), Berger (1985), Lehmann and Casella (1998, Theorem 5.7; although there are some typos), Tan (2015) and elsewhere.

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

$$\widehat{\theta}_i^c = X_i - \widehat{b}(X_i - \overline{X}), \quad \widehat{b} = \min\left(1, c_n \overline{V}/s_n^2\right) \tag{13}$$

where $\overline{X} = \sum_{i=1}^n X_i/n$, $\overline{V} = \sum_{i=1}^n V_i/n$, $s_n^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$ and c_n is a positive constant. Let $V_{\max} = \max_{i \le n} V_i$ and $c_n^* = \{[(n-3) - 2(V_{\max}/\overline{V} - 1)]/(n-1)\}_+ =$

$\{1 - 2(V_{\max}/\overline{V})/(n-1)\}_+$. *Then for* $0 \le c_n \le 2c_n^*$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(\widehat{\theta}_i - \theta_i\Big)^2 \;\le\; \overline{V}\left[1 - (1-1/n)\,\mathbb{E}\left\{(2c_n^* - c_n)\widehat{b} + (2 - 2c_n^* + c_n - s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\le c_n\}}\right\}\right] \;\le\; \overline{V}. \tag{14}$$

Remarks:

1. The main reason for using $\overline{X}$ is analytical simplicity. When θ_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 (14) note that when $s_n^2/\overline{V} \ge c_n$, $(2c_n^* - c_n)\widehat{b} = (2c_n^* - c_n)c_n\overline{V}/s_n^2$ attains maximum at $c_n = c_n^*$. In the homoscedastic case $V_{\max} = \overline{V}$ 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 $0 \le c_n \le 2\{1 - 2(V_{\max}/\overline{V})/n\}$. This is consistent with Lemma 1.

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

We are now ready to introduce an empirical Bayes estimator, which employs the spherically symmetric estimator of Lemma 1 to mimic the oracle rule $\widehat{\boldsymbol{\theta}}^{a^*,b^*}$. When the number of distinct values V_i is very small compared to n, a natural competitor of $\widehat{\boldsymbol{\theta}}^{a_n^*,b_n^*}$ 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 (12). The situation in the general heteroscedastic problem, when the number of

distinct values V_i is not very small compared to n, is not as obvious; still, the expression for the optimal function a^* and b^* in (11) 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 asymptotic adaptive wavelet estimation. However, the estimator of 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. See also Ma et al. (2015). For the heteroscedastic case, Tan (2014) comments 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 our approach.

Definition 1 (Group-linear Empirical Bayes Estimator for a Heteroscedastic Mean)**.** *Let $J_1, \ldots, J_m$ be disjoint intervals. For $k = 1, ..., m$ denote*

$$\mathcal{I}_k = \{i : V_i \in J_k\}, \quad n_k = |\mathcal{I}_k|, \quad \overline{V}_k = \sum_{i\in\mathcal{I}_k} \frac{V_i}{n_k}, \quad \overline{X}_k = \sum_{i\in\mathcal{I}_k} \frac{X_i}{n_k}, \quad s_k^2 = \sum_{i\in\mathcal{I}_k} \frac{(X_i - \overline{X}_k)^2}{n_k \vee 2 - 1}.$$

Define a corresponding group-linear estimator $\widehat{\boldsymbol{\theta}}^{GL}$ componentwise by

$$\widehat{\theta}_i^{GL} = \begin{cases} X_i - \min\left(1, c_k \overline{V}_k / s_k^2\right)(X_i - \overline{X}_k), & i \in \mathcal{I}_k \\ X_i, & \textit{otherwise} \end{cases} \tag{15}$$

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

Theorem 1. *For $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{GL}$ in Definition 1 with $c_k = \left\{1 - 2\left(\max_{i\in\mathcal{I}_k} V_i/\overline{V}_k\right)/(n_k - 1)\right\}_+$ the following holds:*

1. *Under the Gaussian model (1) with deterministic $(\theta_i, V_i), i \le n$, the risk of $\widehat{\boldsymbol{\theta}}$ is no*

greater than that of the naive estimator $\boldsymbol{X}$ *and therefore* $\widehat{\boldsymbol{\theta}}$ *is minimax*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(\widehat{\theta}_i-\theta_i\Big)^2\le\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(X_i-\theta_i\Big)^2=\frac{1}{n}\sum_{i=1}^{n}V_i=\overline{V}. \tag{16}$$

2. *Let* $(X_i,\theta_i,V_i), i=1,\ldots,n$, *be i.i.d. vectors from any fixed (with respect to* n*) population satisfying (1). Let* (Y,ξ,A) *be defined by (6);* $r(a,b)$ *as defined in (8); and* a^* *and* b^* *as defined in (11). Then*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big]\le\frac{1}{n}\sum_{i=1}^{n}r(a^*,b^*|V_i)+o(1) \tag{17}$$

with $\boldsymbol{V}=(V_1,...,V_n)$ *and for any sequence* $V_1,V_2,...$ *such that the following holds:*

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

$$\max_{1\le k\le m}|J_k|\to 0,\ \min_{1\le k\le m}n_k\to\infty,\qquad a^*(v),b^*(v)\ \textit{are uniformly continuous}$$
$$\limsup_{n\to\infty}\frac{\sum_{i=1}^{n}V_i}{n}<\infty,\ \limsup_{n\to\infty}\frac{\sum_{i=1}^{n}V_iI_{\{V_i\notin\cup_{k=1}^{m}J_k\}}}{n}=0 \tag{18}$$

Remarks on the second part of the theorem:

1. Note that when (X_i,θ_i,V_i) are i.i.d., then each triple is distributed as (Y,ξ,A). We assumed that the 'population' distribution (Y,ξ,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,ξ,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 analogue of the second part of the theorem could be stated for the nonrandom situation, $X_i|(\theta_i,V_i)\sim N(\theta_i,V_i), 1\le i\le n$ with deterministic θ_i and V_i. In this case, suppose that the empirical joint distribution

G_n of $\{(\theta_i, V_i) : 1 \le i \le n\}$ 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 $r(\widehat{a}_n, \widehat{b}_n) \to r(a^*, b^*)$ 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, $\max_{1\le k\le m} |J_k| \to 0$, $\min_{1\le k\le m} n_k \to \infty$ 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 $\max_{1\le k\le m} |J_k| \to 0$, $\min_{1\le k\le m} n_k \to \infty$ are satisfied when $\cup_{k=1}^m J_k$ are chosen to cover most of the observations, and at the same time $\limsup_{n\to\infty} \sum_{i=1}^n V_i I_{\{V_i \notin \cup_{k=1}^m J_k\}}/n = 0$, 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 $\boldsymbol{V}$ in (17), can be obtained in a similar way by replacing the conditions on $\boldsymbol{V}$ with bounded expectation of the random variable A. We skip this for simplicity.

For the i.i.d. 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 (17) regarding the rate of convergence of the excess risk. Noting that, trivially, $\xi = a^*(A)$ implies that $\mathbb{E}(\xi|A = v) = a^*(v)$, $X_i|V_i \sim N(a^*(V_i), V_i)$ is a nonparametric regression model, i.e., θ_i is a deterministic measurable function of V_i. In this case, the rate of convergence in (17) depends primarily on the smoothness of the function $a^*(v)$. In the homoscedastic case the smoothing feature of the James-Stein estimator was studied in 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 i.i.d. 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* $|J_k| = |J| = \left(\frac{10V_{\max}^2}{nL}\right)^{\frac{1}{3}}$ *and* $c_n = c_n^*$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \leq 2\Big(\frac{10V_{\max}^2\sqrt{L}}{n}\Big)^{\frac{2}{3}} \tag{19}$$

for any deterministic sequence $\boldsymbol{V} = (V_1, ..., V_n)$.

For the asymptotic results in Theorems 1 and 2 to hold, it is enough to choose bins J_k of equal length $|J| = \left(\frac{10V_{\max}^2}{nL}\right)^{\frac{1}{3}}$. 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 $\big(\max\{V_i : i \in J_k\}\big)/\overline{V}_k$ (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 $n^{-1/3}$ in $|J|$, which may be suitable when the $V_i \in (0, 1]$, we suggest in general to fix the *number* of bins to $n^{1/3}$, i.e., divide $\log(V_i)$ to bins of equal length $[\max_i(\log V_i) - \min_i(\log V_i)]/n^{1/3}$. 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 (17). Otherwise, the method is expected to work reasonably well in the sense of (17) 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 of Xie et al. (2012), and compare the performance of our group-linear estimator to the methods proposed in their work. In each example, we draw n i.i.d. triples $(X_i, \theta_i, V_i) \sim (Y, \xi, A)$ such that $Y|(\xi, A) \sim N(\xi, A)$; the last example is the only exception, with $Y|(\xi, A) \nsim N(\xi, A)$, in order to assess the sensitivity to departures from normality. Various estimators are then applied to the data $(X_i, V_i), 1 \leq i \leq n$, and the normalized sum of squared error is computed. For each value of n in $\{20, 40, 60, ..., 500\}$, this process is repeated $N = 10,000$ times to obtain a good estimate of the (Bayes) risk for each method. Among the empirical Bayes estimators proposed by Xie et al. (2012) we consider the parametric SURE estimator given by

$$\widehat{\theta}_i^M = X_i - \frac{V_i}{V_i + \widehat{\gamma}}(X_i - \widehat{\mu}), \qquad 1 \leq i \leq n$$

where $\widehat{\gamma}$ and $\widehat{\mu}$ minimize an unbiased estimator of the risk (SURE) for estimators of the form $\widehat{\theta}_i^{\mu,\gamma} = X_i - [V_i/(V_i + \gamma)](X_i - \mu)$ over μ and γ. We also consider the semi-parametric SURE estimator of Xie et al. (2012) with shrinkage towards the grand mean, defined by

$$\widehat{\theta}_i^{SG} = X_i - \widehat{b}_i(X_i - \overline{X}), \qquad 1 \leq i \leq n \tag{20}$$

Table 1: Oracle shrinkage locations and shrinkage factors, $\mu^*, v/(v+\gamma^*)$ and $a^*(v), b^*(v)$, corresponding to the family of estimators of Xie et al. (equation (23)) and to the family of estimators that are linear in Y (equation (24)). Columns correspond to simulation examples (a)- (f). Values of μ^*, γ^* for each example are from Xie et al. (2012).

	(a)	(b)	(c)	(d)	(e)	(f)
μ^*, $\frac{v}{v+\gamma^*}$	0, $\frac{v}{v+1}$	.5, $\frac{v}{v+.083}$	0.6, $\frac{v}{v+0.078}$	0.13, $\frac{v}{v+0.0032}$	0.15, $\frac{v}{v+0.84}$	0.6, $\frac{v}{v+0.078}$
$a^*(v)$, $b^*(v)$	0, $\frac{v}{v+1}$	0, $\frac{v}{v+1}$	$v, 0$	$v, 0$	$2\delta_{\{v=0.1\}}(v), 0.5$	$v, 0$

where $\widehat{\boldsymbol{b}} = (\widehat{b}_1, ..., \widehat{b}_n)$ minimize an unbiased estimator of the risk for estimators of the form $\widehat{\theta}_i^{\boldsymbol{b},\mu} = X_i - b_i(X_i - \overline{X})$ with $\boldsymbol{b} = (b_1, ..., b_n)$ restricted to satisfy $b_i \leq b_j$ whenever $V_i \leq V_j$. The group-linear estimator $\widehat{\boldsymbol{\theta}}^{GL}$ of Definition 1 is applied here with the bins J_k formed by dividing the range of $\log(V_i)$ into $\lfloor n^{1/3} \rfloor$ equal length intervals, per the discussion concluding Section 3. As benchmarks, in each example we also compute the two oracle risks

$$r(\mu^*, \gamma^*) = \min_{\mu,\gamma \in \mathbb{R} \,:\, \gamma \geq 0} \mathbb{E}\left\{ \left[Y - \frac{A}{A+\gamma}(Y-\mu) - \xi \right]^2 \right\} \tag{21}$$

and

$$r(a^*, b^*) = \min_{a(\cdot), b(\cdot) \,:\, a(v) \geq 0 \;\forall v} \mathbb{E}\left\{ \left[Y - b(A)\big(Y - a(A)\big) - \xi \right]^2 \right\} \tag{22}$$

corresponding to the optimal rule in the parametric family of estimators considered in Xie et al. (2012, labeled "XKB oracle" in the legend of Figure 1), and to the optimal linear-in-x rule of Section 2, respectively. Note that μ^* and γ^* are numbers whereas a^* and b^* are functions. Table 1 displays the oracle shrinkage locations and shrinkage factors corresponding to (21) and (22); 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, semi-parametric 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 semi-parametric SURE estimator, as $r(a^*, b^*) \leq \inf\{r(a, b) : b(v) \text{ monotone increasing in } v\}$. For moderate n, whenever ξ and A are independent, the SURE estimators are appropriate and achieve smaller risk. By contrast, the situations where ξ and A are dependent are handled best by the group-linear estimator, which indeed achieves significantly smaller risk than both SURE estimators.

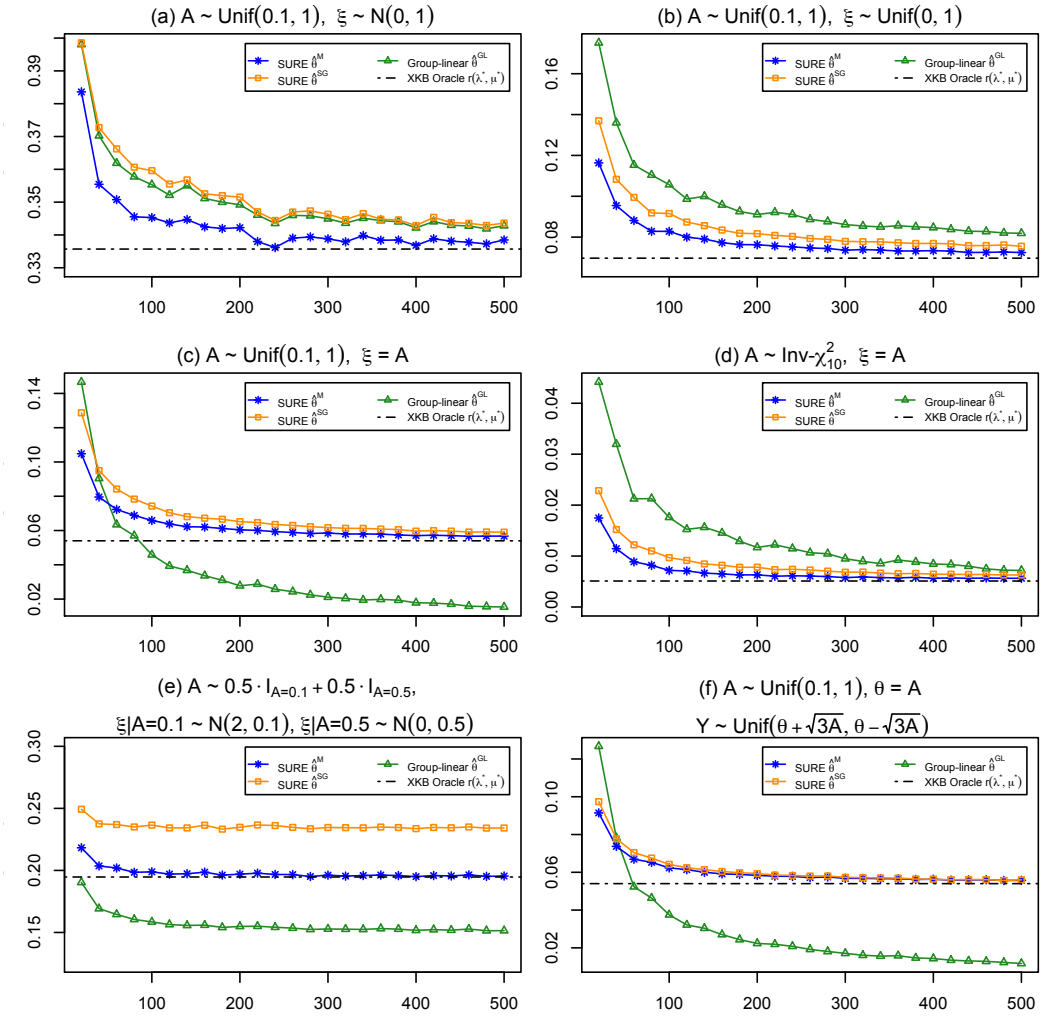

Figure 1: Estimated risk for various estimators vs. number of observations.

In example (a) (7.1 of Xie et al., 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 (4) and, in particular, the functions a^* and b^* are constant in v. The parametric SURE estimator is therefore appropriate, and it performs best, requiring estimation of only two hyperparameters. The group-linear estimator and the semi-parametric 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 ($\approx .3357$). In example (b) (7.2 of Xie et al., 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 v as long as ξ and A are independent. The risk of the group-linear approaches the oracle risk ($\approx .0697$), but here the semi-parametric 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 et al., 2012) takes $A \sim \text{Unif}(0.1, 1)$, $\xi = A$. Here ξ and A are strongly dependent, and indeed the gap between the two oracle risks, $r(\mu^*, \gamma^*) \approx .0540$ 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 semi-parametric SURE is slightly smaller than that of the parametric SURE, because of the improved capability of the semi-parametric oracle to accommodate the dependence between ξ and A. In the fourth case (d) (7.3 of Xie et al., 2012) we take $A \sim \text{Inv-}\chi^2_{10}$, $\xi = A$. ξ 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 et al., 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 semi-parametric SURE estimator of Xie et al. (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.

Lastly, in (f) (7.6 of Xie et al., 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 semi-parametric SURE estimator does just a little better, its risk approaching ≈ 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

$$H_{ji} \sim \text{Bin}(N_{ji}, p_i), \qquad j = 1, 2, \ i = 1, ..., \mathcal{P}_j. \tag{23}$$

As suggested in Brown (2008), a variance-stabilizing transformation is first applied, $X_{ji} = \arcsin\{(H_{ji} + 1/4)^{1/2}/(N_{ji} + 1/2)^{1/2}\}$, resulting in $X_{ji} \dot{\sim} N\big(\theta_i, 1/(4N_{ji})\big)$, $\theta_i = \arcsin(p_i)$, and $\{(X_{1i}, N_{1i}) : i = 1, ..., \mathcal{P}_1\}$ are then used to estimate the means θ_i. We should remark

that there is no reason for using this transformation, and for focusing on estimating θ_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 (23) directly (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 $\widehat{\boldsymbol{\theta}}$, we use the Total Squared Error, $\mathrm{TSE}(\widehat{\boldsymbol{\theta}}) = \sum_i \left[(X_{2i} - \widehat{\theta}_i)^2 - 1/(4N_{2i})\right]$, proposed by Brown (2008) as an (approximately) unbiased estimator of the risk of $\widehat{\boldsymbol{\theta}}$. 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_{1i} at-bats of the first period from a Hypergeometric distribution, $\mathcal{HG}(N_{1i} + N_{2i}, H_{1i} + H_{2i}, N_{1i})$. In the permutation analysis we concentrate on the two SURE methods of Xie et al. (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 et al. (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_{2i} 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.

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.

	All		Pitchers		Non-pitchers	
Naive	1		1		1	
Grand mean	.852		.127		.378	
Nonparametric EB	.508		.212		.372	
Binomial mixture	.588		.156		.314	
Weighted Least Squares	1.07		.127		.468	
Weighted nonparametric MLE	.306		.173		.326	
Weighted Least Squares (AB)	.537		.087		.290	
Weighted nonparametric MLE (AB)	.301		.141		**.261**	
James-Stein	.535	(.543)	.165	(.239)	.348	(.234)
SURE $\widehat{\theta}^{M}$	.421	(.484)	.123	(.211)	.289	(.265)
SURE $\widehat{\theta}^{SG}$	.408	(.468)	**.091**	(**.169**)	**.261**	(.219)
Group-linear $\widehat{\theta}^{GL}$	.302	(.280)	.178	(.244)	.325	(**.175**)
Group-linear (dynamic)	**.288**	(**.276**)	.168	(.283)	.349	(**.175**)

In the table, the Grand mean estimator uses the simple average of all X_{1i}; the extended positive-part James-Stein estimator is given by $\widehat{\theta}_i^{JS+} = \hat{\mu}_{JS+} + \big(1 - \frac{p-3}{\sum_i (X_i - \hat{\mu}_{JS+})}\big)_+ (X_i - \hat{\mu}_{JS+})$ where $\hat{\mu}_{JS+} = (\sum_i X_i/V_i)/(\sum_i 1/V_i)$; $\widehat{\boldsymbol{\theta}}^{M}$ is the parametric empirical Bayes estimator of Xie et al. (2012) using the SURE criterion to choose both the shrinkage and the location parameter; and $\widehat{\boldsymbol{\theta}}^{SG}$ is the semi-parametric SURE estimator of Xie et al. (2012) that shrinks towards 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).

For the group-linear estimator, in addition to the plain estimator $\widehat{\boldsymbol{\theta}}^{GL}$ that uses $k = \lfloor n^{1/3} \rfloor$ equal length bins on $\log(\frac{1}{4N_{1i}})$ (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/MaZhuang/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 in 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 θ_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, θ_i, and the sampling variances of X_{ji} ($\propto 1/N_{ji}$); see Figure 4 in Brown (2008). While the nonparametric MLE method handles well non-normality in the "prior" distribution of the θ_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 semi-parametric SURE estimator (0.325 vs. 0.261); but the opposite emerges for shuffled data (0.175 vs. 0.219). For pitchers only, the estimators of Xie et al. (2012) outperform 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.

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 θ_i has a prior distribution unrelated to the observed sampling variance $V_i = \mathrm{Var}(X_i|\theta_i)$. 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 θ_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 i.i.d. case, the distribution of the population (Y, ξ, 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 (17) should be strengthened to the asymptotic ratio optimality criterion

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(\widehat{\theta}_i - \theta_i\Big)^2 \le (1+o(1))r_n(a_n^*, b_n^*) \tag{24}$$

   as $n \to \infty$. As (24) does not hold uniformly for all joint distributions of (Y, ξ, A), a reasonable target would be to prove (24) when $r_n(a^*, b^*) \ge \eta_n$ for small η_n under suitable side conditions on the joint distribution of (Y, ξ, A). This theory should include (17) as a special case and still maintain the property (16).

2. When $a^*(v)$ satisfies an order α smoothness condition with $\alpha > 1$, a higher-order estimate of $a^*(V_i)$ needs to be used to achieve the optimal rate $n^{-\alpha/(2\alpha+1)}$ in the nonparametric regression case, $r(a^*, b^*) = 0$, e.g., $\widehat{a}(V_i)$ with an estimated polynomial $\widehat{a}(v)$ for each J_k. We speculate that such a group-polynomial estimator might still

always outperform the naive estimator $\widehat{\theta}_i = X_i$ under a somewhat stronger minimum sample size requirement.

Appendix: Proofs

Proof of Lemma 1 It suffices to consider $0 < c_n \le 2c_n^*$. Let $b(x) = \min(1, c_n\overline{V}/x)$ such that $\widehat{b} = b\big(s_n^2\big)$. Notice that $(\partial/\partial X_i)s_n^2 = 2(X_i - \overline{X})/(n-1)$. By Stein's lemma,

$$\mathbb{E}(X_i - \theta_i)(X_i - \overline{X})\widehat{b} = V_i\,\mathbb{E}\Big\{(1-1/n)b(s_n^2) + 2(X_i - \overline{X})^2 b'\big(s_n^2\big)/(n-1)\Big\}.$$

By definition, $2V_i/(n-1) \le \overline{V}(1-c_n^*)$ and $xb'(x) = -b(x)I\{b(x) < 1\}$,

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big(X_i - (X_i - \overline{X})\widehat{b} - \theta_i\Big)^2 \\
= \;& \frac{1}{n}\sum_{i=1}^n \left[V_i + \mathbb{E}(X_i - \overline{X})^2 b^2(s_n^2) - 2V_i\mathbb{E}\left\{(1-1/n)b(s_n^2) + \frac{2(X_i - \overline{X})^2 b'\big(s_n^2\big)}{n-1}\right\}\right] \\
\le \;& \overline{V} + (1-1/n)\,\mathbb{E}\left\{s_n^2 b^2(s_n^2) - 2\overline{V}b(s_n^2) + \overline{V}(1-c_n^*)2b\big(s_n^2\big)I_{\{s_n^2 > c_n\overline{V}\}}\right\} \\
= \;& \overline{V} + (1-1/n)\,\mathbb{E}\,\overline{V}b(s_n^2)\left\{\min\left(s_n^2/\overline{V}, c_n\right) - 2 + 2(1-c_n^*)I_{\{s_n^2 > c_n\overline{V}\}}\right\} \\
= \;& \overline{V} - (1-1/n)\,\mathbb{E}\,\overline{V}b(s_n^2)\left\{(2c_n^* - c_n)I_{\{s_n^2 > c_n\overline{V}\}} + (2 - s_n^2/\overline{V})I_{\{s_n^2 \le c_n\overline{V}\}}\right\} \\
= \;& \overline{V}\left[1 - (1-1/n)\,\mathbb{E}\left\{b(s_n^2)(2c_n^* - c_n) + (2 - 2c_n^* + c_n - s_n^2/\overline{V})I_{\{s_n^2/\overline{V} \le c_n\}}\right\}\right] \le \overline{V}.
\end{aligned}$$

Define $\epsilon_{|J|} = \max\limits_{v_1, v_2 \in J}\big\{|a^*(v_1) - a^*(v_2)|, |b^*(v_1) - b^*(v_2)|\big\}$, $g(v) = \mathrm{Var}(\xi|A = v)$ and $h(v) = \mathbb{E}(\xi^2|A = v)$. Unless otherwise stated, all expectations and variances are conditional on $\boldsymbol{V}$.

Lemma 2 (Analysis within each block)**.** *Let $(X_i, \theta_i, V_i)_{i=1}^n$ be i.i.d. vectors drawn from some population (Y, ξ, A) satisfying (9) with $n \ge 2$. If $V_1, \cdots, V_n \in J$ for some interval J and $\min_{1\le i\le n} b^*(V_i) \ge \varepsilon, b^*(\overline{V}) \ge \varepsilon$ for some $\varepsilon > 0$. Then the spherically symmetric shrinkage*

estimator defined in (15) *with* $c_n = c_n^*$ *satisfies*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{i=1}^{n} r(a^*,b^*|V_i) + \frac{7V_{\max}}{n-1} + (\overline{V}\epsilon_{|J|}+|J|)\frac{\varepsilon^2+1}{\varepsilon^2} + \epsilon_{|J|}^2 \\ + \frac{2}{n-1}\Big\{\sum_{i=1}^{n}V_i^2 + 2\sum_{i=1}^{n}(V_i+\overline{V})h(V_i) + \overline{V}^2\Big\}^{\frac{1}{2}}\end{aligned} \tag{25}$$

where $V_{\max} = \max\{V_1,\cdots,V_n\}$ *and* $\overline{V} = \sum_{i=1}^n V_i/n$.

Proof of Lemma 2 As in the proof of Lemma 1 with $c_n = c_n^*$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(X_i-(X_i-\overline{X})\widehat{b}-\theta_i\Big|\boldsymbol{V}\Big)^2 \\ &\le \overline{V} + \left(1-\frac{1}{n}\right)\mathbb{E}\,\overline{V}b(s_n^2)\Big\{\min\left(s_n^2/\overline{V},c_n^*\right) - 2 + 2(1-c_n^*)I_{\{s_n^2>c_n^*\overline{V}\}}\Big\}\end{aligned}$$

By definition, $r(a^*,b^*|V_i) = V_i(1-b^*(V_i))$ and $\min\left(s_n^2/\overline{V},c_n^*\right) \le c_n^* \le 1$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{i=1}^{n} r(a^*,b^*|V_i) + \frac{1}{n}\sum_{i=1}^{n} b^*(V_i)V_i - \left(1-\frac{1}{n}\right)\overline{V}\mathbb{E}(\widehat{b}) + 2\overline{V}(1-c_n^*)$$

Observing that $0 \le \widehat{b} \le 1$ and $\overline{V}(1-c_n^*) \le 2V_{\max}/(n-1)$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] &\le \frac{1}{n}\sum_{i=1}^{n} r(a^*,b^*|V_i) + 4V_{\max}/(n-1) + \overline{V}/n + \frac{1}{n}\sum_{i=1}^{n} b^*(V_i)V_i - \overline{V}\mathbb{E}(\widehat{b}) \\ &\le \frac{1}{n}\sum_{i=1}^{n} r(a^*,b^*|V_i) + 5V_{\max}/(n-1) + \overline{V}\Big(\max_{1\le i\le n} b^*(V_i) - \mathbb{E}\widehat{b}\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} r(a^*,b^*|V_i) + 5V_{\max}/(n-1) + \overline{V}\big\{\max_{1\le i\le n} b^*(V_i) - b^*(\overline{V})\big\} + \overline{V}\Big(b^*(\overline{V}) - \mathbb{E}\widehat{b}\Big) \\ &\le \frac{1}{n}\sum_{i=1}^{n} r(a^*,b^*|V_i) + 5V_{\max}/(n-1) + \overline{V}\epsilon_{|J|} + \overline{V}\Big(b^*(\overline{V}) - \mathbb{E}\widehat{b}\Big)\end{aligned}$$

where the last inequality is due to the uniformly continuity of $b^*(v)$. Next we will bound $\overline{V}\Big(b^*(\overline{V}) - \mathbb{E}\widehat{b}\Big)$. By definition, $\overline{V}(b^*(\overline{V}) - \mathbb{E}\widehat{b}) = \overline{V}\mathbb{E}\Big\{\overline{V}/\mathrm{Var}(Y|A=\overline{V}) - \min(1, c_n^*\overline{V}/s_n^2)\Big\}$.

Further observe that $\overline{V}/\mathrm{Var}(Y|A=\overline{V}) = \overline{V}/\left(\overline{V}+\mathrm{Var}(\xi|A=\overline{V})\right) \le 1$,

$$\begin{aligned}
\overline{V}(b^*(\overline{V}) - E\widehat{b}) &\le \overline{V}\mathbb{E}\Big\{\Big(\overline{V}/\mathrm{Var}(Y|A=\overline{V}) - c_n^*\overline{V}/s_n^2\Big)I_{\{c_n^*\overline{V}\le s_n^2\}}\Big\} \\
&\le \overline{V}\mathbb{E}\Big\{\left(1 - c_n^*\mathrm{Var}(Y|A=\overline{V})/s_n^2\right)I_{\{c_n^*\overline{V}\le s_n^2\}}\Big\} \\
&= \mathbb{E}\overline{V}\Big\{(1-c_n^*)I_{\{c_n^*\overline{V}\le s_n^2\}} + \frac{c_n^*}{s_n^2}\big[s_n^2 - \mathrm{Var}(Y|A=\overline{V})\big]I_{\{c_n^*\overline{V}\le s_n^2\}}\Big\}
\end{aligned}$$

Also, noting that $1-c_n^* \ge 0$ and $c_n^*\overline{V}/s_n^2 I_{\{c_n^*\overline{V}\le s_n^2\}} \le 1$,

$$\begin{aligned}
\overline{V}\Big(b^*(\overline{V}) - E\widehat{b}\Big) &\le \overline{V}(1-c_n^*) + \mathbb{E}|s_n^2 - \mathrm{Var}(Y|A=\overline{V})| \\
&\le 2V_{\max}/(n-1) + \mathbb{E}|s_n^2 - \mathbb{E}s_n^2| + |\mathbb{E}s_n^2 - \mathrm{Var}(Y|A=\overline{V})| \\
&= 2V_{\max}/(n-1) + \mathbb{E}\Big\{\mathbb{E}_{\boldsymbol{\theta}}|s_n^2 - \mathbb{E}s_n^2|\Big\} + |\mathbb{E}s_n^2 - \mathrm{Var}(Y|A=\overline{V})| \\
&\le 2V_{\max}/(n-1) + \mathbb{E}\sqrt{\mathrm{Var}(s_n^2|\boldsymbol{\theta})} + |\mathbb{E}s_n^2 - \mathrm{Var}(Y|A=\overline{V})| \\
&\le 2V_{\max}/(n-1) + \Big\{\mathbb{E}\Big[\mathrm{Var}(s_n^2|\boldsymbol{\theta})\Big]\Big\}^{\frac{1}{2}} + |\mathbb{E}s_n^2 - \mathrm{Var}(Y|A=\overline{V})|
\end{aligned}$$

where the last two inequalities are due to Jensen's inequality. Conditionally on $\boldsymbol{V} = (V_1, \cdots, V_n)$ and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$, $\overline{X} \sim N(\sum_{i=1}^n \theta_i/n, \sum_{i=1}^n V_i/n^2)$, and therefore

$$\begin{aligned}
\mathbb{E}(s_n^2) &= \frac{1}{n-1}\mathbb{E}\Big\{\mathbb{E}\Big(\sum_{i=1}^n X_i^2 - n\overline{X}^2|\boldsymbol{\theta}\Big)\Big\} = \frac{1}{n-1}\mathbb{E}\Big\{\sum_{i=1}^n(V_i+\theta_i^2) - \frac{(\sum_{i=1}^n\theta_i)^2}{n} - \overline{V}\Big\} \\
&= \overline{V} + \frac{1}{n(n-1)}\Big\{(n-1)\sum_{i=1}^n \mathbb{E}(\xi^2|A=V_i) - \sum_{j\ne k}\mathbb{E}(\xi|A=V_j)\mathbb{E}(\xi|A=V_k)\Big\} \qquad (26) \\
&= \overline{V} + \frac{1}{n(n-1)}\Big\{(n-1)\sum_{i=1}^n \mathrm{Var}(\xi|A=V_i) + n\sum_{i=1}^n\Big[\mathbb{E}(\xi|A=V_i) - \frac{1}{n}\sum_{j=1}^n\mathbb{E}(\xi|A=V_j)\Big]^2\Big\} \\
&\le \overline{V} + \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\xi|A=V_i) + \frac{1}{n-1}\sum_{i=1}^n\Big[\mathbb{E}(\xi|A=V_i) - \frac{1}{n}\sum_{j=1}^n\mathbb{E}(\xi|A=V_j)\Big]^2 \\
&= \overline{V} + \frac{1}{n}\sum_{i=1}^n g(V_i) + \frac{1}{n-1}\sum_{i=1}^n\Big[a^*(V_i) - \frac{1}{n}\sum_{j=1}^n a^*(V_j)\Big]^2
\end{aligned}$$

On the other hand, $\mathrm{Var}(Y|A=\overline{V}) = \overline{V} + \mathrm{Var}(\xi|A=\overline{V}) = \overline{V} + g(\overline{V})$. Hence,

$$|\mathbb{E}(s_n^2) - \mathrm{Var}(Y|A=\overline{V})| \le \frac{1}{n}\sum_{i=1}^n |g(V_i) - g(\overline{V})| + \frac{1}{n-1}\sum_{i=1}^n \big[a^*(V_i) - \frac{1}{n}\sum_{j=1}^n a^*(V_j)\big]^2$$

The uniform continuity of $a^*(v)$ implies that $|a^*(V_i) - \sum_{j=1}^n a^*(V_j)/n| \le (n-1)/n\epsilon_{|J|}$. By definition, $b^*(v) = v/(v+g(v))$, then $g(v) = v/b^*(v) - v$ and therefore

$$\begin{aligned} |g(V_i) - g(\overline{V})| &= \Big|\frac{V_i b^*(\overline{V}) - \overline{V} b^*(V_i)}{b^*(V_i)b^*(\overline{V})} + (V_i - \overline{V})\Big| \\ &\le \frac{\big|V_i\big[b^*(\overline{V}) - b^*(V_i)\big]\big|}{b^*(V_i)b^*(\overline{V})} + \frac{\big|(V_i-\overline{V})b^*(V_i)\big|}{b^*(V_i)b^*(\overline{V})} + \big|V_i - \overline{V}\big| \quad \le \frac{\big(V_i\epsilon_{|J|} + |J|\big)}{\varepsilon^2} + |J| \end{aligned}$$

where the last inequality follows from $\min_{1\le i\le n} b^*(V_i) \ge \varepsilon, b^*(\overline{V}) \ge \varepsilon$. Combining the two inequalities above, $|\mathbb{E}(s_n^2) - \mathrm{Var}(Y|A=\overline{V})| \le \big(\overline{V}\epsilon_{|J|} + |J|\big)/\varepsilon^2 + |J| + \epsilon_{|J|}^2$. Finally, we are going to control $\mathbb{E}\Big\{\mathrm{Var}(s_n^2|\boldsymbol{\theta})\Big\}$. Again, $\overline{X}|\boldsymbol{V},\boldsymbol{\theta} \sim N(\sum_{i=1}^n \theta_i/n, \sum_{i=1}^n V_i/n^2)$, hence

$$\begin{aligned} \mathbb{E}\Big\{\mathrm{Var}(s_n^2|\boldsymbol{\theta})\Big\} &= \frac{1}{(n-1)^2}\mathbb{E}\Big\{\mathrm{Var}\Big(\sum_{i=1}^n X_i^2 - n\overline{X}^2|\boldsymbol{\theta}\Big)\Big\} \\ &\le \frac{2}{(n-1)^2}\mathbb{E}\Big\{\mathrm{Var}\Big(\sum_{i=1}^n X_i^2|\boldsymbol{\theta}\Big) + \mathrm{Var}\Big(n\overline{X}^2|\boldsymbol{\theta}\Big)\Big\} \\ &= \frac{2}{(n-1)^2}\mathbb{E}\Big\{\sum_{i=1}^n \big(2V_i^2 + 4\theta_i^2 V_i\big) + n^2\Big(2\overline{V}^2/n^2 + 4\overline{\theta}^2\overline{V}/n\Big)\Big\} \end{aligned}$$

By definition, $h(v) = \mathbb{E}(\xi^2|A=v)$, and, noting that $n\overline{\theta}^2 \le \sum_{i=1}^n \theta_i^2$,

$$\begin{aligned} \mathbb{E}\Big\{\mathrm{Var}(s_n^2|\boldsymbol{\theta})\Big\} &\le \frac{4}{(n-1)^2}\Big\{\sum_{i=1}^n V_i^2 + 2\sum_{i=1}^n V_i h(V_i) + \overline{V}^2 + 2\overline{V}\sum_{i=1}^n h(V_i)\Big\} \\ &\le \frac{4}{(n-1)^2}\Big\{\sum_{i=1}^n V_i^2 + 2\sum_{i=1}^n (V_i+\overline{V})h(V_i) + \overline{V}^2\Big\} \end{aligned} \tag{27}$$

Put pieces together, we have

$$\overline{V}(b^*(\overline{V}) - \mathbb{E}\widehat{b}) \le \frac{2V_{\max}}{n-1} + \frac{\overline{V}\epsilon_{|J|} + |J|}{\varepsilon^2} + |J| + \epsilon_{|J|}^2 + \frac{2}{n-1}\Big\{\sum_{i=1}^n V_i^2 + 2\sum_{i=1}^n (V_i+\overline{V})h(V_i) + \overline{V}^2\Big\}^{\frac{1}{2}}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big]\le\frac{1}{n}\sum_{i=1}^{n}r(a^*,b^*|V_i)+\frac{7V_{\max}}{n-1}+(\overline{V}\epsilon_{|J|}+|J|)\frac{\varepsilon^2+1}{\varepsilon^2}+\epsilon_{|J|}^2$$
$$+\frac{2}{n-1}\Big\{\sum_{i=1}^{n}V_i^2+2\sum_{i=1}^{n}(V_i+\overline{V})h(V_i)+\overline{V}^2\Big\}^{\frac{1}{2}}$$

Proof of Theorem 1. The first part the Theorem follows from Lemma 1. For the second part, it suffices to show that for all $\varepsilon>0$, the excess risk is $O_n(\varepsilon)$. Notice that the contribution to the normalized risk for observations outside $\cup_{k=1}^m J_k$ is $\sum_{i=1}^n V_i I_{\{V_i\notin\cup_{k=1}^m J_k\}}/n=o(1)$, we only need to consider the case where $\forall 1\le i\le n$, $V_i\in\cup_{k=1}^m J_k$. Without loss of generality, we assume $\forall\,1\le k\le m$, either $J_k\subset[0,\varepsilon)$ or $J_k\subset(\varepsilon,+\infty)$ since we can always reduce ε such that this happens. Due to the assumption that $\limsup_{n\to\infty}\sum_{i=1}^n V_i/n<\infty$, we can also choose M_ε large enough such that $\sum_{i=1}^n V_i I_{\{V_i\ge M_\varepsilon\}}/n\le\varepsilon$ and for any k with $J_k\subset(\varepsilon,+\infty)$, either $J_k\subset(\varepsilon,M_\varepsilon)$ or $J_k\subset(M_\varepsilon,+\infty)$.

For the rest of the proof, we divide all the observations into four disjoint groups and handle them separately. Let $\overline{V}^k=\sum_{i\in\mathcal{I}_k}V_i/n_k$ and define $S_1=\{k|1\le k\le n,J_k\subset(0,\varepsilon)\}$, $S_2=\{k|1\le k\le n,J_k\subset(\varepsilon,M_\varepsilon),\min_{V_i\in J_k}b^*(V_i)\ge\varepsilon,b^*(\overline{V}^k)\ge\varepsilon\}$, $S_3=\{k|1\le k\le n,J_k\subset(\varepsilon,M_\varepsilon),\min_{V_i\in J_k}b^*(V_i)<\varepsilon\ or\ b^*(\overline{V}^k)\le\varepsilon\}$, $S_4=\{k|1\le k\le n,J_k\subset(M_\varepsilon,+\infty)\}$.
Case i) For the small variance part, $V_i\in(0,\varepsilon)$, the contribution to the risk is negligible. Because the group linear shrinkage estimator dominate the MLE in each interval, then

$$\frac{1}{n}\sum_{k\in S_1}\sum_{i\in\mathcal{I}_k}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big]\le\sum_{k\in S_1}\sum_{i\in\mathcal{I}_k}V_i/n\le\sum_{k\in S_1}\sum_{i\in\mathcal{I}_k}\varepsilon/n\le\varepsilon$$

Case ii) For moderate variance with large shrinkage factor, $V_i\in(\varepsilon,M_\varepsilon)$ and $b^*(V_i),b^*(\overline{V})\ge\varepsilon$, shrinkage is necessary to mimic the oracle. Applying Lemma 2 to each interval $J_k,k\in S_2$,

$$\frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big]\le\frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k}r(a^*,b^*|V_i)+\frac{1}{n}\sum_{k\in S_2}n_k\Big\{\frac{7}{n_k-1}(\overline{V}^k+|J_k|)$$
$$+\Big(\overline{V}^k\epsilon_{|J_k|}+|J_k|\Big)\frac{\varepsilon^2+1}{\varepsilon^2}+\epsilon_{|J_k|}^2+\frac{2}{n_k-1}\Big(\sum_{i\in\mathcal{I}_k}V_i^2+2\sum_{i\in\mathcal{I}_k}(V_i+\overline{V}^k)h(V_i)+(\overline{V}^k)^2\Big)^{\frac{1}{2}}\Big\}$$

Let $|J|_{\max} = \max_{1\le k\le m} |J_k|, \epsilon_{\max} = \max_{1\le k\le m} \epsilon_{|J_k|}$. Using the fact that $\max_{1\le k\le m} n_k/(n_k-1) \le 2$,

$$\frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k} r(a^*,b^*|V_i) + \frac{1}{n}\sum_{k\in S_2}\Big\{14(\overline{V}^k + |J|_{\max}) + n_k\epsilon_{\max}^2 \\ + n_k(\overline{V}^k\epsilon_{\max} + |J|_{\max})\frac{\varepsilon^2+1}{\varepsilon^2} + 4\Big(\sum_{i\in\mathcal{I}_k} V_i^2 + 2\sum_{i\in\mathcal{I}_k}(V_i+\overline{V}^k)h(V_i) + (\overline{V}^k)^2\Big)^{\frac{1}{2}}\Big\}$$

For any $k\in S_2, i\in\mathcal{I}_k$, $\overline{V}^k, V_i \le M_\varepsilon$. Because $a^*(v)$ is uniformly continuous on $[0,M_\varepsilon]$, there exists constant C_ε only depending on ε such that $a^*(V_i)\le C_\varepsilon$. Then,

$$h(V_i) = \mathrm{Var}(\xi|A=V_i) + \Big(\mathbb{E}(\xi|A=V_i)\Big)^2 \le V_i/b^*(V_i) - V_i + (a^*(V_i))^2 \le M_\varepsilon/\varepsilon + C_\varepsilon^2$$

$$\frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k} r(a^*,b^*|V_i) + \frac{14|S_2|}{n}\Big(M_\varepsilon + |J|_{\max}\Big) + \epsilon_{\max}^2 \\ + (M_\varepsilon\epsilon_{\max} + |J|_{\max})\frac{\varepsilon^2+1}{\varepsilon^2} + \frac{4}{n}\sqrt{2M_\varepsilon^2(1+\varepsilon^{-1}) + 2M_\varepsilon C_\varepsilon}\sum_{k\in S_2}\sqrt{n_k}$$

By the Cauchy Schwarz inequality: $\sum_{k\in S_2}\sqrt{n_k} \le \sqrt{|S_2|\sum_{k\in S_2} n_k} \le \sqrt{|S_2|n}$. Further observe that $|S_2| \le m \le n/\min_{1\le k\le m} n_k$, then

$$\frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k} r(a^*,b^*|V_i) + \frac{14}{\min_{1\le k\le m} n_k}\Big(M_\varepsilon + |J|_{\max}\Big) + \epsilon_{\max}^2 \\ + (M_\varepsilon\epsilon_{\max} + |J|_{\max})\frac{\varepsilon^2+1}{\varepsilon^2} + \frac{4}{\sqrt{\min_{1\le k\le m} n_k}}\sqrt{2M_\varepsilon^2(1+\varepsilon^{-1}) + 2M_\varepsilon C_\varepsilon}$$

Since $|J|_{\max}, \epsilon_{\max} \to 0$ and $\min_{1\le k\le m} n_k \to +\infty$, we obtain

$$\frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{k\in S_2}\sum_{i\in\mathcal{I}_k} r(a^*,b^*|V_i) + o(\varepsilon)$$

Case iii) For moderate variance with negligible shrinkage factor, $V_i \in (\varepsilon, M_\varepsilon)$ and $\min_{i\in\mathcal{I}_k} b^*(V_i)$ *or* $b^*(\overline{V}) < \varepsilon$. The uniform continuity of $b^*(\cdot)$ implies that $\forall i\in\mathcal{I}_k,\ b^*(V_i) \le$

$\varepsilon + \epsilon_{\max}$. By definition $r(a^*, b^*|V_i) = V_i(1 - b^*(V_i))$, then

$$\frac{1}{n}\sum_{k\in S_3}\sum_{i\in\mathcal{I}_k} r(a^*, b^*|V_i) = \frac{1}{n}\sum_{k\in S_3}\sum_{i\in\mathcal{I}_k} V_i(1 - b^*(V_i)) \geq \frac{1}{n}\sum_{k\in S_3}\sum_{i\in\mathcal{I}_k} V_i - \overline{V}(\varepsilon + \epsilon_{\max})$$

Since the proposed group linear shrinkage estimator dominates MLE in each block,

$$\frac{1}{n}\sum_{k\in S_3}\sum_{i\in\mathcal{I}_k} \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \leq \frac{1}{n}\sum_{k\in S_3}\sum_{i\in\mathcal{I}_k} r(a^*, b^*|V_i) + \overline{V}(\varepsilon + \epsilon_{\max})$$

Case iv) For the large variance part, $V_i \in (M_\varepsilon, +\infty)$, by the definition of M_ε,

$$\frac{1}{n}\sum_{k\in S_4}\sum_{i\in\mathcal{I}_k} \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \leq \sum_{k\in S_4}\sum_{i\in\mathcal{I}_k} V_i/n = \sum_{i=1}^{n} V_i I_{\{V_i\geq M_\varepsilon\}}/n \leq \varepsilon$$

Summing up the inequalities of all four cases

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

which completes the proof by the assumption that $\limsup_{n\to\infty} \sum_{i=1}^{n} V_i/n \leq \infty$ □

Lemma 3 (Analysis within each block)**.** *Let $(X_i, \theta_i, V_i)_{i=1}^n$ be i.i.d. vectors from some population (Y, ξ, A) satisfying (9). If $r(a^*, b^*) = 0, a^*(\cdot)$ is L-Lipschitz continuous and $V_1, \cdots, V_n \in J$ for some interval J, then the estimator defined in (15) with $c_n = c_n^*$ satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \leq L|J|^2 + 3\overline{V}/n + 4V_{\max}/(n \vee 2 - 1)$$

Proof of Lemma 3 As in the proof of Lemma 1 and substitute c_n with c_n^*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(X_i - (X_i - \overline{X})\widehat{b} - \theta_i|\boldsymbol{V}\right)^2 \\
&\leq \overline{V}\left[1 - (1 - 1/n)\,\mathbb{E}\left\{\widehat{b}(2c_n^* - c_n) + (2 - 2c_n^* + c_n - s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\leq c_n\}}\right\}\right] \\
&= \overline{V}\left[1 - (1 - 1/n)\,\mathbb{E}\left\{\widehat{b}c_n^* + (2 - c_n^* - s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\leq c_n^*\}}\right\}\right]
\end{aligned}$$

$$= \overline{V}\mathbb{E}\left\{(1-\widehat{b}c_n^*) - (2-2c_n^*)I_{\{s_n^2/\overline{V}\le c_n^*\}} - (c_n^* - s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\le c_n^*\}}\right\}$$
$$+ \mathbb{E}\left\{\widehat{b}c_n^* + (2-c_n^*-s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\le c_n^*\}}\right\}\overline{V}/n$$

Notice that $2-2c_n^* > 0$ and $\widehat{b}c_n^* + (2-c_n^*-s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\le c_n^*\}} \le 2$.

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \overline{V}\mathbb{E}\left\{(1-\widehat{b}c_n^*) - (c_n^* - s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\le c_n^*\}}\right\} + 2\overline{V}/n$$
$$\le \overline{V}\mathbb{E}\Big\{c_n^*(1-\widehat{b}) - (c_n^* - s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\le c_n^*\}}\Big\} + 2\overline{V}/n + (1-c_n^*)\overline{V}$$
$$\le \mathbb{E}\Big\{c_n^*\overline{V}\left(\frac{s_n^2 - c_n^*\overline{V}}{s_n^2}\right)_+ - \big(c_n^*\overline{V} - s_n^2\big)_+\Big\} + 2\overline{V}/n + (1-c_n^*)\overline{V}$$
$$\le \mathbb{E}\Big\{\big(s_n^2 - c_n^*\overline{V}\big)_+ - \big(c_n^*\overline{V} - s_n^2\big)_+\Big\} + 2\overline{V}/n + (1-c_n^*)\overline{V}$$
$$= \mathbb{E}(s_n^2 - c_n^*\overline{V}) + 2\overline{V}/n + (1-c_n^*)\overline{V}$$

Recall that $\mathbb{E}s_n^2 = \overline{V} + \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\xi|A=V_i) + \frac{1}{n\vee 2-1}\sum_{i=1}^n\big[\mathbb{E}(\xi|A=V_i) - \frac{1}{n}\sum_{j=1}^n \mathbb{E}(\xi|A=V_j)\big]^2$. With $\mathrm{Var}(\xi|A=v)=0$, we have $\mathbb{E}s_n^2 = \overline{V} + \frac{1}{n\vee 2-1}\sum_{i=1}^n\big[a(V_i) - \frac{1}{n}\sum_{j=1}^n a(V_j)\big]^2$ and

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le 2(1-c_n^*)\overline{V} + \frac{1}{n\vee 2-1}\sum_{i=1}^n\big[a(V_i) - \frac{1}{n}\sum_{j=1}^n a(V_j)\big]^2 + 2\overline{V}/n$$
$$\le L|J|^2 + 2\overline{V}/n + 2(1-c_n^*)\overline{V} \le L|J|^2 + 2\overline{V}/n + \frac{4V_{\max}}{n\vee 2-1}$$

Proof of Theorem 2. Apply Lemma 3 to each interval and notice $n_k/(n_k-1)\le 2$,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{k=1}^m \big(n_k L|J_k|^2 + 2\overline{V}^k + 4V_{\max}\frac{n_k}{n_k\vee 2-1}\big)$$
$$\le L|J|^2 + 10mV_{\max}/n = L|J|^2 + 10V_{\max}^2/(n|J|)$$

Letting $|J| = \big(\frac{10V_{\max}^2}{nL}\big)^{\frac{1}{3}}$, we have that $\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le 2\big(\frac{10V_{\max}^2\sqrt{L}}{n}\big)^{\frac{2}{3}}$ □

References

James O Berger. Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *The Annals of Statistics*, pages 223–226, 1976.

James O Berger. Selecting a minimax estimator of a multivariate normal mean. *The Annals of Statistics*, 10(1):81–92, 1982.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

Mary Ellen Bock. Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 209–218, 1975.

Lawrence D Brown. Estimation with incompletely specified loss functions (the case of several location parameters). *Journal of the American Statistical Association*, 70(350):417–427, 1975.

Lawrence D Brown. In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics*, pages 113–152, 2008.

Lawrence D Brown and Eitan Greenshtein. Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pages 1685–1704, 2009.

T Tony Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Annals of statistics*, pages 898–924, 1999.

Bradley Efron and Carl Morris. Combining possibly related estimation problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 379–421, 1973a.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitorsan empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973b.

Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

Wenhua Jiang and Cun-Hui Zhang. Empirical bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 263–273. Institute of Mathematical Statistics, 2010.

Iain M Johnstone. Gaussian estimation: Sequence and wavelet models. *Unpublished manuscript*, 2011.

Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer, 1998.

Ker-Chau Li and Jiunn Tzon Hwang. The data-smoothing aspect of stein estimates. *The Annals of Statistics*, pages 887–897, 1984.

Zhuang Ma, Dean Foster, and Robert Stine. Adaptive monotone shrinkage for regression. *arXiv preprint arXiv:1505.01743*, 2015.

Omkar Muralidharan. An empirical bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, 4(1):422–438, 2010.

Zhiqiang Tan. Steinized empirical bayes estimation for heteroscedastic data. *Preprint*, 2014.

Zhiqiang Tan. Improved minimax estimation of a multivariate normal mean under heteroscedasticity. *Bernoulli*, 21(1):574–603, 02 2015.

Xianchao Xie, SC Kou, and Lawrence D Brown. Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.