1-1-2015

Empirical Bayes Estimation in Cross-Classified Gaussian Models With Unbalanced Design

Asaf Weinstein
University of Pennsylvania, assafweinstein@gmail.com

Recommended Citation

Weinstein, Asaf, "Empirical Bayes Estimation in Cross-Classified Gaussian Models With Unbalanced Design" (2015). *Publicly Accessible Penn Dissertations*. 2091.
http://repository.upenn.edu/edissertations/2091

Empirical Bayes Estimation in Cross-Classified Gaussian Models With Unbalanced Design

Abstract
The James-Stein estimator and its Bayesian interpretation demonstrated the usefulness of empirical Bayes methods in facilitating competitive shrinkage estimators for multivariate problems consisting of nonrandom parameters.

When transitioning from homoscedastic to heteroscedastic Gaussian data, empirical ``linear Bayes" estimators typically lose attractive properties such as minimaxity, and are usually justified mainly from Bayesian viewpoints.

Nevertheless, by appealing to frequentist considerations, traditional empirical linear Bayes estimators can be modified to better accommodate the asymmetry in unequal variance cases.

This work develops empirical Bayes estimators for cross-classified (factorial) data with unbalanced design that are asymptotically optimal within classes of shrinkage estimators, and in particular asymptotically dominate traditional parametric empirical Bayes estimators as well the usual (unbiased) estimator.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Lawrence D. Brown

Keywords
Compound Decision, Cross-Classified Models, Decision Theory, Empirical Bayes, Heteroscedasticity, Shrinkage Estimation

Subject Categories
Statistics and Probability

EMPIRICAL BAYES ESTIMATION IN CROSS-CLASSIFIED GAUSSIAN MODELS WITH UNBALANCED DESIGN

Asaf Weinstein

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

Lawrence D. Brown, Professor of Statistics

Graduate Group Chairperson

Eric T. Bradlow, Professor of Marketing, Statistics, and Education

Dissertation Committee

Lawrence D. Brown, Professor of Statistics

Edward I. George, Professor of Statistics

Mark G. Low, Professor of Statistics

EMPIRICAL BAYES ESTIMATION IN CROSS-CLASSIFIED GAUSSIAN MODELS WITH UNBALANCED DESIGN

ACKNOWLEDGEMENT

Foremost, I would like to express my deep gratitude to my advisor, Professor Larry Brown. I could have never completed this work without his guidance, amazingly knowledgeable advice and encouragement, that kept me on track for the past five years. I will forever be thankful for his generosity with time and for making his students a priority.

I would also like to sincerely thank Professor Cun-Hui Zhang, who I learned so much from in the last year and a half of my doctorate, and who has a significant part in what this thesis has become.

I would like to thank Professor Abba Krieger, Professor Yoav Benjamini, Professor Dani Yekutieli and my committee members, Professor Mark Low and Professor Ed George, for their involvement, attention and advice along the way.

I would also like to thank Professor Meir Feder for short conversations that had a huge influence on my interests and work.

I would like to thank Gourab Mukherjee for his commitment, time, advice, and for being more than just a colleague.

Finally, I would like to thank my family, first and foremost my parents, for being a part of this journey. I cannot express in words my thankfulness to their selfless understanding, love and support.

ABSTRACT

EMPIRICAL BAYES ESTIMATION IN CROSS-CLASSIFIED GAUSSIAN MODELS WITH UNBALANCED DESIGN

Asaf Weinstein

Lawrence D. Brown

The James-Stein estimator and its Bayesian interpretation demonstrated the usefulness of empirical Bayes methods in facilitating competitive shrinkage estimators for multivariate problems consisting of nonrandom parameters. When transitioning from homoscedastic to heteroscedastic Gaussian data, empirical “linear Bayes” estimators typically lose attractive properties such as minimaxity, and are usually justified mainly from Bayesian viewpoints. Nevertheless, by appealing to frequentist considerations, traditional empirical linear Bayes estimators can be modified to better accommodate the asymmetry in unequal variance cases. This work develops empirical Bayes estimators for cross-classified (factorial) data with unbalanced design that are asymptotically optimal within classes of shrinkage estimators, and in particular asymptotically dominate traditional parametric empirical Bayes estimators as well the usual (unbiased) estimator.

TABLE OF CONTENTS

ACKNOWLEDGEMENT iii

ABSTRACT iv

LIST OF TABLES vii

LIST OF ILLUSTRATIONS viii

PREFACE ix

CHAPTER 1 : Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean 1

1.1 Introduction 1

1.2 A Compound Decision Problem for the Heteroscedastic Case 6

1.3 A Spherically Symmetric Shrinkage Estimator 9

1.4 Group Linear Shrinkage Methods 12

1.5 Simulation Study 17

1.6 Real Data Example 21

1.7 Conclusion and Directions for Further Investigation 26

CHAPTER 2 : Empirical Bayes Estimates for a Cross-Classified Additive Model with Unbalanced Design 28

2.1 Introduction 28

2.2 Model Setup and Bayes Estimates 32

2.3 Computation of the SURE Estimator 38

2.4 The Balanced Case 39

2.5 Risk Properties of the SURE Estimator 43

2.6 Simulation Study 45

2.7 The Case of Missing Values 49

2.8 Concluding Remarks 54

APPENDIX 57

A.1 Supplements to Chapter 1 57

A.2 Supplements to Chapter 2 66

BIBLIOGRAPHY 79

LIST OF TABLES

TABLE 1 : Oracle shrinkage locations and shrinkage factors 19

TABLE 2 : Prediction Errors of Batting Averages 23

TABLE 3 : Estimated risks relative to the Least-Squares estimator, $L = 180$. 47

LIST OF ILLUSTRATIONS

FIGURE 1 : Estimated risk for various estimators vs. number of observations. 19

FIGURE 2 : Shrinkage vs. number of at-bats 25

FIGURE 3 : Estimated risk for various estimators vs. number of row/column levels 50

PREFACE

In 1956 Charles Stein published a landmark paper (Stein, 1956) which showed that the natural estimator of the mean of a normal vector with $n \geq 3$ independent components with a common and known variance is inadmissible under sum of squared errors loss. This result was strengthened when James and Stein (1961) gave an explicit form of an estimator whose risk is strictly smaller than that of the usual estimator for any value of the true parameter. The James-Stein estimator demonstrated a deficiency of the (loss-independent) classical methods of least squares and maximum likelihood, and revealed serious limitations of unbiased estimation in multivariate statistical problems. Stein's original discovery incited a flurry of work on shrinkage estimation in the thirty years that followed, with the main focus on developing minimax estimators and admissible estimators under various linear models and different loss criteria. The long list of references includes Stein (1966, 1973); Alam and Thompson (1964); Baranchik (1964, 1970); Bhattacharya (1966); Brown (1966); Thompson (1968); Sclove (1968); Strawderman (1971, 1978); Alam (1973); Bock (1975); Efron and Morris (1976); Berger (1976); Rolph (1976); Berger et al. (1977); George et al. (1986); among many others.

The James-Stein estimator was brought into fame first for being a minimax estimator different than (and hence dominating) the usual estimator. Nevertheless, the actual form of the estimator was a contribution in itself, uncovering the role of Bayesian procedures in constructing shrinkage estimators. The Bayesian interpretation of the James-Stein estimator was recognized already by Stein (1962) as an empirical version of what later became known as the Best Linear Unbiased Predictor (BLUP) in a random-effects model. That is, Stein referred to a hierarchical model with $X_i \sim N(\theta_i, 1)$ independently for $1 \leq i \leq n$ and the unobserved means θ_i themselves coming from an i.i.d. normal distribution with mean zero and common, unknown variance τ^2. The Bayes estimator (under squared loss) based on $X = (X_1, ..., X_n)^\top$ is

$$\widehat{\theta}_i = E_{\tau^2}(\theta_i|X) = \Big(1 - \frac{1}{\tau^2+1}\Big)X_i \tag{1}$$

which produces almost exactly the James-Stein estimator when $b^{-1} = \tau^2 + 1$ is replaced by the unbiased estimate $\|X\|^2/n$ (in fact, if an unbiased estimate is used for b instead of b^{-1}, the exact form of the James-Stein estimator is recovered; see, e.g. Morris et al., 2012). The Bayesian point of view was taken up in Lindley's discussion of Stein's paper (Lindley, 1962), and developed extensively by Efron and Morris in a sequence of papers (Efron and Morris, 1972a,b, 1973b) that promoted an empirical Bayes interpretation of the James-Stein estimator. In Efron and Morris (1973b) they suggested a derivation of Stein-type estimators for a normal mean vector, which was technically equivalent to that briefly mentioned by Stein, but offered another perspective. Efron and Morris considered a two-level model given by

$$\begin{aligned} &\theta_i \overset{iid}{\sim} G \\ &X_i|\theta_i \sim N(\theta_i, 1) \end{aligned} \tag{2}$$

where the distribution G is unknown. Under this setup they targeted the "linear Bayes" rule, namely, the linear rule in X that minimizes the Bayes risk. If $\tau^2 = \int \theta^2 \, G(d\theta)$ denotes the mean of θ_1^2 under G, the minimizer is given by (1), which can be "estimated" by the James-Stein estimator. Thus, from this point of view, the James-Stein estimator is an empirical "linear Bayes" rule for the situation described by the hierarchical model above. (We should remark that the account given above is a somewhat abused version of the source: Efron and Morris did not assume normality even for the likelihood function — which would not change the result; They also allowed the variance of $V_i = \mathrm{Var}(X_i|\theta_i)$ to be different for different coordinates, and even to depend on θ_i; And they considered in fact the optimal affine, not linear, predictor. Yet the simplification that we made suffices for the current discussion).

The pioneering work of Efron and Morris led the way to an empirical Bayes approach to multivariate problems, where strict "Model-I" (i.e., conditional on μ) minimaxity or admissibility are not *necessarily* a primary concern. Since then there has been a lot of effort, in recent years as well, to develop parametric, semi-parametric and non-parametric empirical Bayes procedures for homoscedastic and heteroscedastic normal means problems, some examples being Morris (1983); Edelman (1988); Zhang (1997); Brown and Greenshtein (2009); Jiang and Zhang (2009, 2010); Xie et al. (2012, 2015); Koenker and Mizera (2014).

The Bayesian derivation of the James-Stein estimator makes the need for shrinkage of individual components intuitive, but it still does not explain why the resulting empirical Bayes estimator would have good risk properties conditional on the θ_i. On the other hand, an explicit connection between the original frequentist problem and the Bayesian problem was made by Herbert Robbins. To demonstrate Robbins's ideas in the normal mean problem, consider "separable" estimators of the form $\widehat{\theta}_i(X) = t(X_i)$ for some common function t. If the sum of squares loss is normalized by n, the risk for such an estimator is

$$\frac{1}{n}E\|\widehat{\theta} - \theta\|^2 = \sum_{i=1}^{n} \frac{1}{n} E_{\theta_i}[t(X_i) - \theta_i]^2$$

which is exactly the Bayes risk

$$E[t(\widetilde{X}) - \widetilde{\theta}]^2$$

of the estimator $\widetilde{\theta} = t(\widetilde{X})$ where $(\widetilde{X}, \widetilde{\theta})$ is a pair of univariate random variables jointly distributed as (X_1, θ_1) in (2) with

$$G(A) = G_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(\theta_i \in A), \qquad A \subseteq \mathbb{R}.$$

In other words, estimating the best separable rule for the original $n-$dimensional problem is equivalent to the problem of estimating the Bayes rule for (2) with G taken to be the empirical distribution of the (unknown) nonrandom $\theta_i,\ i \leq n$. Robbins called a problem of the first kind a *compound decision* problem. He developed a theory of *empirical Bayes* to solve problems of the second kind with arbitrary G, after realizing the equivalence between the problems when $G = G_n$. An excellent review of the two intimately related topics is given in Zhang (2003).

Interestingly, Robbins presented such "shrinkage" ideas already in 1951 (Robbins, 1951). However, his target was different from Stein's. Put into the context of the homoscedastic normal means problem, Robbins's goal was to design an estimator $\widehat{\theta}_n$ such that for all sequences $\{\theta_n\}$ (that satisfy minor conditions),

$$\limsup_{n\to\infty} \left\{ R(\theta_n, \widehat{\theta}_n) - \inf_{\delta\in\mathcal{D}_n} R_n(\theta_n, \delta_n) \right\} \leq 0 \tag{3}$$

where

$$R(\theta_n, \widehat{\theta}_n) := \frac{1}{n} E_{\theta_n} \|\widehat{\theta} - \theta_n\|^2$$

and

$$\mathcal{D}_n = \{\delta : \delta_i(X) = t(X_i) \text{ for some function } t_n : \mathbb{R} \to \mathbb{R}\}. \tag{4}$$

Hence, Robbins aimed at an asymptotic goal — he did not address "finite-n" criteria as Stein did. In turn, his target was more ambitious, namely, to asymptotically attain the risk of an oracle who is allowed to base the choice of $\delta \in \mathcal{D}_n$ on the truth θ_n.

A more modest goal is achieving (3) where $\mathcal{D}_n$ of (4) is replaced by a smaller family of estimators $\mathcal{D}_n' \subset \mathcal{D}_n$, for example this could be some parametric or semi-parametric family. The discussion in the two previous paragraphs implies that if $\mathcal{D}_n^L$ is taken to be the family of all estimators for which $\widehat{\theta}_i(X) = t(X_i)$ and t is also required to be *linear*, then the empirical Bayes (or *parametric* empirical Bayes, as it was referred to by Morris, 1983) derivation of the James-Stein estimator based on (2) will also serve to achieve (3) with $\mathcal{D}_n$ replaced by $\mathcal{D}_n^L$ (again, some regularity conditions will be needed for the sequence $\{\theta_n\}$). In other words, for $X \sim N_n(\theta, I)$, the goal of asymptotically attaining the performance of the best separable, linear oracle is aligned with (non-asymptotic) minimaxity. (It can be shown that the James-Stein estimator also satisfies a *non*-asymptotic oracle inequality; See Johnstone, 2011, Section 2.7, Corollary 2.6).

Alas, the situation is not as favorable for heteroscedastic data, $X \sim N_n(\theta, D)$ for a known covariance matrix $D = \text{diag}(V_1, ..., V_n)$ (with usual sum of squared errors used as the loss function). The problem is, essentially, that the minimaxity requirement generally limits considerably the largeness of "sensible" families $\mathcal{D}'_n$ of *linear* rules (not necessarily subsets of (4)) for which (3) can be achieved. The situation is similar when θ is known a-priori to lie in some linear subspace, for example in a two and higher-way cross-classified model with unbalanced design.

Nevertheless, if willing to settle for only *asymptotic* minimaxity, or if non-linear classes $\mathcal{D}'_n$ are considered, (3) can be achieved for much more interesting families. This is the main thrust of our work. We focus on designing empirical Bayes estimators for additive, cross-classified Gaussian models with unequal cell counts, that asympotically achieve (3) where D_n is an appropriate class of parametric or semi-parametric shrinkage estimators. As (3) is a frequentist criterion, we use frequentist considerations to modify standard parametric empirical Bayes procedures that rely on the usual random-effects Gaussian model to produce shrinkage estimators.

These considerations are different for the one-way layout and for the higher-way layout. For the one-way unbalanced layout, Chapter 1 develops empirical Bayes estimators motivated from a compound-decision perspective. It is shown that, under appropriate conditions, our estimator achieves asymptotic oracle optimality with respect to the semi-parametric class

$$\mathcal{D}_n^{SP} = \left\{ \widehat{\theta}_i = Y_i - \frac{V_i}{V_i + g(V_i)} \big(Y_i - m(V_i) \big) : \ g \geq 0, \ m \ \text{ are any real-valued functions} \right\}$$

and at the same time is minimax for all n.

In the two-way unbalanced layout our results apply to the usual parametric family of Bayes estimators, that arise from using a prior reflecting within-factor exchangeability. Specifically, in the second chapter we extend the one-way results of (Xie et al., 2012) and show that for estimating the true cell means, under appropriate conditions, our estimator achieves asymptotic oracle optimality with respect to this parametric family of Bayes estimators. The practically important case of missing values is also treated. The approach immediately extends to the higher-way additive layout, although the computational effort in implementing the estimator may become serious for even moderately large number of factors.

CHAPTER 1 : Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean

Joint work with Zhuang Ma, Lawrence D. Brown and Cun-Hui Zhang

1.1. Introduction

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

$$X_i|(\theta_i, V_i) \sim N(\theta_i, V_i) \tag{1.1}$$

independently for $1 \leq i \leq n$. This includes the case of nonrandom $\boldsymbol{\theta}$ and $\boldsymbol{V}$. In the heteroscedastic normal mean problem, the goal is to estimate the vector $\boldsymbol{\theta}$ based on $\boldsymbol{X}$ and $\boldsymbol{V}$. Hence we assume that in addition to the random observations $X_1, ..., X_n$, the variances $V_1, ..., V_n$ are available. For squared loss, $L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \frac{1}{n}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \frac{1}{n}\sum_{i=1}^n(\widehat{\theta}_i - \theta_i)^2$, this problem has been widely studied for both the special case of equal variances, $V_i \equiv \sigma^2$, and the more general case above, and alternative estimators to the usual (Maximum Likelihood) estimator $\widehat{\boldsymbol{\theta}}^{ML}(\boldsymbol{X}) = \boldsymbol{X}$ have been suggested that perform better, in some sense, in terms of the risk $R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \mathbb{E}[L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}(\boldsymbol{X}))|\boldsymbol{\theta}]$, regardless of $\boldsymbol{\theta}$. Here and elsewhere, unless otherwise stated, we suppress in notation the dependence of the risk function on $\boldsymbol{V}$.

In the homoscedastic case such shrinkage estimators go back, of course, to the James-Stein estimator,

$$\widehat{\boldsymbol{\theta}}^{JS}(\boldsymbol{X}) = \left(1 - \frac{(n-2)\sigma^2}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X} \tag{1.2}$$

which, for $n \geq 3$, has strictly smaller risk than $\widehat{\boldsymbol{\theta}}^{ML}$ for any $\boldsymbol{\theta}$. This estimator can be derived as an empirical Bayes estimator under a model that puts $\boldsymbol{\theta} \sim N_n(\boldsymbol{0}, \gamma\boldsymbol{I})$, independently of $\boldsymbol{V}$, where γ is unspecified and "estimated" from the data $\boldsymbol{X}$. Equivalently, as observed in 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 $\boldsymbol{\theta}$ is only assumed to have i.i.d. 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 $\boldsymbol{\theta}$ really is random with i.i.d. components. Efron and Morris (1973b, Section 9) analyze and quantify relative savings in Bayes risk when using the true linear Bayes rule versus the James-Stein rule.

What is more, the James-Stein estimator has certain attractive asymptotic optimality properties *uniformly* in $\boldsymbol{\theta}$. Let $\mathcal{D}_S = \{\widehat{\boldsymbol{\theta}} : \widehat{\theta}_i(\boldsymbol{X}) = t(X_i) \text{ for some } t : \mathbb{R} \to \mathbb{R}\}$. We say that an estimator is *simple* if $\widehat{\theta}_i(\boldsymbol{X}) = t_i(X_i)$ for functions $t_i : \mathbb{R} \to \mathbb{R}$. We say that an estimator is *symmetric* if $\widehat{\boldsymbol{\theta}}(\tau(\boldsymbol{X})) = \tau(\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))$ for all permutation operators τ. Then $\mathcal{D}_S$ is the class of simple, symmetric estimators. If $\widetilde{\mathcal{D}}_S$ denotes the class of estimators in $\mathcal{D}_S$ that are also linear in $\boldsymbol{X}$, it holds that for all $\boldsymbol{\theta}$ (with a mild restriction on the sequence $\theta_i,\ i = 1, 2, ..$),

$$R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{JS}) = \inf\{R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) : \widehat{\boldsymbol{\theta}} \in \widetilde{\mathcal{D}}_S\} + o(1) = R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{b_n^*}) + o(1) \tag{1.3}$$

where $\widehat{\boldsymbol{\theta}}^b(\boldsymbol{X}) = (1-b)\boldsymbol{X}$ and $b_n^* = \arg\min_b R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^b)$. Herbert Robbins was the first to seek for decision rules that exhibit asymptotic oracle performance of the kind exhibited above, although Robbins considered the entire family of simple and symmetric rules (Robbins, 1951; Zhang, 2003). As observed by Robbins, the striking fact that the property (1.3) is possible without knowing $\boldsymbol{\theta}$ 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. Indeed, as presented in Zhang (2003), for $\widehat{\boldsymbol{\theta}} \in \mathcal{D}_S$ with $\widehat{\theta}_i(\boldsymbol{X}) = t(X_i)$,

$$R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \frac{1}{n} E_{\theta_i}[t(X_i) - \theta_i]^2 = E[t(X) - \theta]^2 \tag{1.4}$$

where the expectation in the last term is taken over the pair (θ, X) of random variables jointly distributed according to

$$\theta \sim G = \frac{1}{n}\sum_{i=1}^{n} I\{\theta_i \leq \theta\}, \qquad X|\theta \sim N(\theta, \sigma^2).$$

As such, the problem is equivalent to a one-dimensional Bayesian estimation problem, and the optimal rule in $\widetilde{\mathcal{D}}_S$ has $\widehat{\theta}_i^*(\boldsymbol{X}) = (1 - b_n^*)X_i$ where $(1 - b_n^*)X$ is the best linear predictor of the random variable θ based on the random variable X, namely $b_n^* = \sigma^2/E_{\boldsymbol{\theta}}(X^2)$. While b_n^* depends on $\boldsymbol{\theta}$, this dependence is only through $1/E_{\boldsymbol{\theta}}(X^2)$, which for large n is well approximated by $(n-2)/\|\boldsymbol{X}\|^2$. This estimator is exactly unbiased for $1/E_{\boldsymbol{\theta}}(X^2)$ under $\boldsymbol{\theta} = \mathbf{0}$.

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. Indeed, existing parametric empirical Bayes estimators, which usually start by putting again 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, essentially undershrinking on 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 have been suggested, that is, estimators that shrink all components by the same factor regardless of V_i; These exist only when the V_i satisfy certain conditions that restrict how much they can be spread out. A precise result was given by Brown (1975). See Tan (2015) for a concise review of some existing estimators and references therein for related literature.

There have been attempts to moderate the respective disadvantages of estimators resulting from either of the two approaches. Xie et al. (2012, XKB hereafter) considered empirical Bayes estimators arising from the hierarchical model

$$\theta_i \overset{\text{iid}}{\sim} N(\mu, \gamma) \qquad X_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, V_i) \qquad 1 \leq i \leq n \tag{1.5}$$

with unspecified μ and γ. They suggested to plug into the Bayes rule

$$\widehat{\theta}_i^{\mu,\gamma} = \mathbb{E}_{\mu,\gamma}(\theta_i|X_i) = X_i - \frac{V_i}{V_i+\gamma}(X_i - \mu) \tag{1.6}$$

the values

$$(\widehat{\mu}, \widehat{\gamma}) = \underset{\mu,\gamma}{\arg\min}\, \mathcal{R}(\mu, \gamma; \boldsymbol{X})$$

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 (1.5) is, as compared to the standard approach that uses Maximum Likelihood or Method-of-Moments estimates of μ, γ under (1.5). 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). This improves Bayesian performance while retaining minimaxity. Tan (2015) recently suggested a minimax estimator with a simpler form, that has similar properties.

As in (1.3), empirical Bayes rules resulting from an exchangeable prior on $\boldsymbol{\theta}$ are well motivated in the homoscedastic case even when θ_i are deterministic, owing to the symmetry of the decision problem with respect to the components $1 \le i \le n$. Indeed, together with the additivity of the loss function, the fact that

$$X_i \sim f(x; \theta_i), \qquad 1 \le i \le n \tag{1.7}$$

for a common distribution f, allows us to write the risk of $\widehat{\boldsymbol{\theta}} \in \widetilde{\mathcal{D}}_S$ as a Bayes risk, and hence set the minimum linear Bayes risk as a benchmark for all $\widehat{\boldsymbol{\theta}} \in \widetilde{\mathcal{D}}_S$. In the unequal variances case, on the other hand, the problem does not immediately admit a compound decision structure as before, because instead of (1.7) we now have

$$X_i \sim f_i(x; \theta_i), \qquad 1 \le i \le n$$

where $f_i(x;\theta) = (2\pi V_i)^{-1/2} \exp[-(x-\theta)^2/(2V_i)]$, violating the symmetry referred to above. Consequently, if $\widehat{\theta}_i(\boldsymbol{X}, \boldsymbol{V}) = t(X_i, V_i)$ is allowed to depend on X_i and V_i only, then it is not immediately evident what oracle rule might set a reasonable benchmark for an empirical Bayes estimator. This raises some questions, for example: how well can the approach pursued by any empirical Bayes estimator starting from (1.5) ever expect to perform? Is there a more ambitious goal that is still asymptotically achievable?

We show that symmetry can be restored in the heteroscedastic case to produce a counterpart of (1.4), which, in turn, gives rise to a useful benchmark. In essense, our observation comes from taking a point of view in which the "observed data" associated with the unknown parameter θ_i is the pair (X_i, V_i) instead of just X_i. This will lead to a connection between the risk of an estimator $\widehat{\theta}_i(\boldsymbol{X}, \boldsymbol{V}) = t(X_i, V_i)$ and the Bayes risk of the estimator $t(X, V)$ for a random triplet (X, θ, V), where $X|(\theta, V) \sim N(\theta, V)$ and the joint distribution of θ and V is determined by $(\theta_i, V_i),\ 1 \le i \le n$.

We then take a similar approach to Efron and Morris (1973b) in setting out to mimic the rule $t(X, V)$ with smallest Bayes risk among all rules that are linear in X, with no normality assumption on the distribution of $\theta|V$. We suggest an empirical Bayes block-linear estimator, that groups together observations with similar variances and applies a spherically symmetric minimax estimator to each group separately. A qualitative desctiption of our results follows in the next section.

The chapter is organized as follows. In Section 1.2 we present the estimation of a heteroscedastic mean as a compound decision problem, for simple, symmetric estimators. Section 1.3 presents a spherically symmetric minimax estimator for a heteroscedastic normal vector. Our group-linear empirical Bayes estimator is introduced in Section 1.4, where we discuss its properties and prove two oracle inequalities that establish its asymptotic optimality within a class in the case where $(X_i, \theta_i, V_i),\ i \le n$ are independent and identically distributed. In Section 2.6 we present a simulation study, and in Section 1.6 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 are generally deferred to the Appendix.

1.2. A Compound Decision Problem for the Heteroscedastic Case

Let $\boldsymbol{X}, \boldsymbol{\theta}$ and $\boldsymbol{V}$ be as in (1.1). Denote by $\mathcal{D}_S$ the set of all simple and symmetric estimators in $(\boldsymbol{X}, \boldsymbol{V})$, namely, $\widehat{\theta}_i(\boldsymbol{X}, \boldsymbol{V}) = t(X_i, V_i)$ for some function t (we reuse the notation $\mathcal{D}_S$ from the previous section for simplicity, hoping this will cause no confusion). If $\widehat{\boldsymbol{\theta}} \in \mathcal{D}_S$ with $\widehat{\theta}_i(\boldsymbol{X}, \boldsymbol{V}) = t(X_i, V_i)$, then

$$R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \frac{1}{n} E_{\theta_i}[t(X_i, V_i) - \theta_i]^2 = E[t(X, V) - \theta]^2 \tag{1.8}$$

where the expectation in the last term is taken over the random vector $(X, \theta, V, I)^\top$ distributed according to

$$\mathbb{P}(I = i) = \frac{1}{n}, \qquad (X, \theta, V)|(I = i) \overset{d}{=} (X_i, \theta_i, V_i) \qquad 1 \leq i \leq n. \tag{1.9}$$

where $\overset{d}{=}$ means equal in distribution. We emphasize the distinction throughout between the vectors $\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V}$ and the random variables X, θ, V. In particular, $\boldsymbol{X}$ is a random vector with random components $X_i,\ 1 \leq i \leq n$, but $\boldsymbol{\theta}$ and $\boldsymbol{V}$ may be nonrandom vectors; whereas $X,\ \theta$ and V are always random variables, by (1.9).

Again we stress that (1.8) holds also when the pairs (θ_i, V_i) are deterministic. The identity (1.8) is easily verified by calculating the expectation on the right hand side when first conditioning on I, and says that for a simple, symmetric estimator in $(\boldsymbol{X}, \boldsymbol{V})$, the risk is again equivalent to the Bayes risk in a one-dimensional estimation problem. Note that (1.8) can be interpreted as an application of (1.4) to a compound decision problem as originally intended by Robbins - consisting of n *identical* copies of a single decision problem - except that the data associated with the unknown parameter θ_i is now the *pair* (X_i, V_i) with a distribution given by the conditional distribution of $(X, V)|(\theta = \theta_i)$ in (1.9).

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 will use the former word for convenience) in X,

$$\widehat{\theta}_i^{a,b}(\boldsymbol{X}, \boldsymbol{V}) = X_i - b(V_i)[X_i - a(V_i)] \qquad 1 \leq i \leq n. \tag{1.10}$$

The corresponding Bayes risk in (1.8) is

$$r_n(a, b) \triangleq \mathbb{E}\Big\{X - b(V)[X - a(V)] - \theta\Big\}^2. \tag{1.11}$$

Since

$$X|(\theta, V) \sim N(\theta, V), \tag{1.12}$$

the minimizers of

$$r_n(a, b|v) \triangleq \mathbb{E}\Big\{\Big(X - b(v)[X - a(v)] - \theta\Big)^2 \Big| V = v\Big\}, \tag{1.13}$$

and hence also of (1.11), are

$$a_n^*(v) = \mathbb{E}(X|V = v), \quad b_n^*(v) = \frac{v}{\mathrm{Var}(X|V = v)} \tag{1.14}$$

and the minimum Bayes risk is

$$R_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{a_n^*, b_n^*}) = r_n(a_n^*, b_n^*) = \mathbb{E}\Big[V\big\{1 - b_n^*(V)\big\}\Big]. \tag{1.15}$$

Therefore, (1.15) is a lower bound on the risk achievable by any estimator of the form (1.10), and $\widehat{\boldsymbol{\theta}}^{a_n^*, b_n^*}$ is the optimal such decision rule. Note that any estimator of the form (1.6) is also of the form (1.10), but not vice versa.

To highlight the difference between the oracle of the form (1.10) and an oracle of the form (1.6), the connection to a one-dimensional Bayesian problem in (1.8) allows us to focus on a 3-tuple of random variables (X, θ, V) with the known (since oracle rules are considered now) joint distribution (1.9). Hence, X and V are observed and an estimator $t(X, V)$ incurs loss $(t(X, V) - \theta)^2$. The optimal rule linear in X is

$$t^*(X, V) = X - \frac{V}{V + \gamma_n^*(V)}(X - \mu_n^*(V)) \tag{1.16}$$

where $\mu_n^*(v) = \mathbb{E}(\theta|V = v)$ and $\gamma_n^*(v) = \mathrm{Var}(\theta|V = v)$; this is just rewriting of (1.14) in terms of $\mu_n^*(\cdot)$ and $\gamma_n^*(\cdot)$ instead of $a_n^*(\cdot)$ and $b_n^*(\cdot)$, which is convenient for the purpose of the current discussion. In contrast, the oracle rule of the form (1.6) looks for the best *constants* γ_n, μ_n in (1.16). If θ and V are independent, $\gamma_n^*(v)$ and $\mu_n^*(v)$ are indeed constant in v, and the oracle rules coincide. However, if θ and V are not independent, (1.16) might have strictly smaller risk. The estimator (1.16) allows different shrinkage factor (through $\gamma_n^*(v)$) and location (through $\mu^*(v)$) for different values of v, as opposed to using a common shrinkage factor and location (regardless of v). To conclude, we demonstrate these differences in an example.

Example 1.2.1 (XKB, Section 7, Example 5)**.** (X, θ, V) are distributed so that $V \sim 0.5 \cdot 1_{\{V=0.1\}} + 0.5 \cdot 1_{\{V=0.5\}}$, $\theta|(V = 0.1) \sim N(2, 0.1)$, $\theta|(V = 0.5) \sim N(0, 0.5)$ and $X \sim N(\theta, V)$. The best rule $t(X, V)$ which is linear in X, i.e., the rule of that form with minimum Bayes risk, is

$$t^*(X, V) = \begin{cases} \frac{X}{2} + 1 & V = 0.1 \\ \frac{X}{2} & V = 0.5\ . \end{cases}$$

This is easily seen noting that conditionally on V the usual normal-normal problem (with only θ random) arises. The corresponding Bayes risk is $\mathbb{E}[V(1 - 1/2)] = 0.15$. On the other hand, the best rule of the form $t(X, V) = X - \frac{V}{V+\gamma}(X - \mu)$ has $\gamma \approx 0.83$ and $\mu \approx 0.15$, with Bayes risk ≈ 0.194, about 30% higher than that of the best linear-in-x rule.

Our results may now be described more precisely. We suggest an estimator which (i) is minimax for all n and, under some conditions, (ii) asymptotically achieves the oracle risk

(1.15) when $(X_i, \theta_i, V_i), 1 \leq i \leq n$ are i.i.d. from some population with $X_i|(\theta_i, V_i) \sim N(\theta_i, V_i)$. Note that if (X_i, θ_i, V_i) are i.i.d., the functions a_n^* and b_n^* and the corresponding risk $r_n(a_n^*, b_n^*)$ indeed do not depend on n. In the case $r(a^*, b^*) = 0$, Theorem 1.4.3 also gives a rate of converges under appropriate smoothness conditions on the functions a^*, b^*. Although it is not considered in the current work, an analogue of (ii) 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 θ_i and V_i. In this case, to ensure that the limit does not depend on n, suppose that the empirical joint distribution G_n of $\{(\theta_i, V_i) : 1 \leq i \leq n\}$ has a limiting distribution G. Define the risk for candidates a_n, b_n to be computed with respect to G. Then Theorem 1.4.3 will say that our estimator has $r(\widehat{a}_n, \widehat{b}_n) \to r(a^*, b^*)$ under appropriate conditions on a^*, b^*.

Finally, a comment is in place regarding nonparametric estimators. Existing nonparametric empirical Bayes estimators, such as the semiparametric estimator of XKB and the nonparametric method of Jiang and Zhang (2010), target the best predictor $g(X, V)$ of θ where g is restricted to some nonparametric class of functions. While the optimal g may indeed be a non-linear function of X, these methods implicitly assume independence between θ and V. If under the the distribution in (1.9), θ and V are "far" from independent, these methods can still suffer from the gap between the optimal predictor $g(X, V)$ assuming independence, and the *true* Bayes rule, namely, $\mathbb{E}(\theta|X, V)$. Therefore, in some cases the oracle rule (1.16) might still have smaller expected loss than the oracle choice of g computed under independence of θ and V.

1.3. A Spherically Symmetric Shrinkage Estimator

In this section suppose that $\boldsymbol{\theta}, \boldsymbol{V}$ and $\boldsymbol{X}$ are as in (1.1) where $\boldsymbol{\theta}$ is nonrandom and unknown and $\boldsymbol{V}$ is nonrandom and known. We present a family of spherically symmetric estimators that shrink toward a data-dependent location. This will serve as a building block for the group-linear estimator of the following section. The version of our estimator that shrinks toward the origin, and sufficient conditions for its minimaxity, were given by Brown (1975) and are reviewed in Tan (2015).

We will need the following definitions before we state the next result. Suppose that $\boldsymbol{X}, \boldsymbol{\theta}$ and $\boldsymbol{V}$ as in (1.1) with nonrandom $\boldsymbol{\theta}$ and $\boldsymbol{V}$, and where $\boldsymbol{V}$ is known. Let

$$\overline{X} = \sum_{i=1}^{n} X_i, \qquad s_n^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$$
$$\overline{V} = \sum_{i=1}^{n} V_i/n, \qquad V_{\max} = \max_{i \le n} V_i.$$

and

$$c_n^* \;=\; \{[(n-3) - 2(V_{\max}/\overline{V} - 1)]/(n-1)\}_+ \;=\; \{1 - 2(V_{\max}/\overline{V})/(n-1)\}_+.$$

Then define a spherically symmetric estimator $\widehat{\boldsymbol{\theta}}^c$ by $\widehat{\theta}_i^c = X_i$ if $n = 1$, and otherwise

$$\widehat{\theta}_i^c = X_i - \widehat{b}(X_i - \overline{X}), \quad \widehat{b} = \min\left(1, c_n\overline{V}/s_n^2\right) \tag{1.17}$$

Lemma 1.3.1. *For* $0 \le c_n \le 2c_n^*$,

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big(\widehat{\theta}_i^c - \theta_i\Big)^2 \\ &\le \overline{V}\left[1 - (1 - 1/n)\,\mathbb{E}\left\{(2c_n^* - c_n)\widehat{b} + (2 - 2c_n^* + c_n - s_n^2/\overline{V})I_{\{s_n^2/\overline{V} \le c_n\}}\right\}\right] \\ &\le \overline{V}. \end{aligned} \tag{1.18}$$

Remarks:

1. In (2.16) 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 $c_n = c_n^*$.

2. 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 $(\sum_{i=1}^n X_i/V_i)/(\sum_{i=1}^n 1/V_i)$. This can be used in place of $\overline{X}$ in (1.17). However, in the following section we will use $\widehat{\boldsymbol{\theta}}^c$ only on subsets of observations with similar variances; Hence for our use, the

difference will not be significant, especially under the continuity assumption on $a^*(\cdot)$ in theorm (1.4.2).

3. 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 an unknown mean. In the heteroscedastic case, a sufficient condition for minimaxity of the version of the estimator above that shrinks toward zero, is reported in Tan (2015) as $0 \le c_n \le 2\{1 - 2(V_{\max}/\overline{V})/n\}$. This is consistent with Lemma 1.3.1.

Proof of Lemma 1.3.1. To carry out the analysis, it suffices to consider $0 < c_n \le 2c_n^*$. Let $b(x) = \min(1, c_n\overline{V}/x)$ so that $\widehat{b} = b\big(s_n^2\big)$. Because $(\partial/\partial X_i)s_n^2 = 2(X_i - \overline{X})/(n-1)$, Stein's lemma yields

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

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

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

□

Estimators $\widehat{\boldsymbol{\theta}}^c$ in the family described above have a risk function that never exceeds $\bar{V}$, but its usefulness in the heteroscedastic case is limited because it includes only the usual estimator $\widehat{\boldsymbol{\theta}}^{ML}$ unless $c_n^* > 0$, i.e., $V_{\max}/\bar{V} < (n-1)/2$. The estimators of $\boldsymbol{\theta}$ that we suggest in the following sections, however, only use block-wise versions of $\widehat{\boldsymbol{\theta}}^{c}$, applying it separately to subsets of observations with similar variances V_i. The magnitude of $V_{\max}/\bar{V}$ may be large when the entire vector $\boldsymbol{V}$ is considered; But when $\boldsymbol{V}$ is partitioned, this ratio is more or less controlled on each bin. Hence the estimator $\widehat{\boldsymbol{\theta}}^{c}$ is potentially much more useful, and likely to provide actual shrinkage.

1.4. Group Linear Shrinkage Methods

Sections 1.1 and 1.3 set the stage for introducing an empirical Bayes estimator, which employs the spherically symmetric estimator to mimic the oracle rule $\widehat{\boldsymbol{\theta}}^{a^*,b^*}$. When the number of distinct values V_i is very small compared to n, as in Example 1.2.1, it is natural to mimic the oracle rule (1.16) by applying a James-Stein estimator separately to each group of homoscedastic observations. As we will show, under appropriate conditions, this estimator asymptotically approaches the oracle risk (1.14). Moreover, as long as the size of any sub-group is bigger than 3, this estimator has risk strictly smaller than the minimax risk $\overline{V}$.

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, but the expression for the optimal function a^* and b^* in (1.14) suggests grouping together observations with *similar* variances V_i, and then applying a James-Stein-type estimator separately to each group. The spherically symmetric estimator of section 1.3 is an appropriate candidate to use for each of the separate groups, as the variances are only approximately, but not exactly, equal to each other. The resulting estimator is also minimax, as it is minimax on each group by Lemma 1.3.1 (in fact, is likely to attain strictly smaller risk than $\overline{V}$ since c_n^*, at least for some intervals, is likely to be strictly positive).

Before defining our group-linear estimator, we remark that block-linear shrinkage has been suggested before for the homoscedastic case by Cai (1999) as an alternative to block-thresholding estimators in the context of wavelet estimation. We mention this approach because of the similarity in structure to our heteroscedastic mean estimator; otherwise, the estimator of Cai (1999) is motivated from an entirely different perspective, and addresses a very different oracle rule (which is itself a blockwise rule, unlike the oracle associated with our procedure). On the other hand, Tan (2014) comments briefly that block shrinkage methods building on his own "minimax Bayes" estimator can be considered to allow different shrinkage patterns for observations with different sampling variances. This is very much in line with the approach we pursue in the current paper.

Definition 1.4.1 (Group-linear Empirical Bayes Estimator for a Heteroscedastic Mean)**.** Let $J_1, \ldots, J_m$ be disjoint intervals and denote

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

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

$$\widehat{\theta}_i^{GL} = \begin{cases} X_i - \min\left(1, c_k \overline{V}_k / s_k^2\right)(X_i - \overline{X}_k), & i \in \mathcal{I}_k \\ X_i, & \text{otherwise} \end{cases} \tag{1.19}$$

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 $c_k = 0$.

Remark. The estimator in definition 1.4.1 is technically not an affine function on a particular interval, as the shrinkage factor $b_k = \min\left(1, c_k \overline{V}_k / s_k^2\right)$ depends on the data $\boldsymbol{X}$. In fact b_k is a highly non-linear function of $\boldsymbol{X}$, involving s_k^2 and a truncation (the estimator is a "positive-part" estimator on each interval). Nevertheless, we call the estimator "group-linear" because it is affine up to the dependency of b_k on $\boldsymbol{X}$.

Theorem 1.4.2. *Let $r(a, b)$ be as defined in* (1.11)*, a^* and b^* as defined in* (1.14) *and c_n^* as defined in Lemma 1.3.1. For $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{GL}$ in Definition 1.4.1 with $c_n = c_n^*$ the following holds.*

1. *Under the Gaussian model* (1.1) *with deterministic* $(\theta_i, V_i), i \le n$, *the risk of* $\widehat{\boldsymbol{\theta}}$ *is no greater than that of the naive estimator* $\widehat{\boldsymbol{\theta}}^{ML}$ *and therefore* $\widehat{\boldsymbol{\theta}}$ *is minimax*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(\widehat{\theta}_i - \theta_i\Big)^2 \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(X_i - \theta_i\Big)^2 = \frac{1}{n}\sum_{i=1}^{n} V_i = \overline{V}. \tag{1.20}$$

2. *Let* $(X_i, \theta_i, V_i), i = 1, \ldots, n$, *be iid vectors from a population* (X, θ, V) *satisfying (1.12). Then*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{i=1}^{n} r(a^*, b^*|V_i) + o(1) \tag{1.21}$$

for any sequence $\boldsymbol{V} = (V_1, ..., V_n)$ *such that the following conditions hold: With* $|J|$ *being the length of interval* J,

$$\begin{gathered}\max_{1\le k\le m}|J_k| \to 0,\ \min_{1\le k\le m} n_k \to \infty \\ a^*(v), b^*(v)\ \textit{are uniformly continuous} \\ \limsup_{n\to\infty}\frac{\sum_{i=1}^{n} V_i}{n} < \infty,\ \limsup_{n\to\infty}\frac{\sum_{i=1}^{n} V_i I_{\{V_i \notin \cup_{k=1}^{m} J_k\}}}{n} = 0\end{gathered} \tag{1.22}$$

Remark 1. The continuity of shrinkage factor and location $a^*(v), b^*(v)$ 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 at the place where 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 V. It ensures that $\max_{1\le k\le m}|J_k| \to 0,\ \min_{1\le k\le m} n_k \to \infty$ is 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), guaranteeing that their contribution to the normalized risk is negligible.

Remark 2. A statement regarding the marginal Bayes risk, when expectation is taken over $\boldsymbol{V}$ in (1.21), can be obtained in a similar way if replacing the conditions on the individual sequence $\boldsymbol{V}$ with bounded expectation of the random variable V. We skip this for simplicity.

For the i.i.d. situation of the second part of theorem (1.4.2), the case $r(a^*, b^*) = 0$ corresponds to $\theta = a^*(V)$, a deterministic function of V (equivalently, $b^*(V) \equiv 1$), and calls for a sharper result than (1.21) regarding the rate of convergence of the excess risk. Note that, when $\boldsymbol{\theta}$ and $\boldsymbol{V}$ are deterministic, $\theta = a^*(V)$ if and only if there are no two distinct values of θ_i with the same variance V_i, in which case the oracle rule indeed sets $a^*(V_i) = \theta_i, b^*(V_i) = 1$ and incurs zero loss. The precision in estimating the function a^*, secondary to that in estimating b^* when $r(a^*, b^*) > 0$, is crucial now. Noting that, trivially, $\theta = a^*(v)$ implies $\mathbb{E}(\theta|V = v) = a^*(v)$,

$$X_i|V_i \sim N(a^*(V_i), V_i) \tag{1.23}$$

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 (1.21) depends primarily on the smoothness of the function $a^*(v)$. We will say that a function $f : \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} \subset \mathbb{R}$ is L-Lipschitz continuous of order $\alpha > 0$ if $|f(x) - f(y)| \leq L(|x-y|)^{\alpha}$. If $\alpha = 1$, we will simply say that f is L-Lipschitz continuous.

The following theorem states that our group-linear estimator attains the optimal convergence rate under a Lipschitz condition, at least when V is bounded. In the homoscedastic case the smoothing feature of the James-Stein estimator was studied in Li and Hwang (1984).

Theorem 1.4.3. *Let $(X_i, \theta_i, V_i), i = 1, \ldots, n$, be iid vectors from a population (X, θ, V) satisfying (1.12). If $r(a^*, b^*) = 0$ and $a^*(\cdot)$ is L-Lipschitz continuous, then the group linear estimator in Definition 1.4.1 with equal block size $|J_k| = |J| = \left(\frac{11V_{\max}^2}{nL}\right)^{\frac{1}{3}}$ and $c_n = c_n^*$*

attains optimal nonparametric rate of convergence

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big]\leq 2\Big(\frac{11V_{\max}^2\sqrt{L}}{n}\Big)^{\frac{2}{3}}. \tag{1.24}$$

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

For the asymptotic results in Theorems 1.4.2 and 1.4.3 to hold, it is enough to choose bins J_k of equal length $|J|=\big(\frac{11V_{\max}^2}{nL}\big)^{\frac{1}{3}}$. 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, Lemma 1.3.1 and the first remark that follows it, suggest that binning such that $\big(\max\{V_i : i\in J_k\}\big)/\overline{V}_k$, rather than $\max\{V_i : i\in J_k\}-\min\{V_i : i\in J_k\}$, is fixed, is more appropriate. Hence we propose to bin observations to windows of equal lengths in $\log(V_i)$ instead of V_i.

Furthermore, the constant multiplying $n^{-1/3}$ in $|J|=\big(\frac{11V_{\max}^2}{nL}\big)^{\frac{1}{3}}$, is appropriate when the V_i range between 0 and 1; Otherwise, we suggest to scale the partition to the range of the V_i by fixing the *number* of bins to $n^{1/3}$, i.e., divide $\log(V_i)$ to bins of equal length $|\text{range}\{\log(V_i)\}|/n^{1/3}$. 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 (cf. Efron and Morris, 1973a).

More ambitiously, one might try to choose the common bin length (or bins of unequal length) with a data-dependent method, for example, by considering a group-linear estimator $\widehat{\boldsymbol{\theta}}^k$ using k equal-length bins on $\log(V_i)$, and ultimately setting $\widehat{\boldsymbol{\theta}}=\widehat{\boldsymbol{\theta}}^{k^{SURE}}$ with

$$k^{SURE}=\arg\min\mathcal{R}(k;\boldsymbol{X})$$

and where $\mathcal{R}(k;\boldsymbol{X})$ is an unbiased estimator of the risk of $\widehat{\boldsymbol{\theta}}^k$. The disadvantage of such data-based methods is that the minimaxity of the group-linear estimator is typically lost. On the other hand, minimaxity is preserved when the values of V_i, but not X_i, are used in deciding how to bin the observations, and it certainly makes sense to use this information to choose bins J_k of unequal lengths, when it seems appropriate from the empirical distribution of V_i.

1.5. Simulation Study

In this section we carry out a simulation study using the examples of XKB, and compare the performance of our group-linear estimator to the methods proposed in their work. In each example, $(X_i, \theta_i, V_i),\ 1 \leq i \leq n$ are drawn i.i.d. from a joint distribution such that $X_i | (\theta_i, V_i) \sim N(\theta_i, V_i)$; various estiamtors are then applied to the data $(X_i, V_i), 1 \leq i \leq n$, and the normalized sum of sqrared error is computed (the last example is the only exception, with X_i drawn from a different distribution than $N(\theta_i, V_i)$ given θ_i and V_i, to assess sensitivity to departures from the basic model). 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 XKB we conside 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 $\widehat{\mu}$ mimimize 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 semiparametric SURE estimator of XKB with shrinkage towards the grand mean, defined by

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

where $\widehat{\boldsymbol{b}} = (\widehat{b}_1, ..., \widehat{b}_n)$ minimize an unbiased estimator of the risk (SURE) 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 $V_i \leq V_j \Rightarrow b_i \leq b_j$. The group-linear estimator $\widehat{\boldsymbol{\theta}}^{GL}$ of Definition 1.4.1 is applied here with the bins J_k formed by dividing the range of $\log(V_i)$ into $n^{1/3}$ equal length intervals, as per the discussion concluding section 1.4.

As benchmarks, in each example we also compute the two oracle risks

$$r(\mu^*,\gamma^*) = \min_{\mu,\gamma\in\mathbb{R}\ \ni\ \gamma\geq 0}\ \mathbb{E}\left\{\left[X - \frac{V}{\gamma+V}(X-\mu) - \theta\right]^2\right\} \tag{1.25}$$

and

$$r(a^*,b^*) = \min_{a(\cdot),b(\cdot)\ \ni\ a(v)\geq 0\ \forall v}\ \mathbb{E}\left\{\left[X - b(V)\big(X - a(V)\big) - \theta\right]^2\right\} \tag{1.26}$$

corresponding to the optimal rule in the parametric family of estimators considered in XKB, and to the optimal linear-in-x rule of section 1.2, respectively. (μ^* and γ^* are numbers whereas a^* and b^* are functions; the notation on the left hand sides of (1.25) and (1.26) should be understood here simply as the Bayes risk indexed by the appropriate quantities, and not as defined in (1.11)). In (1.25) and (1.26) the expected value is taken over (X,θ,V) distributed as (X_i,θ_i,V_i) in each example. Table 1 displays the oracle shrinkage location and shrinkage factors corresponding to (1.25) and (1.26): μ^* and $v/(v+\lambda^*)$ for the XKB family of estimators, and $a^*(v)$ and $b^*(v)$ for the family of estimators linear in X.

Figure 1 shows the average loss across the N 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 $r(a^*,b^*) \leq \inf\{r(a,b) : b(v)$ monotone increasing in $v\}$. For moderate n, whenever θ and V are independent, the SURE estimators are appropriate and achieve smaller risk, and when θ is furthermore normally distributed, the parametric SURE performs substantially better than the rest due to increased precision in estimating the shrinkage factor and shrinkage location. In contrast, the situations where θ and V are dependent are handled best by the group-linear estimator, and it indeed achieves significantly smaller risk than both SURE estimators.

Table 1: Oracle shrinkage locations and shrinkage factors, $(\mu^*, v/(v+\gamma^*))$ and $(a^*(v), b^*(v))$, corresponding to the family of estimators of XKB (equation (1.25)) and to the family of estimators that are linear in X (equation (1.26)). Table columns correspond to simulation examples (a)- (f). Values of μ^*, γ^* for each example are from Xie et al. (2012). Table shows value of γ^* in $v/(v+\gamma^*)$.

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

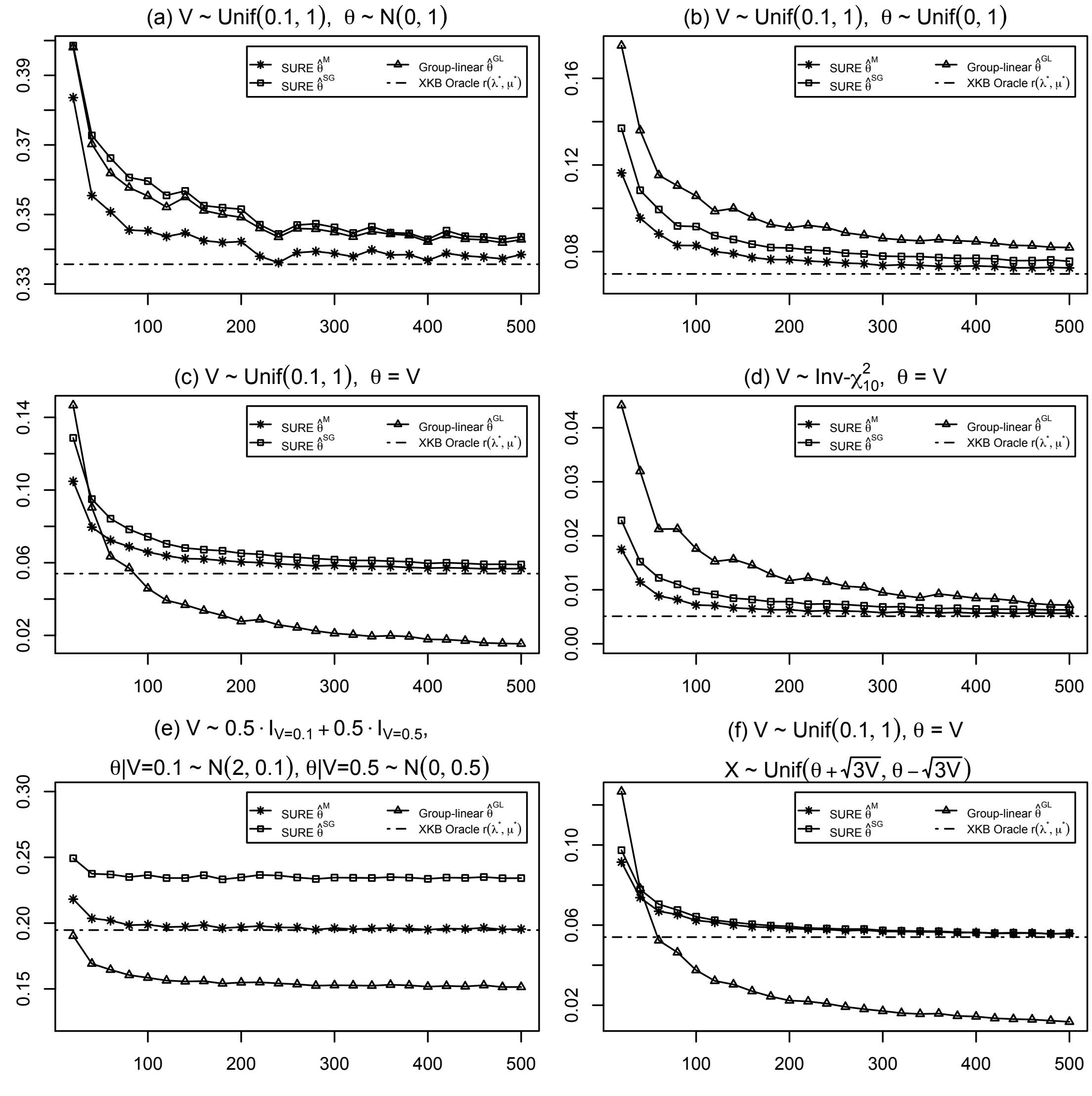


Figure 1: Estimated risk for various estimators vs. number of observations.

(a) Example 7.1 of XKB. In this example $V \sim \text{Unif}(0.1, 1)$ and $\theta \sim N(0, 1)$, independently, and $X \sim N(\theta, V)$. The parametric form of the estimators used in XKB is appropriate here as θ and V are independent and θ is normally distributed, and indeed the parametric SURE estimator performs best among the estimators considered. Still, the grouplinear estimator does at least as good as the semiparametric SURE across values of n, both estimators having some nonparametric aspect to them. As $n \to \infty$, the group-linear and the parametric SURE estimator have the same limiting risk $\approx .3357$. The asymptotic performance of the semiparametric SURE estimatoar is comparable.

(b) Example 7.2 of XKB. In this example $V \sim \text{Unif}(0.1, 1)$ and $\theta \sim \text{Unif}(0, 1)$, independently, and $X \sim N(\theta, V)$. Similarly to the pervious example, the parametric SURE estimator has substantially smaller risk than the group-linear estimator because of independence of θ and V. The semiparametric SURE estimator also performs better in this example. Nevertheless, the group-linear estimator again has the same asymptotic risk$\approx .0697$ as the the parametric SURE estimator, and the semiparametric SURE estimator performs comparably as n tends to infinity.

(c) Example 7.3 of XKB. This time $V \sim \text{Unif}(0.1, 1)$, $\theta = V$ and $X \sim N(\theta, V)$. θ and V are strongly dependent here, 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 estiator 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 θ and V.

(d) Example 7.4 of XKB. Here $V \sim \text{Inv-}\chi^2_{10}$, $\theta = V$ and $X \sim N(\theta, V)$. θ is still a deterministic function of V, but it takes larger values of n for the group-linear to outperform the SURE estimators; this is not seen before $n = 500$. This seems to be cuased because of the non-uniform distribution of the V_i, and is somewhat mitigated by considering $\log(V_i)$

when binning the observations, but not completely. For reference, when we used the (oracle knowledge of the) fact that $V \sim$ Inv-χ^2_{10} and applied the group-linear estimator to the transformed variables $F(V_i)$ where F is the distribution function of a Inv-χ^2_{10} random variable, the average loss approached the oracle risk 0 much faster in n. Still, the risk of the group-linear estimator approaches $r(a^*, b^*) = 0$ while the risk of the parametric SURE estimator approaches .0051.

(e) Example 7.5 of XKB. In this example, with probability 0.5 $V = 0.1$ and with probability 0.5 $V = 0.5$; $\theta|(V = 0.1) \sim N(2, 0.1)$ and $\theta|(V = 0.5) \sim N(0, 0.5)$; and $X \sim N(\theta, V)$. In this "two-groups" case, in each variance group, $\{i : V_i = 0.1\}$ and $\{i : V_i = 0.5\}$, 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 of XKB, which uses a SURE criterion to choose also the shrinkage location, achieves significantly smaller risk than the SURE estimators we considered here; still, its limiting risk is 0.1739, which is about 16% more than that of the group-linear estimator. The limiting risks of the parametric SURE estimator and of the group-linear estimator are $r(\mu^*, \gamma^*) = 0.1947$ and $r(a^*, b^*) = 0.15$, respectively.

(f) Example 7.6 of XKB. Lastly, $V \sim \text{Unif}(0.1, 1)$, $\theta = V$ and $X \sim \text{Unif}(\theta - \sqrt{3V}, \theta + \sqrt{3V})$, 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, with its risk approaching ≈ 0.0423.

1.6. Real Data Example

We now turn to a real data example to test our group-linear methods. We use the popular baseball data of 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 only. 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

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

As suggested in Brown (2008), a variance-stabilizing transformation is first applied,

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

resulting in

$$X_{ji} \dot{\sim} N(\theta_i, \frac{1}{4N_{ji}}), \qquad \theta_i = \arcsin(p_i)$$

and $\{(X_{1i}, N_{1i})\}$ are then used to estimate the means $\{\theta_i\}$. To measure the performance of an estimator $\widehat{\boldsymbol{\theta}}$, we use the Total Squared Error,

$$\text{TSE}(\widehat{\boldsymbol{\theta}}) = \sum_i \left[(X_{2i} - \widehat{\theta}_i)^2 - 1/(4N_{2i}) \right],$$

suggested by Brown (2008) as an 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.

Table 2 shows TSE for various estimators when applied (i) to all players, (ii) to pitchers only and (iii) to nonpitchers only. The values in the table are fractions of the TSE for the naive estimator, which, in each of the cases (i)-(iii), simply predicts X_{2i} by X_{1i}. In the table, the Grand mean estimator uses the simple average of all X_{1i}; the extended positive-part

Table 2: Prediction Errors of Batting Averages

	All	Pitchers	Nonpitchers
Naive	1	1	1
Grand mean	.852	.127	.378
James-Stein	.525	.164	.359
Nonparametric EB	.508	.212	.372
Binomial mixture	.588	.156	.314
Weighted Least Squares	1.074	.127	.468
Weighted nonparametric MLE	.306	.173	.326
Weighted Least Squares (AB)	.537	.087	.290
Weighted nonparametric MLE (AB)	.301	.141	.261
SURE $\widehat{\theta}^{M}$	.422	.123	.282
SURE $\widehat{\theta}^{SG}$	.409	.081	.261
Semi-parametric URE	0.414	0.045	.259
Group-linear $\widehat{\theta}^{GL}$	**.3017**	**.1784**	**.3246**

James-Stein estimator is given by

$$\widehat{\theta}_i^{JS+} = \hat{\mu}_{JS+} + \big(1 - \frac{p-3}{\sum_i (X_i - \hat{\mu}_{JS+})}\big)_+ (X_i - \hat{\mu}_{JS+}), \qquad \hat{\mu}_{JS+} = \frac{\sum_i X_i/A_i}{\sum_i A_i};$$

$\widehat{\theta}^{M}$ is the parametric empirical Bayes estimator of XKB using the SURE criterion to choose both the shrinkage and the location parameter; $\widehat{\theta}^{SG}$ is the semiparametric SURE estimator of XKB that shrinks towards the grand mean. The table also includes values for various estimators reported in Table 2 of XKB, who surveyed some of the best-performing parametric and nonparametric estimators that had been previously applied to this dataset: The nonparametric shrinkage methods of Brown and Greenshtein (2009), the weighted least squares and 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 et al. (2010). Finally, we also included the values for the semiparametric URE of Xie et al. (2015), applied directly to the binomial averages H_{ji}.

As in the simulations, the group-linear estimator is applied to the data using equal length bins on $\log(\frac{1}{4N_{1i}})$, partitioning the observations into eight groups.

The table shows that the group-linear estimator performs very well in predicting batting averages for all players relative to the other estimators. It has virtually the same prediction error as the nonparametric MLE method, which achieves the minimum error overall, although the two estimators are derived from very different perspectives. 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, pitchers tend to have better batting averages than non-pitchers, making it more plausible to believe that the θ_i come from a mixture of two normal distributions than from a single normal distribution. 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_i. 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, on the other hand, performs well in this example exactly because it is able to accommodate statistical dependence between the true means and the sampling variances.

Figure 2, a counterpart of Figure 2 in XKB, plots the coefficient of X_i (one minus the shrinkage factor) for the parametric SURE estimator $\widehat{\theta}^M$ and the group-linear estimator when each is applied to all batters; As opposed to the monotone decreasing shrinkage factor $V_i/(\widehat{\gamma} + V_i)$, $V_i = 1/(4N_i)$ of $\widehat{\theta}^M$, the shrinkage factors of group-linear estimator do not at all exhibit a monotone behavior as a function of N_i. The corresponding shrinkage location (not shown in figure) is constant for $\widehat{\theta}^M$, $\hat{\mu} = 0.45$, while it is piecewise constant and nondecreasing with N_i for the group-linear estimator: $\hat{\mu} = 0.42, 0.43, 0.43, 0.49, 0.52, 0.53, 0.54, 0.56$ corresponding to the eight consecutive segments of N_i in figure 2. Hence the estimates of the grouplinear estimator are in line with the behavior indicated by Brown (2008): "True" batting average seems to increase with number of at-bats (or decrease with V_i), and the variances are also not independent of N_i (otherwise, as long as the binomial model is appropriate, the shrinkage factors are expected to be decreasing across the segments of N_i).

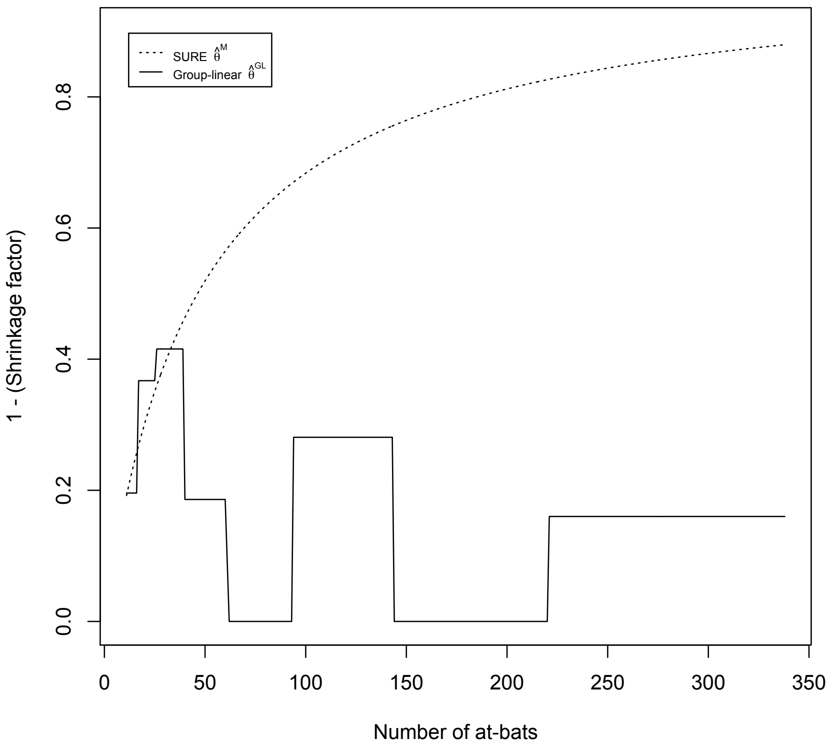


Figure 2: Shrinkage vs. number of at-bats. $\{1 - \text{shrinkage factor}\}$ increases with N_i according to $\widehat{\gamma}/(V_i + \widehat{\gamma}),\ V_i = 1/(4N_i)$ for the SURE estimator $\widehat{\theta}^M$; it is piecewise constant for the group-linear estimator, and exhibits no monotonicity. The corresponding shrinkage location is constant with N_i for $\widehat{\theta}^M$; for the group linear estimator $\{1 - \text{shrinkage factor}\}$ is constant on each segment of N_i, and nondecreasing.

Not surprisingly, the group linear estimator is not doing as well on the separate analyses for pitchers and nonpitchers. The parametric SURE estimator already has substantially smaller prediction error in both cases, and the semiparametric SURE estimator does even better. Intuitively, this again confirms that much of the heterogeneity in the data is accounted for by the type of player, pitcher or nonpitcher; it pays off to presume that independence holds between θ_i and N_i conditional on player type, when considering a linear versus a group-linear estimator.

1.7. Conclusion and Directions for Further Investigation

For a heteroscedastic mean, empirical Bayes estimators that have been suggested, both parametric and nonparametric, usually rely on a hierarchical model in which the parameter θ has a prior distribution unrelated to the observed sampling variance $V = \mathrm{Var}(X|\theta)$. 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, at least asymptotically. Group-linear methods, on the other hand, are capable of capturing dependency between θ and V, and therefore are more appropriate for problems where it exists.

There is certainly room for futher investigation and refinement of the results presented in this paper. We point out a few possible directions for extending Theorems 1.4.2 and 1.4.3, that are outside the scope of the current paper.

(i) When the distribution of the population (X, θ, V) is allowed to depend on n, the asymptotic optimality criterion (1.21) should be strengthened to the asymptotic ratio optimality criterion

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(\widehat{\theta}_i - \theta_i\Big)^2 \leq (1+o(1))r_n(a_n^*, b_n^*) \tag{1.27}$$

as $n \to \infty$. As (1.27) does not hold uniformly for all (X, θ, V), the aim is to prove this ratio optimality when $r_n(a^*, b^*) \geq \eta_n$ for small η_n under suitable side conditions on the joint distribution of (X, θ, V). This theory should include (1.21) as a special case and still maintain the property (1.20).

(ii) 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. For a strict improvement over the naive estimator, if the number of observations in a certain block is n, then the requirement on n may depend on the sequence $\{V_i\}$ in a more complicated way than the condition $n > 1 + 2V_{\max}/\overline{V}$ (i.e., $c_n^* > 0$) in Lemma 1.3.1.

CHAPTER 2 : Empirical Bayes Estimates for a Cross-Classified Additive Model with Unbalanced Design

Joint work with Lawrence D. Brown and Gourab Mukherjee

2.1. Introduction

The James-Stein estimator and its Bayesian interpretation revealed the usefulness of employing hierarchical models in estimation of a vector parameter with nonrandom components. As a tool to facilitate shrinkage, hierarchical models are appealing because they make evident the need for adjusting likelihood-based estimation, and possibly pooling of information from (conditionally) independent observations.

A great contribution to the understanding of the appropriateness of such models in some "fixed effects" situations was the work of Herbert Robbins, who drew an explicit connection between a frequentist compound decision problem - consisting of n symmetric copies to be solved simultaneously under some additive loss - and a one-dimensional Bayesian problem. Although not addressing strict minimaxity, ideas that appeared already in Robbins (1951) demonstrate the shortcomings of unbiased estimation, and are illuminating in the context of Stein's solution to the normal mean problem.

The point risk of any Bayes estimator resulting from a hierarchical structure posited for the data depends, of course, on the actual configuration of the true unknown parameters. Equivalently, for the homoscedastic normal means example, the Bayes estimator will be effective as compared to the usual estimator when the hierarchical structure does a good job in the second ("prior") level accommodating the empirical distribution of the unknown parameters. Empirical "linear Bayes" estimators, an example of which is the James-Stein estimator (see Efron and Morris, 1973b), attempt to mimic an optimal (oracle) linear estimator by considering a hierarchical model with a normal distribution at the (first and) second level, specified up to a set of hyperparameters to be estimated from the observed data and plugged back into the Bayes rule. Alternatively, a fully Bayes approach can be taken by considering the hyperparameters as random themselves with some vague prior distribution.

An advantage of the empirical Bayes approach is that in using the data to estimate the hyperparameter, one can appeal to frequentist considerations. In other words, use the Bayesian formalism to obtain a parametric family of estimators, but choose among this family relying on the likelihood function (and loss criterion) only. The entirely Bayes approach, by contrast, produces an estimator that is tied to the postulated model, hence its performance might deteriorate when the model does not reflect well the empirical distribution of the unknown parameters. In the extension of the normal mean problem to unequal variances, $y_i \sim N(\theta_i, \sigma_i^2),\ \sigma_i^2$ known, $i \leq n$ with sum-of-sqares loss, this motivated Xie et al. (2012) to suggest a parametric empirical Bayes estimator by minimizing an unbiased estimator of the (point) risk among Bayes rules with respect to an i.i.d. normal prior on $\theta = (\theta_1, ..., \theta_n)^\top$.

Technically, the ideas above carry over to the Gaussian linear model, $y = X\beta + \epsilon$ where $\epsilon \sim N_n(0, \Sigma)$ and $X_{n \times p}$ is a fixed and known matrix of covariates. Lindley and Smith (1972) were perhaps the first to employ a conjugate normal prior to obtain Bayes estimates for nonrandom β. They pursued a fully Bayes approach, considering in general a multilevel normal model conditional on 'dispersion' hyperparameters, which are themselves assigned a prior distribution meant to allow their estimation from the observed data. Extending the work of Xie et al. (2012) from the sequence model to the linear model with heteroscedastic error, Kou and Yang (2015) recently suggested to estimate the hyperparameters by minimizing an unbiased risk estimate (URE) among the parametric class of Bayes estimators of $E_\beta[y] = X\beta$ indexed by the hyperparameters. Under a set of sufficient conditions, they prove that the URE estimator is asymptotically optimal uniformly over β in terms of the point risk. In fact, their URE estimator achieves the performance of the optimal *loss* oracle within the class, which can base the choice of the hyperparameters also on y, and hence in particular achieves smaller risk than any hierarchical Bayes estimator as suggested in Lindley and Smith (1972).

The criterion of Kou and Yang (2015) leads to an asymptotically optimal estimator once the covariance structure of β has been specified up to a scaling hyperparameter: they assume $\text{Cov}(\beta) = \lambda W$ where W is a known positive definite matrix. However, "empirical" versions of the Bayes rule, which use data-dependent values for λ and the location of shrinkage, might be ineffective altogether when W is inadequate for representing the structure of the true parameters, as discussed before. Lindley and Smith (1972) promote the use of exchangeable structures in the covariance of β; But they emphasize throughout that exchangeability is not always reasonable. Indeed, for a typical linear regression problem, it is usually hard to justify exchangeability between the β_j, even after rescaling the columns of X. On the other hand, in the case of factorial experiments, exchangeability does make sense for the effects within each factor. Lindley and Smith (1972) considered a two-factor model with no interactions, where an exchangeable normal prior is used for the row effects and another exchangeable normal prior is used for the column effects. For the balanced design case, they proceed to derive the Bayes estimates under a hierarchical model with conjugate priors for the variance components corresponding to the row effects and column effects.

The two-factor additive model with *unbalanced* design is the focus of the current paper. We propose empirical Bayes estimates for the cell means under squared loss, as an alternative to the standard empirical Best Linear Unbiased Predictors (BLUP) in that we use the URE minimization criterion to "estimate" the hyperparameters. We emphasize that we are working in the "fixed effects" setting: The performance of an empirical Bayes (or any other) estimator is evaluated in terms of the point risk rather than the Bayes risk, which can explain why it might be desirable to estimate the hyperparameters differently than in a random effects model.

The complications that arise due to nonorthogonality in unbalanced factorial experiments are well documented in the literature. The evolution of the theory for the analysis of variance (ANOVA) in the unbalanced case is reviewed in Herr (1986) , who credits Yates's seminal paper (Yates, 1934) as the origin of the different methods (different sums-of-squares) used today. As for estimation, the difficulties presented in the classical approach are computational: the maximum likelihood estimates do not have a closed form as in the balanced case, and are much more complicated, but they have a familiar exact characterization (See, e.g., Searle, 2006). When shrinkage estimators are considered, however, the difficulties are not only computational. For balanced design, Bayes estimators that put separate i.i.d. priors on the row effects and on the column effects reduce, by sufficiency, to two separate one-way balanced problems, for which the standard empirical BLUP (James-Stein, if we want to emphasize that we are estimating nonrandom parameters) estimates are appropriate. Since this is not the case for unbalanced design, the empirical Bayes methods developed for the sequence model (i.e., where the mean of y is unrestricted) need to be extended.

Kou and Yang (2015) consider empirical Bayes estimation where the mean of y may be restricted to a given linear subspace, referred to in their paper as "Model II". While this includes the setup we consider in the current paper, their results do not really cover the (additive) factorial design because the asymptotics are carried out fixing the dimension of the linear subspace. The analysis in the current paper produces a counterpart of Xie et al. (2012) by letting the number of row and the number of column effects grow to infinity.

Another issue that is not addressed in Kou and Yang (2015) is the actual computation of SURE estimators in "Model II". Obtaining the actual SURE estimates in this case is in fact much more computationally intensive because it requires working with matrices throughout. We offer an implementation of the two-way SURE estimator which (for the case of no empty cells) is as efficient and fast as the computation of the standard empirical BLUP in the popular R package `lme4`. In conclusion, the additive cross-classified setup with unbalanced design merits a separate consideration.

The Chapter is organized as follows. In Section 2.2 we set up the model and present empirical Bayes estimators for the two-way unbalanced layout. Section 2.3 discusses the computation of the SURE estimator and provides the essential details. The balanced case is analyzed in Section 2.4. Section 2.5 includes the asymptotic optimality results for the SURE estimator, which are demonstrated in a simulation study in Section 2.6. The case of missing values is discussed in Section 2.7.

2.2. Model Setup and Bayes Estimates

Consider a two-way crosse-classified additive model,

$$y_{ij} = \eta_{ij} + \epsilon_{ij} \qquad \eta_{ij} = \mu + \alpha_i + \beta_j \qquad \epsilon_{ij} \sim N(0, \sigma^2 K_{ij}^{-1}) \qquad 1 \le i \le r,\ 1 \le j \le c \tag{2.1}$$

where $\sigma^2 > 0$ is known. Above, the nonrandom quantity α_i will be referred to as the i-th "row" effect, and the nonrandom quantity β_j as the j-th "column" effect; K_{ij} represents the number of observation, or the "count", in the ij cell. As notation suggests, there is no assumption that the K_{ij} are equal (if the K_{ij} are equal the design is said to be balanced). We do assume, for now, that $K_{ij} \ge 1$ for all i and j; The case of missing values is dealt with in section 2.7. In the overparametrized model (2.1), $\mu, \alpha_1, ..., \alpha_r, \beta_1, ..., \beta_c$ are not identifiable, however the cell means η_{ij} always are, and make the object of our inference. Specifically, the target is to estimate, based on $y = (y_{11}, y_{12}, ..., y_{rc})$, the vector $\eta = E(y) = (\eta_{11}, \eta_{12}, ..., \eta_{rc})^\top$ under the (normalized) sum-of-squares loss

$$L(\eta, \widehat{\eta}) = \frac{1}{rc}\|\widehat{\eta} - \eta\|^2 = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c}(\widehat{\eta}_{ij} - \eta_{ij})^2. \tag{2.2}$$

The risk of an estimator $\widehat{\eta}$ is then

$$R_{r,c}(\eta, \widehat{\eta}) = \frac{1}{rc}E\|\widehat{\eta} - \eta\|^2 = \frac{1}{rc}E\Big\{\sum_{i=1}^{r}\sum_{j=1}^{c}(\widehat{\eta}_{ij} - \eta_{ij})^2\Big\}.$$

The usual estimate of η is the weighted Least Squares (WLS) estimate (this is also maximum-likelihood under the Gaussian model (2.1)), which is unbiased and minimax.

Shrinkage estimators for the general linear model $y \sim N_n(X\theta, \sigma^2 I), X \in \mathbb{R}^{n\times p}$, of which (2.1) is a special case (if considering individual homoscedastic observations y_{ijk} instead of cell averages), have been suggested by extension of the James-Stein estimator (See, e.g., Rolph, 1976). Indeed, the general linear model can be reduced to the problem of estimating the mean of a *heteroscedastic* normal vector with known variances by applying orthogonal transformations to θ and y (See also Johnstone, 2011, Section 2.9). From there one can obtain Stein-type estimators as empirical Bayes rules, putting a prior which is either i.i.d. on the transformed coordinates or i.i.d. on the original coordinates of θ (Rolph, 1976, refers to these two choices as "proportional prior" and "constant prior", resprectively). In the case of factorial designs, however, neither of these choices is very sensible: A more reasonable choice of prior, as suggested by Lindley and Smith (1972), is one under which exhchageablility holds separately for the α_is and for the β_js. Hence, A linear shrinkage estimators for η_{ij} is obtained by adding to (2.1) a second level,

$$\alpha_i \sim N(0, \sigma_A^2), \qquad \beta_j \sim N(0, \sigma_B^2). \tag{2.3}$$

for some parameters σ_A^2, σ_B^2. In vector form, the Bayesian model under consideration is thus

$$y|\eta \sim N_p(\eta, \sigma^2 M), \qquad \eta = 1\mu + Z\theta, \qquad \theta \sim N_q(0, \sigma^2 \Lambda\Lambda^\top) \tag{2.4}$$

where

$$\theta = (\alpha_1, ..., \alpha_R, \beta_1, ..., \beta_C)^\top, \quad Z = [Z_a \ Z_b], \quad Z_a = I_R \otimes 1_C, \quad Z_b = 1_R \otimes I_C$$
$$M = \text{diag}(K_{11}^{-1}, K_{12}^{-1}, ..., K_{rc}^{-1}), \qquad \Lambda = \begin{bmatrix} \lambda_a I_R & 0 \\ 0 & \lambda_b I_C \end{bmatrix}$$

and where $p = rc$, $q = R + C$ and $\lambda_A = \sigma_A/\sigma$ and $\lambda_B = \sigma_B/\sigma$ are the square root of the relative variance components of the row and column effects, respectively. Denoting

$$V = Z\Lambda\Lambda^\top Z^\top + M, \tag{2.5}$$

we have from (2.4) that

$$y|\eta \sim N_{rc}(\eta, \sigma^2 M), \qquad y \sim N_{rc}(1\mu, \sigma^2 V), \tag{2.6}$$

which immediately gives the Bayes rule for η as

$$\widehat{\eta}^{\mu,\lambda_a,\lambda_b} = E_{\mu,\lambda_a,\lambda_b}(\eta|y) = y - MV^{-1}(y - 1\mu) \tag{2.7}$$

using Tweedie's formula (see, e.g. Johnstone, 2011, Section 2.3). Note that we suppressed in notation the dependency of Λ and V on λ_a and λ_b.

The Bayes estimator (2.7) was derived for arbitrary constants $\mu, \lambda_a, \lambda_b$. Instead of fixing the values of $\mu, \lambda_a, \lambda_b$ in advance, we may now return to the model (2.1) and consider the parametric family of estimators

$$\{\widehat{\eta}^{\mu,\lambda_a,\lambda_b} : \mu \in \mathbb{R},\ \lambda_a > 0,\ \lambda_b > 0\} \tag{2.8}$$

for the nonrandom vector η. An empirical Bayes approach uses the observed data y to select a candidate from the family to use as the estimate. In other words, an empirical Bayes estimator corresponding to the family (2.8) takes on the form

$$\widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b} = y - M\widehat{V}^{-1}(y - 1\widehat{\mu}), \qquad \widehat{V} = Z\widehat{\Lambda}\widehat{\Lambda}^\top Z^\top + M \tag{2.9}$$

where $\widehat{\Lambda} = \Lambda_{\widehat{\lambda}_a,\widehat{\lambda}_b}$ and where $\widehat{\mu}, \widehat{\lambda}_a \geq 0, \widehat{\lambda}_b \geq 0$ depend on y only.

Usual empirical Bayes estimators are derived relying on the random effects model (2.4). Hence, the fixed effect μ and the relative variance components λ_a^2 and λ_b^2 are treated as unknown fixed parameters to be estimated based on the marginal distribution of y and substituted into (2.7). For any set of estimates substituted for λ_a^2 and λ_b^2, the general mean μ is customarily estimated by generalized Least Squares, thereby producing an empirical version of what is known as the BLUP (best linear unbiased predictor). There is extensive literature on the estimation of the variance components, with the main methods being maximum-likelihood (ML), Restricted maximum-likelihood (REML) and the ANOVA methods (Method-of-Moments), including the three original ANOVA methods of Henderson. All these methods and some of their properties in the balanced and in the unbalanced random-effects model are discussed in detail in Chapters 6 and 5 of Searle et al. (2009). We concentrate on the commonly used maximum-likelihood estimates, which are implemented (as are the REML estimates) in the popular `R` package `lme4` (Bates et al., 2014). Thus, if $L(\mu, \lambda_a, \lambda_b; y)$ denotes the (marginal) likelihood of y according to (2.4), the maximum-likelihood estimates are

$$(\widehat{\mu}^{ML}, \widehat{\lambda}_a^{ML}, \widehat{\lambda}_b^{ML}) = \underset{\mu, \lambda_a \geq 0, \lambda_b \geq 0}{\arg\max} \{L(y; \mu, \lambda_a, \lambda_b)\}. \tag{2.10}$$

and the corresponding empirical Bayes estimator is obtained by plugging (2.10) into (2.7)

$$\widehat{\eta}^{ML} = \widehat{\eta}^{\widehat{\mu}^{ML}, \widehat{\lambda}_a^{ML}, \widehat{\lambda}_b^{ML}}.$$

Taking the partial derivatives of the log-likelihood

$$\log L(y; \mu, \lambda_a, \lambda_b) = -\frac{rc}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log |V| - \frac{1}{2\sigma^2} (y - 1\mu)^\top V^{-1} (y - 1\mu) \tag{2.11}$$

with respect to μ, λ_a^2 and λ_b^2 and observing that

$$V = \lambda_a^2 Z_a Z_a^\top + \lambda_b^2 Z_b Z_b^\top + M \tag{2.12}$$

yields, on equating to 0, that the maximum likelihood estimator for μ is

$$\widehat{\mu} = (1^\top \widehat{V}^{-1} y)/(1^\top \widehat{V}^{-1} 1) \tag{2.13}$$

and that if the maximum likelihood estimates $(\hat{\lambda_a^2}), (\hat{\lambda_b^2})$ are both strictly positive, they satisfy

$$\begin{aligned} \mathrm{tr}(\widehat{V}^{-1} Z_a Z_a^\top) - \frac{1}{\sigma^2} y^\top (I - \widehat{P})^\top \widehat{V}^{-1} Z_a Z_a^\top \widehat{V}^{-1} (I - \widehat{P}) y = 0 \\ \mathrm{tr}(\widehat{V}^{-1} Z_b Z_b^\top) - \frac{1}{\sigma^2} y^\top (I - \widehat{P})^\top \widehat{V}^{-1} Z_b Z_b^\top \widehat{V}^{-1} (I - \widehat{P}) y = 0 \end{aligned} \tag{2.14}$$

where

$$\widehat{P} = 1(1^\top \widehat{V}^{-1} 1)^{-1} 1^\top \widehat{V}^{-1} \tag{2.15}$$

and where $\widehat{V}$ is obtained from (2.12) by replacing λ_a^2, λ_b^2 with $(\hat{\lambda_a^2}), (\hat{\lambda_b^2})$. The derivation is standard, and we provide details in the Appendix. If the solution to the estimating equations (2.14) includes a negative component, it needs to be appropriately adjusted to produce the maximum-likelihood estimates $(\hat{\lambda_a^2}), (\hat{\lambda_b^2})$.

Designed for the random effects setup, the empirical Bayes estimators described so far will perform well, at least asymptotically, in terms of the Bayes (or prediction) risk $E_{\mu,\lambda_a,\lambda_b} \|\widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b} - \eta\|^2$ associated with the model (2.4). However, as we are interested in the risk function conditional on η - not the Bayes risk - of an estimator, the methods described above for choosing data-based substitutes $\widehat{\mu}, \widehat{\lambda}_a, \widehat{\lambda}_b$ are not necessarily adequate. Hence, taking the approach of Xie et al. (2012), we suggest to choose among the estimators in (2.8) by minimizing an unbiased estimator of the risk. Specifically, invoking a standard fomula (see, e.g., Berger, 1985, p. 362), we obtain Stein's unbiased estimator of the risk of $\widehat{\eta}^{\mu,\lambda_a,\lambda_b}$ as

$$SURE(y; \mu, \lambda_a, \lambda_b) = \frac{1}{rc}\{\sigma^2 \mathrm{tr}(M) - 2\sigma^2 \mathrm{tr}(V^{-1} M^2) + (y - 1\mu)^\top [V^{-1} M^2 V^{-1}](y - 1\mu)\} \tag{2.16}$$

and set

$$(\widehat{\mu}^S, \widehat{\lambda}_a^S, \widehat{\lambda}_b^S) = \mathop{\arg\min}_{\mu,\lambda_a \geq 0, \lambda_b \geq 0} \{SURE(y; \mu, \lambda_a, \lambda_b)\}. \tag{2.17}$$

The corresponding empirical Bayes estimator is obtained by plugging (2.10) into (2.7)

$$\widehat{\eta}^S = \widehat{\eta}^{\widehat{\mu}^S, \widehat{\lambda}_a^S, \widehat{\lambda}_b^S}.$$

As in the case of maximum likelihood estimation, there is no closed-form solution to (2.17), but we can characterize the solutions by the corresponding estimating equations. Taking the partial derivatives of $SURE(\mu, \lambda_a, \lambda_b; y)$ with respect to μ, λ_a^2 and λ_b^2 and using again the representation of V in (2.12), one finds on equating to 0 that the SURE estimate of μ is given by

$$\widehat{\mu} = (1^\top[\widehat{V}^{-1}M^2\widehat{V}^{-1}]y)/(1^\top[\widehat{V}^{-1}M^2\widehat{V}^{-1}]1) \tag{2.18}$$

and the SURE estimates $(\hat{\lambda_a^2}), (\hat{\lambda_b^2})$, if both are strictly positive, satisfy

$$\begin{aligned} &\mathrm{tr}(\widehat{V}^{-1}Z_aZ_a^\top\widehat{V}^{-1}M^2) - \frac{1}{\sigma^2}y^\top(I-\widehat{P})^\top\widehat{V}^{-1}Z_aZ_a^\top\widehat{V}^{-1}M^2\widehat{V}^{-1}(I-\widehat{P})y = 0 \\ &\mathrm{tr}(\widehat{V}^{-1}Z_bZ_b^\top\widehat{V}^{-1}M^2) - \frac{1}{\sigma^2}y^\top(I-\widehat{P})^\top\widehat{V}^{-1}Z_bZ_b^\top\widehat{V}^{-1}M^2\widehat{V}^{-1}(I-\widehat{P})y = 0 \end{aligned} \tag{2.19}$$

where

$$\widehat{P} = 1(1^\top[\widehat{V}^{-1}M^2\widehat{V}^{-1}]1)^{-1}1^\top\widehat{V}^{-1}M^2\widehat{V}^{-1} \tag{2.20}$$

and where $\widehat{V}$ is obtained from (2.12) by replacing λ_a^2, λ_b^2 with $(\hat{\lambda_a^2}), (\hat{\lambda_b^2})$. Details of the derivation are provided in the appendix.

Note that the two systems of equations (2.14) and (2.19) can be compared to study the difference between the estimates of the (relative) variance components produces by the two approaches. For this purpose it is perhaps easier to compare the following less explicit forms of the the first equation of (2.14) and the first equation of (2.19), without substituting the closed expression for μ,

$$\text{MLE:}\quad \mathrm{tr}(\widehat{V}^{-1}Z_aZ_a^\top) - \frac{1}{\sigma^2}(y-1\widehat{\mu})^\top\widehat{V}^{-1}Z_aZ_a^\top\widehat{V}^{-1}(y-1\widehat{\mu}) = 0 \tag{2.21}$$

$$\text{SURE}:\quad \mathrm{tr}(\widehat{V}^{-1}Z_aZ_a^\top\widehat{V}^{-1}M^2) - \frac{1}{\sigma^2}(y-1\widehat{\mu})^\top\widehat{V}^{-1}Z_aZ_a^\top\widehat{V}^{-1}M^2\widehat{V}^{-1}(y-1\widehat{\mu}) = 0. \tag{2.22}$$

Indeed, it can be seen that the SURE equation involves an extra term $\widehat{V}^{-1}M^2$ in both summands of the left-hand-side, as compared to the ML equation.

2.3. Computation of the SURE Estimator

To compute the SURE estimator, one could attempt to solve the system of equations (2.19), which involves only λ_a and λ_b as the unknowns (but has no closed-form solution). For example, one could fix the value of λ_a to some initial positive value and solve the first equation in λ_b; Then plug the solution into the second equation and solve for λ_a, and keep iterating between the two equations until convergence. If this approach is taken, a non-trivial issue to overcome will be obtaining the actual SURE estimates of λ_a and λ_b when one of the solutions to (2.19) is negative.

The main difficulty, in any case, is the occurrence of the $(rc) \times (rc)$ matrix $\widetilde{V}^{-1}$, which depends on λ_a and λ_b: Inverting this matrix can be a prohibitive task for even moderately large values of r anc c, and it needs to be inverted many times during the numerical computation.

For the case of no empty cells, we offer a fast and efficient computation that works as fast as the computation of the EBMLE estimate with the `lme4` R-package (Bates et al., 2014). Our implementation uses an adaptation of some of the key elements from the `lme4` package, which we learned from the excellent documentation in Bates (2010, Sec. 5.4). For the empty-cells case, the implementation is very similar after using the reduction to quadratic loss in η described in section 2.7. Unfortunately, up to this point we have not found a way to make the computation work as fast as it does in the case of no empty cells. Indeed, the presence of the p.s.d. matrix Q of section 2.7 in the expression for SURE imposes further difficulty to the methods described below. This is not to say that there is no way to overcome these difficulties, that had not occurred to us at this point. We highlight the main steps of our implementation for the no-empty-cells case below; More detail is given in the Appendix.

Using the matrix inverse identity, we show in the Appendix that (2.16) can be written as

$$SURE = -\sigma^2 \text{tr}(M) + 2\sigma^2 \text{tr}\{(\Lambda^\top Z^\top M^{-1} Z \Lambda + I_q)^{-1} (\Lambda^\top Z^\top Z \Lambda)\} + \|MV^{-1}(y - 1\mu)\|^2.$$

The expression above is numerically minimized jointly over (λ_a, λ_b), where the key step in evaluating it for a particular pair (λ_a, λ_b) is employing a sparse Cholesky decomposition for the matrix $\Lambda^\top Z^\top M^{-1} Z \Lambda + I_q$, as suggested in the documentation of the `lme4` package (for a slightly different matrix). This decomposition takes advantage of the high sparsity of $\Lambda^\top Z^\top M^{-1} Z \Lambda + I_q$; It first determines the locations of non-zero elements in the Cholesky factor, which do not depend on the values of (λ_a, λ_b) and hence this stage is needed only once during the numerical optimization. This is the costly stage of the decomposition; Determining the values of the non-zero components is repeated during the numerical optimization.

2.4. The Balanced Case

Sufficiency arguments suggest that in the balanced case, $K_{ij} \equiv K$, taking either of the two approaches, MLE or SURE, leads to solving two separate balanced one-way problems, and hence to similar estimates for $\mu, \lambda_a, \lambda_b$. We now turn to show that versions of the MLE and SURE estimates, which take into account centering of the row and column effects, indeed coincide when the design is balanced. Interestingly, the analysis will suggest another class of shrinkage estimators for the general, unbalanced, two-way problem by utilizing the one-way estimates of Xie et al. (2012).

Suppose, without loss of generality, that $K = 1$. We begin with a few definitions. The grand mean and the row and column main effects in terms of the parameters in the model (2.1) are

$$m = \mu + \alpha. + \beta.,\qquad a_i = \alpha_i - \alpha.,\qquad b_j = \beta_j - \beta.\ . \tag{2.23}$$

and the corresponding vectors are $\alpha = \{\alpha_i\},\ a = \{a_i\},\ \beta = \{\beta_i\},\ b = \{b_i\}$. Let also

$$\widehat{\eta}^{y..,\lambda_a,\lambda_b} := \widehat{\eta}^{\mu,\lambda_a,\lambda_b}\Big|_{\mu = y..} \tag{2.24}$$

be the (Bayes) estimator obtained by substituting the mean of y for μ in (2.7), and define similarly the estimates $\widehat{\alpha}^{y..,\lambda_a,\lambda_b}$ and $\widehat{\beta}^{y..,\lambda_a,\lambda_b}$, so that

$$\widehat{\eta}_{ij}^{y..,\lambda_a,\lambda_b} = y.. + \widehat{\alpha}_i^{y..,\lambda_a,\lambda_b} + \widehat{\beta}_j^{y..,\lambda_a,\lambda_b}.$$

Finally, denote by $\widehat{m}^{\mathrm{LS}}, \widehat{a}^{\mathrm{LS}}, \widehat{b}^{\mathrm{LS}}$ the weighted least squares estimates of m, a, b under (2.1). Then in the balanced case,

$$\widehat{m}^{\mathrm{LS}} = y..,\qquad \widehat{a}_i^{\mathrm{LS}} = y_{i}. - y..,\qquad \widehat{b}_i^{\mathrm{LS}} = y._j - y.. \tag{2.25}$$

and the Bayes estimates are

$$\widehat{\mu}^{y..,\lambda_a,\lambda_b} = \widehat{m}^{\mathrm{LS}},\qquad \widehat{\alpha}^{y..,\lambda_a,\lambda_b} = c_\alpha \widehat{a}^{\mathrm{LS}},\qquad \widehat{\beta}^{y..,\lambda_a,\lambda_b} = c_\beta \widehat{b}^{\mathrm{LS}} \tag{2.26}$$

where $c_\alpha = c_\alpha(\lambda_a)$ and $c_\beta = c_\beta(\lambda_b)$.

Therefore,

$$
\begin{aligned}
R(\eta, \widehat{\eta}^{y_{\cdot\cdot},\lambda_a,\lambda_b}) &= \frac{1}{rc} E\Big\{ \sum_{i=1}^{r} \sum_{j=1}^{c} (\widehat{\mu}^{y_{\cdot\cdot},\lambda_a,\lambda_b} + \widehat{\alpha}^{y_{\cdot\cdot},\lambda_a,\lambda_b} + \widehat{\beta}^{y_{\cdot\cdot},\lambda_a,\lambda_b} - m - a_i - b_j)^2 \Big\} && (2.27) \\
&= \frac{1}{rc} E\Big\{ \sum_{i=1}^{r} \sum_{j=1}^{c} (\widehat{m}^{\mathrm{LS}} + c_\alpha \widehat{a}_i^{\mathrm{LS}} + c_\beta \widehat{b}_j^{\mathrm{LS}} - m - a_i - b_j)^2 \Big\} && (2.28) \\
&= \frac{1}{rc} E\Big\{ \sum_{i=1}^{r} \sum_{j=1}^{c} [(\widehat{m}^{\mathrm{LS}} - m) + (c_\alpha \widehat{a}_i^{\mathrm{LS}} - a_i) + (c_\beta \widehat{b}_j^{\mathrm{LS}} - b_j)]^2 \Big\} && (2.29) \\
&= \frac{1}{rc} E\Big\{ rc(\widehat{m}^{\mathrm{LS}} - m)^2 + c \sum_{i=1}^{r} (c_\alpha \widehat{a}_i^{\mathrm{LS}} - a_i)^2 + r \sum_{j=1}^{c} (c_\beta \widehat{b}_j^{\mathrm{LS}} - b_j)^2 \Big\} && (2.30) \\
&= E\Big\{ (\widehat{m}^{\mathrm{LS}} - m)^2 \Big\} + \frac{1}{r} E\Big\{ \sum_{i=1}^{r} (c_\alpha \widehat{a}_i^{\mathrm{LS}} - a_i)^2 \Big\} + \frac{1}{c} E\Big\{ \sum_{j=1}^{c} (c_\beta \widehat{b}_j^{\mathrm{LS}} - b_j)^2 \Big\} \\
& && (2.31)
\end{aligned}
$$

where equality (2.30) is due to orthogonality of the vectors corresponding to the three sums-of-squares. Since, marginally,

$$
\widehat{m}^{\mathrm{LS}} \sim N(m, \sigma^2 \lambda_m^2), \qquad \widehat{a}^{\mathrm{LS}} \sim N_r(a, \sigma^2 \Lambda_a), \qquad \widehat{b}^{\mathrm{LS}} \sim N_c(b, \sigma^2 \Lambda_b), \tag{2.32}
$$

with known $\lambda_m^2, \Lambda_\alpha, \Lambda_\beta$ (and σ^2), SURE can be written as the sum of three separate SURE expressions, one for each of the summands in (2.30). Minimizing SURE for $\widehat{\eta}^{y_{\cdot\cdot},\lambda_a,\lambda_b}$ jointly over c_α, c_β therefore consists of minimizing separately the "row" term over c_α and the "column" term over c_β. Each of these is a "one-way" Gaussian homoscedastic problem, except that the covariance matrices $\Lambda_\alpha, \Lambda_\beta$ are singular (because main effects are centered). This will be taken into account in writing SURE for each, namely, the SURE estimate will have the "correct" degrees-of-freedom.

The maximum-likelihood estimates for the two-way random-effects, additive model do not have a closed-form solution even for balanced data (see Searle et al., 2009, Ch. 4.7 d.), which already rules them out. On the other hand, the REML estimates coincide with the positive-part Moments method estimates (Searle et al., 2009, Ch. 4.8), which, in turn, reduce (for known σ^2) to solving separately two one-way problems involving $\widehat{a}^{\text{LS}}$ for the rows and $\widehat{b}^{\text{LS}}$ for the columns. These have the usual closed-form solutions and are easily seen to coincide with the SURE solutions (and in particular, have the "correct" degrees-of-freedom). We conclude that, for balanced data, if SURE is written for the estimator that shrinks towards the overall mean [1] then minimizing SURE produces the same estimates for λ_a, λ_b as REML.

Note that that independence of $\widehat{m}^{\text{LS}}, \widehat{a}^{\text{LS}}, \widehat{b}^{\text{LS}}$ (in the balanced case) was not needed for any of (2.27)-(2.32). Specifically, (2.30) holds because of the side conditions satisfied by a, b and $\widehat{a}^{\text{LS}}, \widehat{b}^{\text{LS}}$; and (2.32) holds, with some known covariance matrices, in general for the GLS estimators. Hence the calculation goes through for unbalanced data as well (importantly, note that since y is the vector of cell *averages*, the design matrix is the same for balanced and unbalanced data), where $\widehat{m}^{\text{LS}}, \widehat{a}^{\text{LS}}, \widehat{b}^{\text{LS}}$ still denote the GLS estimates. In the unbalanced case, however, (2.26) no longer holds, i.e., the *Bayes* estimates for α (β) no longer depend on $\widehat{a}^{\text{LS}}$ ($\widehat{b}^{\text{LS}}$) alone. Additionally, Λ_a adn Λ_b in (2.32) do not have a constant on their diagonals, that is, the GLS estimators are heteroscedastic (and have correlated components) for unbalanced data. Giving up Bayes optimality, we can nevertheless concentrate on shrinkage estimates of the form (2.26) and look for "optimal" constants $c_\alpha = c_\alpha(\lambda_a)$ and $c_\beta = c_\beta(\lambda_b)$ in terms of the risk. By (2.31), the solution for c_α must be optimal for the one-way problem involving only $\widehat{a}^{\text{LS}}$ (and similarly for c_β with $\widehat{b}^{\text{LS}}$), and is asymptotically attained by the SURE estimate of Xie et al. (2012). Hence we define

$$\widehat{\eta}_{ij}^{\text{XKB}} = \widehat{m}^{\text{LS}} + \widehat{c}_\alpha^{\text{XKB}} \widehat{a}_i^{\text{LS}} + \widehat{c}_\beta^{\text{XKB}} \widehat{b}_j^{\text{LS}}, \qquad 1 \le i \le r,\ 1 \le j \le c \tag{2.33}$$

with

$$\widehat{c}_\alpha^{\text{XKB}} = \arg\min_{c_\alpha} \text{SURE}\Big\{ \sum_{i=1}^{r} (c_\alpha \widehat{a}_i^{\text{LS}} - a_i)^2 \Big\}, \quad \widehat{c}_\beta^{\text{XKB}} = \arg\min_{c_\beta} \text{SURE}\Big\{ \sum_{j=1}^{c} (c_\beta \widehat{b}_j^{\text{LS}} - b_j)^2 \Big\} \tag{2.34}$$

[1] note the difference $\text{SURE}(\{\widehat{\eta}^{\mu,\lambda_a,\lambda_b}\}\big|_{\mu=y_{\cdot\cdot}}) \neq \{\text{SURE}(\widehat{\eta}^{\mu,\lambda_a,\lambda_b})\}\big|_{\mu=y_{\cdot\cdot}}$

Slight modification of the parametric SURE estimate of Xie et al. (2012) that shrinks towards 0, will be required to accommodate the covariance structure of the centered random vectors $\widehat{a}^{\mathrm{LS}}, \widehat{b}^{\mathrm{LS}}$. As shown in Xie et al. (2012), the estimates of c_α and c_β produced by maximum-likelihood empirical Bayes (EBMLE) and Moments-method empirical Bayes (EBMOM) are generally different for heteroscedastic observations, and therefore do not admit the same asymptotic properties.

2.5. Risk Properties of the SURE Estimator

In this section we provide some theoretical results that establish asymptotic optimality of the SURE estimator within the class (2.9) of empirical Bayes estimators. Attention is restricted here to the "all-cells-filled" situation, $K_{ij} \geq 1 \; \forall \; i \leq r$ and $j \leq c$.

For technical reasons, the optimality results in the current section regard the family of estimators

$$\{\widehat{\eta}^{\mu,\lambda_a,\lambda_b} : |\mu| \leq B, \; \lambda_a > 0, \; \lambda_b > 0\} \tag{2.35}$$

which differs from (2.8) in that the absolute value of μ is restricted to be bounded by some positive constant B. From a practical viewpoint, if the role of $\widehat{\mu}$ is, loosely speaking, to capture the overall mean, then the restriction above does not seem very limiting because it is reasonable to assume that the the overall mean is not really affected by the growing dimensions r and c. Therefore, while the empirical Bayes estimators presented in section 2.2 do not impose that restriction on μ, the concern should not be too serious about the extent to which the following results are practically applicable to the unrestricted family of EB estimators.

Theorem 2.5.1. *Under the following conditions:*

I. $\displaystyle \lim_{r,c\to\infty} \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} \eta_{ij}^2 < \infty$

II. $\displaystyle \lim_{r,c\to\infty} \frac{1}{rc} \left\{ \frac{\max\{K_{ij} : i \leq r, \; j \leq c\}}{\min\{K_{ij} : i \leq r, \; j \leq c\}} \right\} = 0$

it holds that:

(a) $$\sup_{|\mu|\le B;\ \lambda_1,\lambda_2\ge 0} E_{r,c}\Big[SURE(y;\mu,\lambda_a,\lambda_b)-R_{r,c}(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})\Big]^2 \to 0 \qquad \textit{as } r,\ c\to\infty.$$

(b) $$\sup_{|\mu|\le B;\ \lambda_1,\lambda_2\ge 0} E_{r,c}\Big[L(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})-R_{r,c}(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})\Big]^2 \to 0 \qquad \textit{as } r,\ c\to\infty.$$

As an immediate consequence of Theorem 2.5.1, we have

Corollary 2.5.2. *Under the conditions of Theorem 2.5.1, it holds that*

$$\sup_{|\mu|\le B;\ \lambda_1,\lambda_2\ge 0} E_{r,c}\Big[SURE(y;\mu,\lambda_a,\lambda_b)-L(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})\Big]^2 \to 0 \qquad \textit{as } r,\ c\to\infty.$$

The unbiased risk estimator and the loss have the same expected value for any η; Corollary 2.5.2 asserts that the these random variables are also close to each other in L^1. Note that the supremum is taken outside the expectation.

As a benchmark for the performance ever achievable by an estimator in the family (2.35) we consider a loss-oracle, which uses the knowledge of the true value of η to choose the values of $\mu, \lambda_a, \lambda_b$ for any realization of y. Hence, let

$$\left(\ \widetilde{\mu}^{OL},\widetilde{\lambda}_a^{OL},\widetilde{\lambda}_b^{OL}\ \right)=\operatorname*{arg\,min}_{|\mu|\le B;\ \lambda_a,\ \lambda_b\ge 0} L(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})=\operatorname*{arg\,min}_{|\mu|\le B;\ \lambda_a,\ \lambda_b\ge 0}\big\|y-MV^{-1}(y-1\mu)-\eta\big\|^2$$

resulting in the rule

$$\widetilde{\eta}^{OL}=\widehat{\eta}^{\widetilde{\mu}^{OL},\widetilde{\lambda}_a^{OL},\widetilde{\lambda}_b^{OL}}. \tag{2.36}$$

Corollary 2.5.2 provides the basis for the following asymptotic optimality results.

Theorem 2.5.3. *Under the conditions of Theorem 2.5.1, it holds that for any $\epsilon > 0$*

$$\lim_{\substack{r\to\infty\\ c\to\infty}} P_{r,c}\big\{L(\eta,\widehat{\eta}^S)\ge L(\eta,\widetilde{\eta}^{OL})+\epsilon\big\}=0 \qquad \textit{as } r,\ c\to\infty.$$

Theorem 2.5.4. *Under the conditions of Theorem 2.5.1, it holds that*

$$\lim_{\substack{r\to\infty \\ c\to\infty}} \left\{R_{r,c}(\eta, \widehat{\eta}^{S}) - E_{r,c}[L(\eta, \widetilde{\eta}^{OL})]\right\} = 0 \qquad \text{as } r,\ c \to \infty.$$

As the loss-oracle performs better than any empirical Bayes estimator of the form considered, a consequence of Theorems (2.5.3) and (2.5.4) is

Corollary 2.5.5. *Under the conditions of Theorem 2.5.1, it holds that for any estimator* $\widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b}$ *of the form* (2.9),

(a) $$\lim_{\substack{r\to\infty \\ c\to\infty}} P_{r,c}\left\{L(\eta, \widehat{\eta}^{S}) \geq L(\eta, \widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b}) + \epsilon\right\} = 0 \qquad \text{as } r,\ c \to \infty.$$

(b) $$\limsup_{r\to\infty,\ c\to\infty} \left\{R_{r,c}(\eta, \widehat{\eta}^{S}) - R_{r,c}(\eta, \widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b})\right\} \leq 0 \qquad \text{as } r,\ c \to \infty.$$

2.6. Simulation Study

We now turn to a simulation study to compare the performance of the SURE estimator to that of different cell-means estimators discussed in the previous sections. As the standard technique we consider the weighted Least-Squares estimator $\widehat{\eta}^{LS} = 1\widehat{\mu}^{LS} + Z\widehat{\theta}^{LS}$ where $(\widehat{\mu}^{LS}, \widehat{\theta}^{LS})$ is any pair that minimizes

$$(y - 1\mu - Z\theta)^{\top} M^{-1} (y - 1\mu - Z\theta).$$

The shrinkage estimators reported are the maximum-likelihood empirical Bayes (EBMLE) estimator $\widehat{\eta}^{ML}$, characterized by equations (A.13)-(2.14), and the SURE empirical Bayes estimator $\widehat{\eta}^{S}$, characterized by equations (A.19)-(2.19), as well as versions of these two estimators that shrinks towards a fixed (prespecified) loacation $\mu = 0$. In addition, we consider the two-way empirical Bayes estimator $\widehat{\eta}^{XKB}$ derived in section 2.4 by reduction to a one-way problem and applying the SURE estimator of Xie et al. (2012) which shrinks towards a general data-driven location. For a benchmark we consider the oracle rule $\widetilde{\eta}^{OL}$

obtained by plugging into the parametric estimator (2.7) values

$$\left(\widetilde{\mu}, \widetilde{\lambda}_a, \widetilde{\lambda}_b \right) = \underset{\mu, \lambda_a \geq 0, \lambda_b \geq 0}{\arg\min} \left\| y - MV^{-1}(y - 1\mu) - \eta \right\|^2 \tag{2.37}$$

where in the right hand side $V = Z\Lambda\Lambda^\top Z^\top + M$ and $\Lambda = \Lambda_{\lambda_a, \lambda_b}$. Since for any y the oracle rule minimizes the loss over all members of the parametric family (2.8), its expected loss lower bounds the risk achievable by any empirical Bayes estimator of the form (2.9).

Referring to the likelihood model (2.1) and denoting $\alpha = (\alpha_1, ..., \alpha_r)^\top$, $\beta = (\beta_1, ..., \beta_c)^\top, M^{-1} = \text{diag}(K_{11}, K_{12}, ..., K_{rc})$, in each simulation example (a)-(d) we draw (α, β, M^{-1}) jointly from some distribution such that the cell counts K_{ij} are i.i.d. and (α, β) are drawn from some conditional distribution given the K_{ij}s. We then draw $y_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2 K_{ij})$ independently, fixing $\mu = 0$ throughout and setting σ^2 to some (known) constant value. This process is repeated for $N = 100$ time for each pair (r, c) in a range of values, and the average squared loss over the N rounds is computed for each of the estimators mentioned above. The SURE estimate is computed using the implementation described in [], and the oracle "estimate" is computed employing a similar technique. The EBMLE estimate is computed using the `R` package `lme4` (Bates et al., 2014).

We remark that the relatively small number of repetitions, $N = 100$, is used due to the computational effort in obtaining the empirical Bayes estimates above that are of the form (2.9). Nevertheless, $N = 100$ is large enough for the standard error of the average loss (for each estimator) to be at least one order-of-magnitude smaller than the estimated differences between the risks, hence the differences can be safely considered significant.

Table 3 shows the estimated risk for $L = 180$ as a fraction of the estimated risk of the Least-Squares estimator. In all examples except from the first (the only case in which the effects and the cell counts are drawn from the "correct" model (2.4)), the SURE estimator attains significantly smaller risk than that of the EBMLE, and comes close to the performance of the loss-oracle. Perhaps surprisingly, the "one-way" XKB estimator seems to have an asymptotically smaller risk than that of the EBMLE in all but the first example. As the table suggests, in extreme cases of dependency between the effects and the cell counts, even the Least-Squares estimator is preferable to the EBMLE (even asymptotically).

	(a)	(b)	(c)	(d)	(e)	(f)
LS	1.00	1.00	1.00	1.00	1.00	1.00
EBML	0.31	1.79	0.48	1.37	0.21	0.96
SURE	**0.31**	**0.45**	**0.19**	**0.21**	**0.18**	**0.58**
EBML (fixed μ)	0.31	0.69	0.45	1.42	0.58	0.95
SURE (fixed μ)	0.31	0.46	0.20	0.53	0.57	0.63
XKB	0.31	0.58	0.28	0.44	0.20	-
OL	0.30	0.42	0.16	0.20	0.17	0.56

Table 3: Estimated risks relative to the Least-Squares estimator, $L = 180$. The columns in the table correspond to the 6 simulation examples described in section 2.6.

(a) For $L \in \{20, 60, ..., 180\}$ we set $r = c = L$ and $\sigma^2 = 25$. K_{ij} are independent such that $P(K_{ij} = 1) = 0.9$ and $P(K_{ij} = 9) = 0.1$. For $1 \leq i, j \leq L$, α_i, β_j are drawn from a $N(0, \sigma^2/(4L))$ distribution independently of the K_{ij}s. The joint distribution of the row effects, column effects and the K_{ij}s in this example obeys the Bayesian model under which the parametric estimator (2.7) is derived. Hence the true Bayes rule is of that form, and the EBMLE is expected to perform well estimating the hyperparameters from the marginal distribution of y according to (2.6). Indeed, the risk curve of the EBMLE approaches that of the oracle rule and seems to perform best for relatively small value of L. The risk of the SURE estimator, however, still converges to the oracle risk as l increases. Interestingly, the performance of the XKB estimator seems to be comparable to that of SURE and EBMLE for large values of L.

(b) For $L \in \{20, 60, ..., 180\}$ we set $r = c = L$ and $\sigma^2 = 25$. In this example the K_{ij} are no longer independent of the random effects. We take $K_{ij} = 1 \cdot (1 - Z_i) + 25 \cdot Z_i$ where $Z_i \sim Bin(1, 0.5)$ independently, so that the cell frequencies are constant in each row. If $Z_i = 1$, α_i is drawn from a $N(1, \sigma^2/(100 \cdot 2L))$ distribution, and otherwise from a $N(0, \sigma^2/(2 \cdot L))$ distribution. β_j are drawn independently from a $N(0, \sigma^2/(2L))$ distribution. The advantage of the SURE estimator over the EBMLE is clear in figure 3; In fact, even the

Least-Squares estimator seems to do better than the EBMLE for the values of L considered here, a consequence of the strong dependency between the cell frequencies and the random effects. Again the XKB estimator performs surprisingly well.

(c) This example is the same as example (b), except that we fix $c = 40$ throughout and study the performance of the different estimators as number of row levels $r = L \in \{20, 60, ..., 180\}$ varies. We remark that this situation is not covered in the theoretical results of section 2.5. The Least-Squares estimator performs much worse, relatively to the other methods, than in the previous examples. The risk of the SURE estimator still seems to get closer to that of the oracle as $r = L$ increases, although we have not studied the behavior of the SURE estimator when c is fixed and $r \to \infty$ (neither theoretically nor numerically). The risk of the XKB is significantly higher than that of the SURE estimator for large values of $r = L$, but still much lower than that of the EBMLE.

(d) For $L \in \{20, 60, ..., 180\}$ we set $r = c = L$ and $\sigma^2 = 25$. In this example the row effects are *determined* by the K_{ij}. We take $K_{ij} = 1 \cdot (1 - Z_i) + 25 \cdot Z_i$ where $Z_i \sim Bin(1, 0.5)$ independently, and set $\alpha_i = 1 \cdot (1 - Z_i) + (1/25) \cdot Z_i$. β_j are drawn independently from a $N(0, \sigma^2/(2L))$ distribution. The SURE estimator performs significantly better than the other estimators for large values of l, with about 50% smaller estimated risk for $L = 180$ than that of the XKB estimator, and even much better for the other methods. The LS estimator again attained smaller estimated risk for the largest two values of L than EBMLE.

(e) For $L \in \{20, 60, ..., 180\}$ we set $r = c = L$ and $\sigma^2 = 25$. In this example both the row and the column effects are determined by the K_{ij}. The cell frequencies $K_{lj} = \max(T_l, 1),\ 1 \leq l \leq L, 1 \leq j \leq L$, where $T_l,\ 1 \leq l \leq L$, are drawn independently from a mixture of a $Pois(1)$ and $Pois(5)$ distributions with weights 0.9 and 0.1, respectively. The row and column effects are $\alpha_l, \beta_l = 1/T_l,\ 1 \leq l \leq L$. The estimated risk for the SURE estimator is still smaller than that of EBMLE by 14.7% ($\hat{sd}(\text{diff}) < 4 \cdot 10^{-5}$) for $L = 200$, but difference is not as big as in earlier examples. The estimated risk for the XKB estimator is smaller by just % than that of EBMLE. The Least Squares estimator performs considerably worse than the rest.

(f) In the last example we study the performance of the estimators when some cells are empty. Details of the calculation of the SURE estimator in the case of empty cells are described in section 2.7. The setting is exactly as in example (b), except that after the K_{ij} are drawn, each K_{ij} is independently set to 0 (corresponding to an empty cell) with probability 0.2. In accordance with the theory of sectionr̃efsec:missing, the risk of the SURE estimator approaches the expected oracle loss, and for $L = 180$ achieves significantly smaller risk than that of the EBMLE, although not as significantly smaller as in example (b) with all cells filled (40% vs 75% smaller than EBMLE for examples (f) and (b), respectively). The performance of the Least Squares estimator is comparable to that of the EBMLE. The XKB estimator is not considered here as it is not applicable when some data are missing (the argument made in section 2.4 is invalid since the columns of the design matrix for the cell averages $y_{ij\cdot}$ are not orthogonal when some cells are empty).

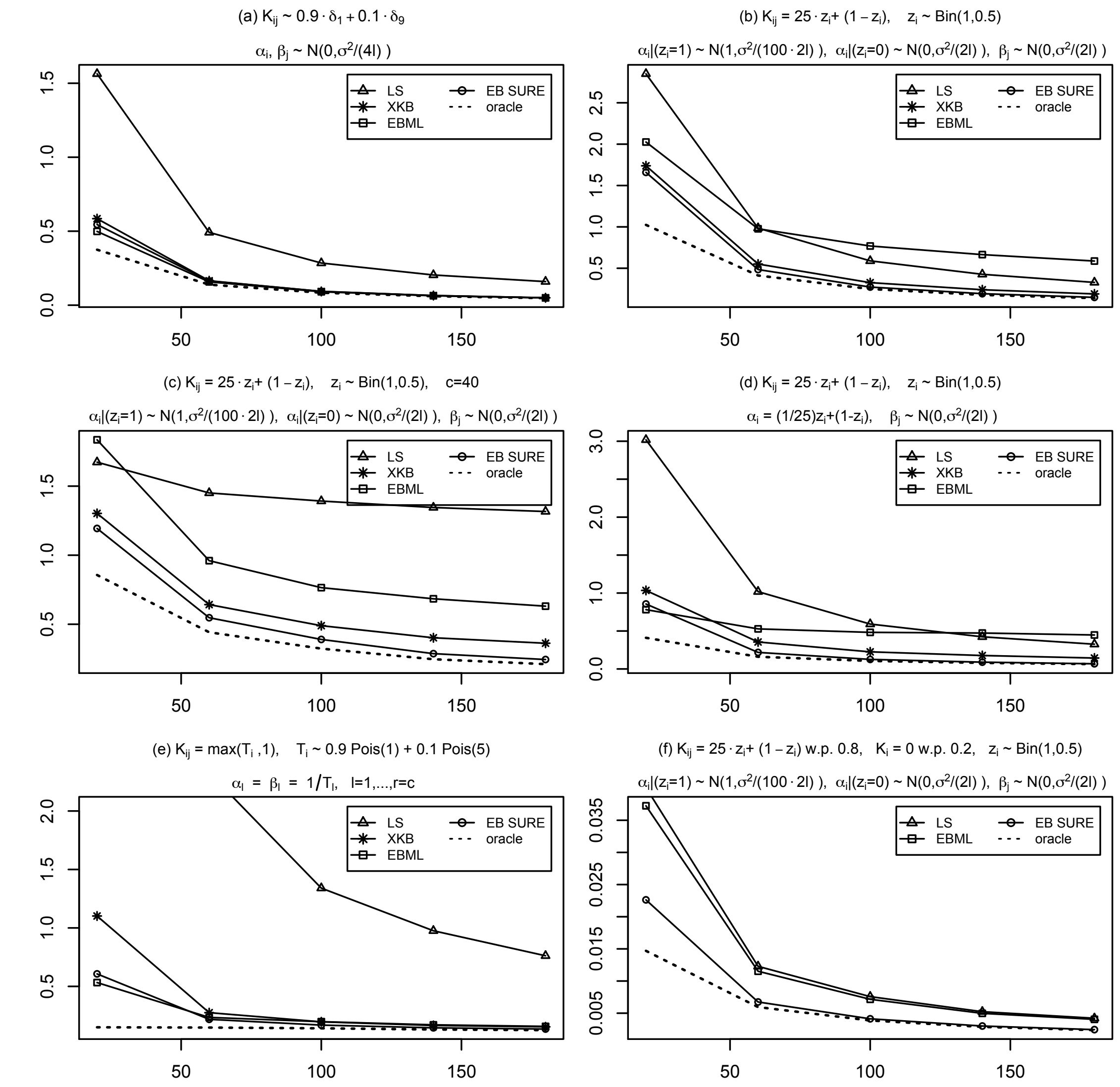
2.7. The Case of Missing Values

An important special case of unbalanced data is that of missing values, i.e., the occurrence of some empty cells in the two-way table. In this section we extend the SURE estimator of 2.2 to the case where some cells may be empty and the target of inference is all *estimable* cell means.

Model (2.1) can be extended to accommodate missing values by simply restricting the index set for the pairs (i, j), namely

$$y_{ij} = \eta_{ij} + \epsilon_{ij} \qquad \eta_{ij} = \mu + \alpha_i + \beta_j \qquad \epsilon_{ij} \sim N(0, \sigma^2 K_{ij}^{-1}) \qquad (i, j) \in S \qquad (2.38)$$

where $S = \{(i, j) : K_{ij} \geq 1\} \subseteq \{1, ..., r\} \times \{1, ..., c\}$ is the set of indices corresponding to the nonempty cells, and where σ^2 is known. Since the no-interaction model is considered, the means of some empty cells can still be estimable functions of the α_i and β_j in (2.38), hence



(a) $K_{ij} \sim 0.9 \cdot \delta_1 + 0.1 \cdot \delta_9$

$\alpha_i, \beta_j \sim N(0,\sigma^2/(4l))$

(b) $K_{ij} = 25 \cdot z_i + (1 - z_i), \quad z_i \sim Bin(1,0.5)$

$\alpha_i|(z_i=1) \sim N(1,\sigma^2/(100 \cdot 2l)), \ \alpha_i|(z_i=0) \sim N(0,\sigma^2/(2l)), \ \beta_j \sim N(0,\sigma^2/(2l))$

(c) $K_{ij} = 25 \cdot z_i + (1 - z_i), \quad z_i \sim Bin(1,0.5), \quad c=40$

$\alpha_i|(z_i=1) \sim N(1,\sigma^2/(100 \cdot 2l)), \ \alpha_i|(z_i=0) \sim N(0,\sigma^2/(2l)), \ \beta_j \sim N(0,\sigma^2/(2l))$

(d) $K_{ij} = 25 \cdot z_i + (1 - z_i), \quad z_i \sim Bin(1,0.5)$

$\alpha_i = (1/25)z_i + (1-z_i), \quad \beta_j \sim N(0,\sigma^2/(2l))$

(e) $K_{ij} = \max(T_i, 1), \quad T_i \sim 0.9\ Pois(1) + 0.1\ Pois(5)$

$\alpha_l = \beta_l = 1/T_l, \quad l=1,\ldots,r=c$

(f) $K_{ij} = 25 \cdot z_i + (1 - z_i)$ w.p. 0.8, $K_i = 0$ w.p. 0.2, $z_i \sim Bin(1,0.5)$

$\alpha_i|(z_i=1) \sim N(1,\sigma^2/(100 \cdot 2l)), \ \alpha_i|(z_i=0) \sim N(0,\sigma^2/(2l)), \ \beta_j \sim N(0,\sigma^2/(2l))$

Figure 3: Estimated risk for various estimators vs. number of row/column levels $r = c \triangleq L$. Example (c) is the only exception, with $c = 40$ fixed and l representing the number of row levels r. The versions of EBMLE and SURE estimators that shrink towards a fixed location μ are not plotted.

we consider the problem of estimating $\{\eta_{ij} : \eta_{ij}$ is estimable$\}$ rather than only the means of observed cells, $\{\eta_{ij} : (i, j) \in S\}$. To simplify matters, we will assume in the remainder of this section that S is such that all rc cell means are estimable (commonly referred to as the case of "connected" data), in which case the loss is still given by (2.2).

To work with vector forms, we introduce new notation to distinguish between the observed model and "unobserved" model (in which all cell are filled). Define $\theta = (\mu, \alpha_1, ..., \alpha_R, \beta_1, ..., \beta_C)^\top$ and note the inclusion of μ as the first element. Define $M = \mathrm{diag}(K_{ij}^{-1} : (i,j) \in S)$ where the indices of diagonal elements are in lexicographical order. Let $\widetilde{Z} = [1_{rc} \;\; I_R \otimes 1_C \;\; 1_R \otimes I_C] \in \mathbb{R}^{(rc)\times(r+c+1)}$ and let $Z \in \mathbb{R}^{|S|\times(r+c+1)}$ be obtained from $\widetilde{Z}$ by deleting the subset of rows corresponding to S^c. Finally, let $\widetilde{\eta} = \widetilde{Z}\theta \in \mathbb{R}^{rc}$ be the vector of all cell means and $\eta = Z\theta \in \mathbb{R}^{|S|}$ be the vector of cell means for only the observed data.

Since $\widetilde{\eta}$ is assumed to be estimable, it must be a linear function of η. Specifically, it must hold that

$$\widetilde{\eta} = \widetilde{Z}Z^\dagger \eta$$

where $Z^\dagger$ is the Moore-Penrose pseudo-inverse of Z. Any other generalized inverse could be used in $(Z^\top Z)^- Z^\top$ to replace $Z^\dagger$; The Moore-Penrose inverse is just a convenient choice. For a proof that the relation above holds, the reader is referred to Theorem 5 in Searle (1966). Now consider an estimate of $\widetilde{\eta}$ of the form $\hat{\widetilde{\eta}} = \widetilde{Z}Z^\dagger \widehat{\eta}$. Then (2.2) can be written as

$$L(\widetilde{\eta}, \hat{\widetilde{\eta}}) = \frac{1}{rc}\|(\hat{\widetilde{\eta}} - \widetilde{\eta})\|^2 = \frac{1}{rc}(\hat{\eta} - \eta)^\top Q(\hat{\eta} - \eta), \qquad Q = (\widetilde{Z}Z^\dagger)^\top \widetilde{Z}Z^\dagger.$$

Namely, for estimators of the form $\hat{\widetilde{\eta}} = \widetilde{Z}Z^\dagger \widehat{\eta}$, the problem is equivalent to estimating η under the quadratic loss above. Note that with $\widehat{\eta} = \widehat{\eta}^{LS}$ this form gives the (Generalized) Least Squares estimate of $\widetilde{\eta}$; And for a Bayes rule $\widehat{\eta} = \widehat{\eta}^B$ with respect to any prior on θ, this form gives the Bayes rule for $\widetilde{\eta}$ with respect to the same prior.

Now that the loss is given in terms of $\widehat{\eta}$ and η, we turn to the Bayes model under the prior in (2.3). The Bayes estimate is still given by (2.7) as

$$\widehat{\eta}^{\mu,\lambda_a,\lambda_b} = y - MV^{-1}(y - 1\mu)$$

where η, M and Z in (2.5) are as defined in the current section. Under (2.38) and the quadratic loss above, an unbiased estimator of the risk of $\widehat{\eta}^{\mu,\lambda_a,\lambda_b}$ is

$$\begin{aligned} SURE^Q(y;\mu,\lambda_a,\lambda_b) = \sigma^2\mathrm{tr}(QM) - 2\sigma^2\mathrm{tr}(V^{-1}MQM) \\ + (y-1\mu)^\top[V^{-1}MQMV^{-1}](y-1\mu). \end{aligned} \tag{2.39}$$

The corresponding SURE estimator of η is

$$\widehat{\eta}^{S_Q} = \widehat{\eta}^{\widehat{\mu}^{S_Q},\widehat{\lambda}_a^{S_Q},\widehat{\lambda}_b^{S_Q}}$$

where

$$(\widehat{\mu}^{S_Q},\widehat{\lambda}_a^{S_Q},\widehat{\lambda}_b^{S_Q}) = \underset{\mu,\lambda_a\geq 0,\lambda_b\geq 0}{\arg\min}\ \{SURE^Q(y;\mu,\lambda_a,\lambda_b)\}.$$

Estimating equations analogous to (2.19) can be derived by taking the partial derivatives with respect to λ_a,λ_b and plugging in the closed-form solution for $\widehat{\mu}$ (which depends on λ_a,λ_b).

The risk properties of the SURE estimator from section 2.5 can be extended to accommodate missing values, by replacing the second condition of theorem 2.5.1 with a slightly stronger one.

Theorem 2.7.1. *Denote by $\lambda_1(A)$ the largest eigenvalue of a diagonalizable matrix A. Under the following conditions:*

I. $\displaystyle\lim_{r,c\to\infty}\frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\eta_{ij}^2 < \infty$

II. $\displaystyle\lim_{r,c\to\infty}\frac{1}{rc}\lambda_1(M^{-1})\lambda_1(M^{1/2}QM^{1/2}) = 0$ *and* $\displaystyle\lim_{r,c\to\infty}\frac{1}{rc}\lambda_1(M^{-1})\lambda_1^2(M^{1/2}QM^{1/2}) = 0$

it holds that:

(a) $\displaystyle\sup_{|\mu|\leq B;\ \lambda_1,\lambda_2\geq 0} E_{r,c}\Big[SURE^Q(y;\mu,\lambda_a,\lambda_b) - R^Q_{r,c}(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})\Big]^2 \to 0 \qquad$ *as $r,\ c\to\infty$.*

(b) $$\sup_{|\mu|\le B;\ \lambda_1,\lambda_2\ge 0} E_{r,c}\Big[L^Q(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b}) - R^Q_{r,c}(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})\Big]^2 \to 0 \qquad \textit{as } r,\ c\to\infty.$$

where $R^Q(\eta,\hat{\eta}) = E\big[L^Q(\eta,\hat{\eta})\big]$ *and* $L^Q(\eta,\hat{\eta}) = \frac{1}{rc}(\hat{\eta}-\eta)^\top Q(\hat{\eta}-\eta)$.

Remark. As $\lambda_1(M) \le 1$, for the second condition above to hold it suffices that $\lambda_1(M^{-1})[\lambda_1(Q)\vee\lambda_1^2(Q)] = \max(K_{ij})[\lambda_1(Q)\vee\lambda_1^2(Q)] = o(rc)$. Note that when there are no missing values, Q can be replaced by I in the quadratic loss, and the second condition reduces to that of theorem 2.5.1.

Theorem 2.7.1 is proved in the appendix. By the remark above, it is unnecessary to restate it in Theorem 2.5.1 for the special case of all-cells-filled. We nevertheless do so and provide an independent proof for the special case, as both the assumptions and the proof of the theorem take a simpler form when there are no missing values, which makes the proof easier to follow.

The following assertions are counterparts of those following Theorem 2.5.1 of section 2.5. To save space, we do not include the proofs; They follow from Theorem 2.7.1 by exactly the same arguments of the proofs for section 2.5.

Corollary 2.7.2. *Under the conditions of Theorem 2.5.1, it holds that*

$$\sup_{|\mu|\le B;\ \lambda_1,\lambda_2\ge 0} E_{r,c}\Big[SURE(y;\mu,\lambda_a,\lambda_b) - L(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})\Big]^2 \to 0 \qquad \textit{as } r,\ c\to\infty.$$

Now let the loss-oracle be defined by

$$(\ \widetilde{\mu}^{OL},\widetilde{\lambda}_a^{OL},\widetilde{\lambda}_b^{OL}\ ) = \operatorname*{arg\,min}_{|\mu|\le B;\ \lambda_a,\ \lambda_b\ge 0} L^Q(\eta,\widehat{\eta}^{\mu,\lambda_a,\lambda_b})$$

resulting in the rule

$$\widetilde{\eta}^{OL} = \widehat{\eta}^{\widetilde{\mu}^{OL},\widetilde{\lambda}_a^{OL},\widetilde{\lambda}_b^{OL}}. \tag{2.40}$$

Then we have

Theorem 2.7.3. *Under the conditions of Theorem 2.7.1, it holds that for any* $\epsilon > 0$

$$\lim_{\substack{r\to\infty \\ c\to\infty}} P_{r,c}\big\{L^Q(\eta, \widehat{\eta}^{S_Q}) \geq L^Q(\eta, \widetilde{\eta}^{OL}) + \epsilon\big\} = 0 \qquad \text{as } r,\ c \to \infty.$$

Theorem 2.7.4. *Under the conditions of Theorem 2.7.1, it holds that*

$$\lim_{\substack{r\to\infty \\ c\to\infty}} \big\{R^Q_{r,c}(\eta, \widehat{\eta}^{S_Q}) - E_{r,c}[L^Q(\eta, \widetilde{\eta}^{OL})]\big\} = 0 \qquad \text{as } r,\ c \to \infty.$$

As the loss-oracle performs better than any empirical Bayes estimator of the form considered, a consequence of Theorems (2.7.3) and (2.7.4) is

Corollary 2.7.5. *Under the conditions of Theorem 2.7.1, it holds that for any estimator* $\widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b}$ *of the form* (2.9)*,*

(a) $\displaystyle\lim_{\substack{r\to\infty \\ c\to\infty}} P_{r,c}\big\{L^Q(\eta, \widehat{\eta}^{S_Q}) \geq L^Q(\eta, \widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b}) + \epsilon\big\} = 0 \qquad \text{as } r,\ c \to \infty.$

(b) $\displaystyle\limsup_{r\to\infty,\ c\to\infty} \big\{R^Q_{r,c}(\eta, \widehat{\eta}^{S_Q}) - R^Q_{r,c}(\eta, \widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b})\big\} \leq 0 \qquad \text{as } r,\ c \to \infty.$

2.8. Concluding Remarks

We considered a parametric family of Bayes estimators for the two-way unbalanced layout that is based on exchangeability. We suggested an empirical Bayes estimator that uses a criterion directly related to the point risk (conditonal on η) of an estimator, for choosing data-dependent values to substitute for the hyperparameters. In the unbalanced case, the resulting estimator differs from standard empirical versions of the so-called Best Linear Unbiased Predictor (BLUP), and is shown to be asymptotically optimal within the matching family of empirical Bayes estimators.

The theory developed in section 2.2 can be easily extended for the higher-way additive layout. The Bayes estimates will be obtained by considering a prior under which the effects of the i-th factor are i.i.d. Normal variables with mean zero and variance $\sigma^2\lambda_i^2$. However, difficulties might be encountered in the actual computation of the SURE estimator. If there are p factors and the path suggested in section 2.3 is followed, a joint numerical optimization over p variables is required; And the function evaluated in each round involves matrices of rapidly growing dimension with p if there are even moderately large number of levels for each factor. We should mention that the `lmer` (and `blmer`) functions for computing the (ML/REML) EBMLE in the `lme4` R-package are able to handle higher-way situations more efficiently. For example, no inversion of matrices is needed in this implementation. Unfortunately, we did not find a way to completely avoid inverting the matrix V when computing the SURE estimator, which may account for a significant part of the computational burden.

To conclude, we consider estimation under weighted loss. Consider the case where there are no unobserved cells. We concentrated throughout on the usual (normalized) sum-of-squares loss. Alternatively, one might be interested in estimation under a weighted loss function,

$$L(\eta, \widehat{\eta}) = \frac{1}{n}(\widehat{\eta} - \eta)^\top M^{-1}(\widehat{\eta} - \eta) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c} K_{ij}(\widehat{\eta}_{ij} - \eta_{ij})^2. \tag{2.41}$$

where $n = \sum_{i=1}^{r}\sum_{j=1}^{c} K_{ij}$. If x is the vector consisting of the individual homoscedastic observations in the cells x_{ijk} (so that $y_{ij} = \frac{1}{K_{ij}}\sum_{k=1}^{K_{ij}} x_{ijk}$), this corresponds to the usual (normalized) squared loss for the mean of x. Under the weighted loss, applying a linear transformation

$$\widetilde{y} = M^{-1/2}y, \qquad \widetilde{\eta} = 1\widetilde{\mu} + \widetilde{Z}\theta = M^{-1/2}\eta, \qquad \widetilde{Z} = M^{-1/2}Z, \qquad 1\widetilde{\mu} = M^{-1/2}1\mu$$

the problem is equivalent to estimating $\widetilde{\eta}$ from $\widetilde{y} \sim N(\widetilde{\eta}, \sigma^2 I)$ under (normalized) sum-of-squares loss. Assuming the prior in (2.3) with fixed λ_a, λ_b,

$$\widetilde{\eta} \sim N_{rc}(1\widetilde{\mu}, \sigma^2 M^{-1/2} Z \Lambda \Lambda^\top Z^\top M^{-1/2})$$

and the corresponding Bayes estimate of $\widetilde{\eta}$ is

$$\hat{\widetilde{\eta}} = \widetilde{y} - \widetilde{V}^{-1}(\widetilde{y} - 1\widetilde{\mu}), \qquad\qquad \widetilde{V} = M^{-1/2} Z \Lambda \Lambda^{\top} Z^{\top} M^{-1/2} + I.$$

In this case the second condition of Theorem (2.5.1) may be dropped. Indeed, it can be checked in the appendix that the proof goes through if

$$\frac{1}{r^2 c^2} \big\{ 2\mathrm{tr}(G^{\top} G) + 4\eta^{\top} G^{\top} G G^{\top} G \eta \big\} \to 0 \quad \text{ as } r, c \ \to \infty.$$

where $G = \widetilde{V}^{-1}$. Since in the current situation $G^{\top} G - I$ is positive semi-definite, it suffices that

$$\frac{1}{r^2 c^2} \big\{ 2\mathrm{tr}(I) + 4\eta^{\top} \eta \big\} \to 0 \quad \text{ as } r, c \ \to \infty.$$

which is satisfied under the first condition of Theorem (2.5.1).

APPENDIX

A.1. Supplements to Chapter 1

A.1.1. Connection to Efron and Morris (1973b)

Efron and Morris (1973b, Section 9) consider empirical linear Bayes estimates for $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\top$ under the hierarchical model

$$\begin{aligned} \theta_i &\sim (\mu_i, \gamma_i) \\ X_i|\theta_i &\sim (\theta_i, V_i) \end{aligned} \tag{A.1}$$

for $i = 1, ..., n$ where the notation $z \sim (m, \sigma^2)$ is used to indicate that Z is a random variable with no assumptions on its distribution other than mean equal to m and variance equal to σ^2. The hyperparameters $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^\top, \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_n^\top$ are unknown, and V_i are not necessarily known either (in fact, in their setup V_i is allowed to be a function of θ_i, but we assume here that the V_i are constant). They consider empirical versions of the "linear Bayes rule"

$$\theta_i^* = \mu_i + (1 - B_i)(X_i - \mu_i), \qquad B_i = V_i/(\gamma_i + V_i) \tag{A.2}$$

by plugging into (A.2) estimates $\widehat{B}_i(\boldsymbol{X})$ and $\widehat{\mu}_i(\boldsymbol{X})$.

Our empirical Bayes approach to the heteroscedastic normal mean problem fits the framework (A.3) in that θ_i is allowed to depend on V_i, hence according to our model, *conditional* on V_i

$$\begin{aligned} 7\theta_i &\sim (\mu(V_i), \gamma(V_i)) \\ X_i|\theta_i &\sim (\theta_i, V_i) \end{aligned} \tag{A.3}$$

for $i = 1, ..., n$ where V_i are known. As opposed to (A.3), we restrict μ_i and γ_i (and, in fact, the entire distribution of θ_i) to depend on i only through V_i, and the V_i are also assumed to be known. While these assumptions are not necessary for estimators of the form (A.2) with smaller Bayes risk than $\theta_i = X_i$ to exist, they make achievable the more ambitious goal that we pursue, namely, mimicking the best rule $\theta_i = t(X_i, V_i)$ that is linear in X_i (if $\mu_i \neq \mu_j$ even when $V_i = V_j$, this violates the exchaneability assumed between (X_i, θ_i, V_i)).

A.1.2. Proofs

Notations:

1. $\epsilon_{|J|} \doteq \max\limits_{v_1, v_2 \in J} \left\{|a^*(v_1) - a^*(v_2)|, |b^*(v_1) - b^*(v_2)|\right\}$
2. $g(v) \doteq Var(\theta|V = v), h(v) \doteq \mathbb{E}(\theta^2|V = v)$
3. All the expectations in this section are conditional on V.

Lemma A.1.1 (*Analysis within each interval*)**.** *Let $(X_i, \theta_i, V_i)_{i=1}^n$ be iid vectors from a population satisfying (1.12). 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* (1.17) *with $c_n = c_n^*$ satisfies*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\widehat{\theta}_i - \theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{i=1}^n r(a^*, b^*|V_i) + \frac{7V_{\max}}{n \vee 2 - 1} + (\overline{V}\epsilon_{|J|} + |J|)\frac{\varepsilon^2+1}{\varepsilon^2} + \epsilon_{|J|}^2 \\ + \frac{2}{n\vee 2 - 1}\Big\{\sum_{i=1}^n V_i^2 + 2\sum_{i=1}^n (V_i + \overline{V})h(V_i) + \overline{V}^2\Big\}^{\frac{1}{2}}\end{aligned} \tag{A.4}$$

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

Proof of Lemma A.1.1 . As in the proof of Lemma 1.3.1 with $c_n = c_n^*$,

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

By definition of the oracle rule, $r(a^*, b^*|V_i) = V_i(1 - b^*(V_i))$ and $\min\big(s_n^2/\overline{V}, c_n^*\big) \le c_n^* \le 1$, hence

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

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

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

where the last inequality is due to the uniform continuity of $b^*(v)$. Then we are going to bound $\overline{V}(b^*(\overline{V})-E\widehat{b})$, by definition of $b^*(v)$ and $\widehat{b}$

$$\overline{V}(b^*(\overline{V})-E\widehat{b})=\overline{V}\mathbb{E}\Big\{\frac{\overline{V}}{Var(X|V=\overline{V})}-\min(1,\frac{c_n^*\overline{V}}{s_n^2})\Big\}$$

Notice that $\frac{\overline{V}}{Var(X|V=\overline{V})}=\frac{\overline{V}}{\overline{V}+Var(\theta|V=\overline{V})}\leq 1$, then

$$\begin{aligned}\overline{V}(b^*(\overline{V})-E\widehat{b})&\leq\overline{V}\mathbb{E}\Big\{(\frac{\overline{V}}{Var(X|V=\overline{V})}-\frac{c_n^*\overline{V}}{s_n^2})I_{\{c_n^*\overline{V}\leq s_n^2\}}\Big\}\\
&=\frac{\overline{V}}{Var(X|V=\overline{V})}\mathbb{E}\Big\{\frac{\overline{V}s_n^2-c_n^*\overline{V}Var(X|V=\overline{V})}{s_n^2}I_{\{c_n^*\overline{V}\leq s_n^2\}}\Big\}\\
&=\frac{\overline{V}}{Var(X|V=\overline{V})}\mathbb{E}\Big\{\frac{\overline{V}s_n^2-c_n^*\overline{V}s_n^2+c_n^*\overline{V}s_n^2-c_n^*\overline{V}Var(X|V=\overline{V})}{s_n^2}I_{\{c_n^*\overline{V}\leq s_n^2\}}\Big\}\\
&=\frac{\overline{V}}{Var(X|V=\overline{V})}\mathbb{E}\Big\{\overline{V}(1-c_n^*)I_{\{c_n^*\overline{V}\leq s_n^2\}}+\frac{c_n^*\overline{V}}{s_n^2}\big[s_n^2-Var(X|V=\overline{V})\big]I_{\{c_n^*\overline{V}\leq s_n^2\}}\Big\}\end{aligned}$$

Also notice that $1 - c_n^* \geq 0$ and $\frac{c_n^* \overline{V}}{s_n^2} I_{\{c_n^* \overline{V} \leq s_n^2\}} \leq 1$

$$\begin{aligned}
\overline{V}(b^*(\overline{V}) - \widehat{Eb}) &\leq \overline{V}(1 - c_n^*) + \mathbb{E}|s_n^2 - Var(X|V = \overline{V})| \\
&\leq \frac{2V_{\max}}{n \vee 2 - 1} + \mathbb{E}|s_n^2 - Es_n^2| + |\mathbb{E}s_n^2 - Var(X|V = \overline{V})| \\
&= \frac{2V_{\max}}{n \vee 2 - 1} + \mathbb{E}\Big\{\mathbb{E}_{\boldsymbol{V},\boldsymbol{\theta}}|s_n^2 - Es_n^2|\Big\} + |\mathbb{E}s_n^2 - Var(X|V = \overline{V})| \\
&\leq \frac{2V_{\max}}{n \vee 2 - 1} + \mathbb{E}\sqrt{Var(s_n^2|\boldsymbol{V}, \boldsymbol{\theta})} + |\mathbb{E}s_n^2 - Var(X|V = \overline{V})| \\
&\leq \frac{2V_{\max}}{n \vee 2 - 1} + \Big\{\mathbb{E}\Big[Var(s_n^2|\boldsymbol{V}, \boldsymbol{\theta})\Big]\Big\}^{\frac{1}{2}} + |\mathbb{E}s_n^2 - Var(X|V = \overline{V})|
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}(s_n^2) &= \frac{1}{n \vee 2 - 1}\mathbb{E}\Big\{\mathbb{E}(\sum_{i=1}^n X_i^2 - n\overline{X}^2|\boldsymbol{V}, \boldsymbol{\theta})\Big\} \\
&= \frac{1}{n \vee 2 - 1}\mathbb{E}\Big\{\sum_{i=1}^n (V_i + \theta_i^2) - \frac{\sum_{i=1}^n \theta_i^2}{n} - \overline{V}\Big\} \\
&= \overline{V} + \frac{1}{n(n \vee 2 - 1)}\Big\{n\sum_{i=1}^n \mathbb{E}(\theta_i^2|V = V_i) - \sum_{j \neq k} \mathbb{E}(\theta|V = V_j)\mathbb{E}(\theta|V = V_k)\Big\} \\
&= \overline{V} + \frac{1}{n(n \vee 2 - 1)}\Big\{(n-1)\sum_{i=1}^n Var(\theta|V = V_i) + n\sum_{i=1}^n \big[\mathbb{E}(\theta|V = V_i) - \frac{1}{n}\sum_{j=1}^n \mathbb{E}(\theta|V = V_j)\big]^2\Big\} \\
&\leq \overline{V} + \frac{1}{n}\sum_{i=1}^n Var(\theta|V = V_i) + \frac{1}{n \vee 2 - 1}\sum_{i=1}^n \big[\mathbb{E}(\theta|V = V_i) - \frac{1}{n}\sum_{j=1}^n \mathbb{E}(\theta|V = V_j)\big]^2 \\
&= \overline{V} + \frac{1}{n}\sum_{i=1}^n g(V_i) + \frac{1}{n \vee 2 - 1}\sum_{i=1}^n \big[a^*(V_i) - \frac{1}{n}\sum_{j=1}^n a^*(V_j)\big]^2
\end{aligned}$$

On the other hand, $Var(X|V = \overline{V}) = \overline{V} + Var(\theta|V = \overline{V}) = \overline{V} + g(\overline{V})$. Hence,

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

By uniform continuity of $a^*(v)$, $|a^*(V_i) - \frac{1}{n}\sum_{j=1}^n a^*(V_j)| \leq \frac{n-1}{n}\epsilon_{|J|}$. By definition, $b^*(v) =$

$\frac{v}{v+g(v)}$; then $g(v) = \frac{v}{b^*(v)} - v$ and therefore

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

where the last inequality is due to the assumption that $\min_{1\leq i \leq n} b^*(V_i) \geq \varepsilon, b^*(\overline{V}) \geq \varepsilon$. Hence,

$$|\mathbb{E}(s_n^2) - Var(X|V = \bar{V})| \leq \frac{\overline{V}\epsilon_{|J|} + |J|}{\varepsilon^2} + |J| + \epsilon_{|J|}^2$$

Finally, we are going to control $\mathbb{E}\Big\{Var(s_n^2|\boldsymbol{V}, \boldsymbol{\theta})\Big\}$. Agin, use the fact that $\overline{X}|\boldsymbol{V}, \boldsymbol{\theta} \sim N(\frac{\sum_{i=1}^n \theta_i}{n}, \frac{\sum_{i=1}^n V_i}{n^2})$

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

By the definition that $h(v) = Var(\theta^2|V = v)$ and the fact that $n\overline{\theta}^2 \leq \sum_{i=1}^n \theta_i^2$

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

Combining all the inequalities, we have

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

and therefore,

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

$\square$

Proof of Theorem 1.4.2. The first part of Theorem 1.4.2 is direct consequence of Lemma 1.3.1. For the second part, it suffices to prove that $\forall\,\varepsilon>0$, the risk is $O(\varepsilon)$ for large enough n. Noticing that the contribution to the risk for observations outside $\cup_{k=1}^{m}J_k$ is $\sum_{i=1}^{n}V_i I_{\{V_i\notin\cup_{k=1}^{m}J_k\}}/n=o(1)$, then we only need to consider the case where $\forall 1\leq i\leq n$, $V_i\in\cup_{k=1}^{m}J_k$. WLOG, we can assume $\forall\,1\leq k\leq m$, either $J_k\subset[0,\varepsilon)$ or $J_k\subset(\varepsilon,+\infty)$ because 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 always choose M_ε such that $\sum_{i=1}^{n}V_i I_{\{V_i\geq M_\varepsilon\}}/n\leq\varepsilon$ and $\forall\,k$ with $J_k\subset(\varepsilon,+\infty)$, either $J_k\subset(\varepsilon,M_\varepsilon)$ or $J_k\subset(M_\varepsilon,+\infty)$. Let $\overline{V}^k=\sum_{i\in\mathcal{I}_k}V_i/n_k$ and define $S_1=\{k|1\leq k\leq n, J_k\subset(0,\varepsilon)\}, S_2=\{k|1\leq k\leq n, J_k\subset(\varepsilon,M_\varepsilon), \min_{V_i\in J_k}b^*(V_i)\geq\varepsilon, b^*(\overline{V}^k)\geq\varepsilon\}, S_3=\{k|1\leq k\leq n, J_k\subset(\varepsilon,M_\varepsilon), \min_{V_i\in J_k}b^*(V_i)<\varepsilon \textit{ or } b^*(\overline{V}^k)\leq\varepsilon\}, S_4=\{k|1\leq k\leq n, J_k\subset(M_\varepsilon,+\infty)\}$. Then we divide all the observations into four disjoint groups S_1,S_2,S_3,S_4 and now we are going to handle them separately.

Case i) For low 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

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

Case ii) For moderate variance with large shrinkage factor, $V_i\in(\varepsilon,M_\varepsilon)$ and $b^*(V_i),b^*(\overline{V})\geq\varepsilon$, shrinkage is necessary to mimic the performance of the oracle rule. Apply Lemma A.1.1

to each interval J_k such that $k \in S_2$,

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

Let $|J|_{\max} \doteq \max\limits_{1\le k\le m}|J_k|, \epsilon_{\max} \doteq \max\limits_{1\le k\le m}\epsilon_{|J_k|}$ and notice that $\max\limits_{1\le k\le m}\frac{n_k}{n_k\vee 2-1}\le 2$

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

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

$$h(V_i) = Var(\theta|V=V_i)+(\mathbb{E}(\theta|V=V_i))^2 \le \frac{V_i}{b^*(V_i)}-V_i+(a^*(V_i))^2 \le \frac{M_\varepsilon}{\varepsilon}+C_\varepsilon^2$$

Therefore,

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

By Cauthy Schwarz inequality: $\sum_{k\in S_2}n_k^{\frac{1}{2}} \le \sqrt{|S_2|\sum_{k\in S_2}n_k} \le \sqrt{|S_2|n}$. Also notice that $|S_2|\le m\le \frac{n}{\min\limits_{1\le k\le m}n_k}$, then

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

Since $|J|_{\max}, \epsilon_{\max} \to 0$ and $\min\limits_{1\le k\le m} n_k \to +\infty$, 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* $b^*(\overline{V}) < \varepsilon$. By uniform continuity of $b^*(\cdot)$, $\forall i \in \mathcal{I}_k$, $b^*(V_i) \le \varepsilon + \epsilon_{\max}$. Notice that $r(a^*, b^*|V_i) = V_i(1 - b^*(V_i))$, then

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

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

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

Case iv) For high variance part, $V_i \in (M_\varepsilon, +\infty)$, the contribution to the risk is also negligible. By definition of M_ε,

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

Sum the inequalities of all four cases

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] \le \frac{1}{n}\sum_{i=1}^{n} r(a^*,b^*|V_i) + (\overline{V}+2)\varepsilon + o(\varepsilon) \tag{A.5}$$

which finishes the proof by the assumption that $\limsup\limits_{n\to\infty} \frac{\sum_{i=1}^n V_i}{n} \le \infty$ □

Lemma A.1.2 (*Analysis within each interval*)**.** *Given $V_1, \cdots, V_n \in J$ and θ is a deterministic function of V with $a^*(\cdot)$ L-Lipschitz continuous. Under the normal model $X_i|\theta_i, V_i \sim N(\theta_i, V_i)$, the spherically symmetric shrinkage* (1.17) *with $c_n = c_n^*$ satisfies*

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

Proof of Lemma A.1.2. As in the proof of Lemma 1.3.1 and substitue c_n with c_n^*

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

Notice that $2-2c_n^*>0$ and $\widehat{b}c_n^*+(2-c_n^*-s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\leq c_n^*\}}\leq 3$, therefore

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] &\leq \overline{V}\mathbb{E}\left\{(1-\widehat{b}c_n^*)-(c_n^*-s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\leq c_n^*\}}\right\}+\frac{3}{n}\overline{V} \\
&\leq \overline{V}\mathbb{E}\Big\{c_n^*(1-\widehat{b})-(c_n^*-s_n^2/\overline{V})I_{\{s_n^2/\overline{V}\leq c_n^*\}}\Big\}+\frac{3\overline{V}}{n}+(1-c_n^*)\overline{V} \\
&\leq \mathbb{E}\Big\{c_n^*\overline{V}\big(\frac{s_n^2-c_n^*\overline{V}}{s_n^2}\big)_+-\big(c_n^*\overline{V}-s_n^2\big)_+\Big\}+\frac{3\overline{V}}{n}+(1-c_n^*)\overline{V} \\
&\leq \mathbb{E}\Big\{\big(s_n^2-c_n^*\overline{V}\big)_+-\big(c_n^*\overline{V}-s_n^2\big)_+\Big\}+\frac{3\overline{V}}{n}+(1-c_n^*)\overline{V} \\
&= \mathbb{E}(s_n^2-c_n^*\overline{V})+\frac{3\overline{V}}{n}+(1-c_n^*)\overline{V}
\end{aligned}$$

Recall that

$$\mathbb{E}s_n^2=\overline{V}+\frac{1}{n}\sum_{i=1}^n Var(\theta|V=V_i)+\frac{1}{n\vee 2-1}\sum_{i=1}^n\big[E(\theta|V=V_i)-\frac{1}{n}\sum_{j=1}^n E(\theta|V=V_j)\big]^2$$

In the case that $\theta(V)=a^*(V)$, $Es_n^2=\bar{V}+\frac{1}{n\vee 2-1}\sum_{i=1}^n\big[a(V_i)-\frac{1}{n}\sum_{j=1}^n a(V_j)\big]^2$. Therefore,

$$\begin{aligned}
R(\hat{a},\hat{b}|V) &\leq (1-c_n^*)\overline{V}+\frac{1}{n\vee 2-1}\sum_{i=1}^n\big[a(V_i)-\frac{1}{n}\sum_{j=1}^n a(V_j)\big]^2+\frac{3\overline{V}}{n}+(1-c_n^*)\overline{V} \\
&\leq L|J|^2+\frac{3\overline{V}}{n}+2(1-c_n^*)\overline{V} \\
&\leq L|J|^2+\frac{3\overline{V}}{n}+\frac{4V_{\max}}{n\vee 2-1}
\end{aligned}$$

□

Proof of Theorem 2. Apply Lemma A.1.2 to each interval and use the fact that $\frac{n_k}{n_k \vee 2-1} \leq 2$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] &\leq \frac{1}{n}\sum_{k=1}^{m}\big(n_k L|J_k|^2+3\overline{V}^k+4V_{\max}\frac{n_k}{n_k\vee 2-1}\big)\\ &\leq L|J|^2+\frac{3mV_{\max}}{n}+\frac{8mV_{\max}}{n}\\ &= L|J|^2+\frac{11V_{\max}^2}{n|J|}\end{aligned}$$

Let $|J| = \big(\frac{11V_{\max}^2}{nL}\big)^{\frac{1}{3}}$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\widehat{\theta}_i-\theta_i\Big)^2\Big|\boldsymbol{V}\Big] \leq 2\big(\frac{11V_{\max}^2\sqrt{L}}{n}\big)^{\frac{2}{3}}$$

□

A.2. Supplements to Chapter 2

A.2.1. Proofs

Denote $\gamma_{r,c} = \frac{\max\{K_{ij}: i\leq r,\ j\leq c\}}{\min\{K_{ij}: i\leq r,\ j\leq c\}}$. Denote $G = MV^{-1}$ and $G^2 = G^\top G$. We use the notation $\sigma_1(A)$ to denote the k-th largest singular value of a matrix A. We use the notation $\lambda_k(B)$ to denote the k-th largest eigenvalue of a square diagonalizable matrix B.

Proof of Theorem 2.5.1. (a) Fix $\lambda_a, \lambda_b \geq 0$. We consider first the case $\mu = 0$ and show that

$$E[SURE(y;0,\lambda_a,\lambda_b) - R_{r,c}(\eta;\widehat{\eta}^{0,\lambda_a,\lambda_b})]^2 \to \infty \quad \text{as } r, c \to \infty.$$

We have

$$E[SURE(y;0,\lambda_a,\lambda_b) - R_{r,c}(\eta;\widehat{\eta}^{\mu,\lambda_a,\lambda_b})]^2 = \mathrm{Var}[SURE(y;0,\lambda_a,\lambda_b)] = \frac{1}{r^2c^2}\mathrm{Var}(y^\top G^2 y)$$
$$= \frac{1}{r^2c^2}\{2\mathrm{tr}(G^2MG^2M) + 4\eta^\top G^2MG^2\eta\} \tag{A.6}$$

Letting $W = M^{\frac{1}{2}}V^{-1}M^{\frac{1}{2}}$ and noting that $W \leq I$,

$$\begin{aligned}\lambda_1(G^2MG^2) &= \lambda_1(V^{-1}M^2V^{-1}MV^{-1}M^2V^{-1}) = \lambda_1(V^{-1}M^{\frac{3}{2}}W^2M^{\frac{3}{2}}V^{-1})\\ &\leq \lambda_1(V^{-1}M^3V^{-1}) = \lambda_1(M^{-\frac{1}{2}}WM^2WM^{-\frac{1}{2}}) = \sigma_1^2(M^{-\frac{1}{2}}WM)\\ &\leq \{\sigma_1(M^{-\frac{1}{2}})\sigma_1(W)\sigma_1(M)\}^2 = \{\lambda_1^{1/2}(M^{-1})\lambda_1(W)\lambda_1(M)\}^2\\ &\leq \lambda_1^2(M)\lambda_1(M^{-1}) = (\max K_{ij})/(\min K_{ij})^2 \leq \gamma_{r,c}\end{aligned}$$

Since $\eta^\top G^2MG^2\eta \leq \lambda_1(G^2MG^2)\|\eta\|^2$, it follows that

$$\eta^\top G^2MG^2\eta \leq \gamma_{r,c}\|\eta\|^2.$$

Also,

$$\begin{aligned}\mathrm{tr}(G^2MG^2M) &= \mathrm{tr}(M^{\frac{1}{2}}G^2MG^2M^{\frac{1}{2}}) = \mathrm{tr}(M^{\frac{1}{2}}V^{-1}M^2V^{-1}MV^{-1}M^2V^{-1}M^{\frac{1}{2}})\\ &= \mathrm{tr}(M^{\frac{1}{2}}V^{-1}M^{\frac{3}{2}}W^2M^{\frac{3}{2}}V^{-1}M^{\frac{1}{2}}) \leq \mathrm{tr}(M^{\frac{1}{2}}V^{-1}M^3V^{-1}M^{\frac{1}{2}})\\ &= \mathrm{tr}(WM^2W) = \mathrm{tr}(MW^2M) \leq \mathrm{tr}(M^2)\end{aligned}$$

Hence the RHS of (A.6) is no more than $\frac{1}{r^2c^2}\{2\mathrm{tr}(M^2) + 4\gamma_{r,c}\|\eta\|^2\} \to 0$ as $r, c \to \infty$ by the assumptions of the theorem.

Now, for arbitrary $\mu \in \mathbb{R}$, we have

$$SURE(y;\mu,\lambda_a,\lambda_b) = SURE(y;0,\lambda_a,\lambda_b) + \frac{1}{rc}\mu^2 1^\top G^2 1 - \frac{1}{rc}2\mu y^\top G^2 1,$$

hence

$$\begin{aligned} E[SURE(y;\mu,\lambda_a,\lambda_b) - R_{r,c}(\eta;\widehat{\eta}^{\mu,\lambda_a,\lambda_b})]^2 &= \mathrm{Var}[SURE(y;\mu,\lambda_a,\lambda_b)] \\ &\leq 4\{\mathrm{Var}[SURE(y;0,\lambda_a,\lambda_b)] + 4\frac{1}{rc}\mu^2 \mathrm{Var}(y^\top G^2 1)\}. \end{aligned}$$

The first term on the RHS was shown above to tend to 0. As for the second term, we need to show that $\frac{4\mu^2}{r^2c^2} 1^\top G^2 M G^2 1 \to 0$ as $r, c \to \infty$.

Since $\lambda_1(G^2 M G^2) \leq \gamma_{r,c}$ as was shown above, it follows that

$$\frac{1}{r^2c^2} 1^\top G^2 M G^2 1 \leq \frac{1}{r^2c^2} rc\gamma_{r,c} \leq \frac{1}{rc}\gamma_{r,c} \to 0 \text{ as } r, c \to \infty$$

implying that $\frac{4\mu^2}{r^2c^2} 1^\top G^2 M G^2 1 \to 0$ as $r, c \to \infty$ for any μ.

As μ is bounded, and since all bounds above on terms in (A.6) are indpendent of λ_a or λ_b, it follwos that $\sup\limits_{\lambda_a,\lambda_b \geq 0} \mathrm{Var}[SURE(y;\mu,\lambda_a,\lambda_b)]$ also tends to zero as $r, c \to \infty$.

(b) Fix $\lambda_a, \lambda_b \geq 0$. We consider first the case $\mu = 0$ and show that

$$E\big\{|L(\eta,\widehat{\eta}^{0,\lambda_a,\lambda_b}) - R_{r,c}(\eta;\widehat{\eta}^{0,\lambda_a,\lambda_b})|\big\} \to 0 \text{ as } r, c \to \infty.$$

For $\widehat{\eta} = \widehat{\eta}^{0,\lambda_a,\lambda_b}$ we have

$$\begin{aligned} L(\eta,\widehat{\eta}) &= \frac{1}{rc}(\widehat{\eta} - \eta)^\top(\widehat{\eta} - \eta) = \frac{1}{rc}(y - \eta - Gy)^\top(y - \eta - Gy) \\ &= \frac{1}{rc}\big\{(y-\eta)^\top(y-\eta) + y^\top G^2 y - 2(y-\eta)^\top G y\big\} \\ &= \frac{1}{rc}\big\{\underbrace{(y-\eta)^\top(y-\eta)}_{Q_1} + \underbrace{y^\top G^2 y}_{Q_2} - \underbrace{2y^\top G y}_{Q_3} + \underbrace{2\eta^\top G y}_{Q_4}\big\} \end{aligned}$$

hence it suffices to show that for each of the four terms above $E|\frac{1}{rc}Q_i - E(\frac{1}{rc}Q_i)| \to \infty$ as $r, c \to \infty$, which in turn will follow if we show that $\mathrm{Var}(\frac{1}{rc}Q_i) \to 0$ as $r, c \to \infty$.

Q_1: $\frac{1}{rc}\mathrm{Var}[(y-\eta)^\top(y-\eta)] = \frac{1}{rc}2\mathrm{tr}(M^2) \to 0$ as $r, c \to \infty$.

Q_2: $\frac{1}{rc}\mathrm{Var}(y^\top G^2 y) \to 0$ as $r, c \to \infty$ was already shown in the proof of part (a).

Q_3: $4\mathrm{Var}(y^\top G y) = 2\mathrm{tr}(\widetilde{G}M\widetilde{G}M) + 4\eta^\top \widetilde{G}M\widetilde{G}\eta, \quad \widetilde{G} = G + G^\top.$

For the second of the two terms on the RHS, as $\widetilde{G} = G + G^\top = MV^{-1} + V^{-1}M$,

$$\widetilde{G}M\widetilde{G} = \underbrace{MV^{-1}MMV^{-1}}_{Q_{31}} + \underbrace{MV^{-1}MV^{-1}M}_{Q_{32}} + \underbrace{V^{-1}MMMV^{-1}}_{Q_{33}} + \underbrace{V^{-1}MMV^{-1}M}_{Q_{34}}.$$

We show that $\frac{1}{r^2c^2}|\eta^\top Q_{3i}\eta| \to 0$ as $r, c \to \infty$ for $i = 1, ..., 4$. Under the assumption of the theorem,

$$\begin{aligned}\lambda_1(G^2) = \lambda_1(V^{-1}M^2V^{-1}) = \lambda_1(M^{-\frac{1}{2}}WMWM^{-\frac{1}{2}}) &= \sigma_1^2\{M^{-\frac{1}{2}}WM^{\frac{1}{2}}\} \\ &\leq \{\sigma_1(M^{-\frac{1}{2}})\sigma_1(W)\sigma_1(M^{\frac{1}{2}})\}^2 \\ &\leq \{\sigma_1(M^{-\frac{1}{2}})\sigma_1(M^{\frac{1}{2}})\}^2 = \gamma_{r,c}\end{aligned}$$

hence

$$\lambda_1(MV^{-1}MMV^{-1}) \leq \lambda_1(M)\lambda_1(V^{-1}M^2V^{-1}) = \lambda_1(M)\lambda_1(G^2) \leq \lambda_1(G^2) \leq \gamma_{r,c}$$

and we note that $MV^{-1}MMV^{-1}$ is indeed diagonalizable as a product of two p.s.d. matrices. Therefore,

$$\frac{1}{r^2c^2}|\eta^\top(MV^{-1}MMV^{-1})\eta| \leq \frac{1}{r^2c^2}\lambda_1(MV^{-1}MMV^{-1})\|\eta\|^2 \leq \frac{1}{r^2c^2}\gamma_{r,c}\|\eta\|^2 \to 0$$

as $r, c \to \infty$.

Since $\eta^\top(V^{-1}MMV^{-1}M)\eta = \eta^\top(MV^{-1}MMV^{-1})\eta$, it follows that $\frac{1}{r^2c^2}|\eta^\top(V^{-1}MMV^{-1}M)\eta| \to 0$ as well.

Also,

$$\begin{aligned}\lambda_1(MV^{-1}MV^{-1}M) &\leq \lambda_1(MV^{-2}M)\\ &= \lambda_1\{M^{\frac{1}{2}}(M^{\frac{1}{2}}V^{-2}M^{\frac{1}{2}})M^{\frac{1}{2}}\}\\ &\leq \lambda_1(M^{-1})\lambda_1(M) = \gamma_{r,c}\end{aligned}$$

where the last inequality follows from

$$M^{\frac{1}{2}}V^{-2}M^{\frac{1}{2}} = (M^{\frac{1}{2}}V^{-2}M^{\frac{1}{2}})M^{-1}(M^{\frac{1}{2}}V^{-2}M^{\frac{1}{2}}) \leq \lambda_1(M^{-1})I.$$

Hence, $\frac{1}{r^2c^2}\eta^\top(MV^{-1}MV^{-1}M)\eta \leq \frac{1}{r^2c^2}\gamma_{r,c}\|\eta\|^2 \to 0$ as $r, c \to \infty$.

Finally, $\lambda_1(V^{-1}M^3V^{-1}) \leq \gamma_{r,c}$ was shown in the proof of part (a) of the theorem, implying

$$\frac{1}{r^2c^2}\eta^\top V^{-1}M^3V^{-1}\eta \leq \frac{1}{r^2c^2}\gamma_{r,c}\|\eta\|^2 \to 0$$

as $r, c \to \infty$.

Now, as for $\mathrm{tr}(\widetilde{G}M\widetilde{G}M)$ we have

$$\widetilde{G}M\widetilde{G}M = GMGM + GMG^\top M + G^\top MGM + G^\top MG^\top M$$

hence

$$\mathrm{tr}(\widetilde{G}M\widetilde{G}M) = 2\mathrm{tr}(GMGM) + 2\mathrm{tr}(GMG^\top M).$$

Since

$$\begin{aligned}\operatorname{tr}(GMGM) = \operatorname{tr}(MV^{-1}M^2V^{-1}M) = \operatorname{tr}(MG^2M) &\le \operatorname{tr}\{M[\lambda_1(G^2)I]M\} \\ &= \lambda_1(G^2)\operatorname{tr}(M^2) = \gamma_{r,c}\operatorname{tr}(M^2),\end{aligned}$$

then

$$\frac{1}{r^2c^2}\operatorname{tr}(GMGM) \le \frac{1}{r^2c^2}\gamma_{r,c}\operatorname{tr}(M^2) \to 0$$

as $r, c \to \infty$.

Also,

$$\begin{aligned}\operatorname{tr}(GMG^\top M) = \operatorname{tr}(MV^{-1}MV^{-1}MM) &= \operatorname{tr}(MM^{\frac{1}{2}}V^{-1}MV^{-1}M^{\frac{1}{2}}M) \\ &= \operatorname{tr}(MW^2M) \le \operatorname{tr}(M^2)\end{aligned}$$

and so

$$\frac{1}{r^2c^2}\operatorname{tr}(GMG^\top M) \le \frac{1}{r^2c^2}\operatorname{tr}(M^2) \to 0$$

as $r, c \to \infty$.

Together, we have

$$\frac{1}{r^2c^2}\operatorname{tr}(\widetilde{G}M\widetilde{G}M) \to 0$$

as $r, c \to \infty$

We conclude that $\frac{1}{r^2c^2}\operatorname{Var}(y^\top Gy) \to 0$ as $r, c \to \infty$.

Q_4: $\operatorname{Var}(\eta^\top Gy) = \eta^\top GMG^\top \eta$.

Since

$$\lambda_1(GMG^\top) = \lambda_1(MV^{-1}MV^{-1}M) = \lambda_1(M^{\frac{1}{2}}W^2M^{\frac{1}{2}}) \leq \lambda_1(M) \leq 1$$

we have

$$\frac{1}{r^2c^2}\eta^\top GMG^\top \eta \leq \frac{1}{r^2c^2}\lambda_1(GMG^\top)\|\eta\|^2 \to 0$$

as $r, c \to \infty$.

Turning to the case of arbitrary $\mu \in \mathbb{R}$, we first note that

$$\begin{aligned}(\widehat{\eta}^{\mu,\lambda_a,\lambda_b} - \eta)^\top(\widehat{\eta}^{\mu,\lambda_a,\lambda_b} - \eta) =&(\widehat{\eta}^{0,\lambda_a,\lambda_b} - \eta)^\top(\widehat{\eta}^{0,\lambda_a,\lambda_b} - \eta) + \mu 1^\top G^\top G1 \\ &+ 2\mu 1^\top G^\top(I-G)y - 2\mu 1^\top G\eta.\end{aligned} \tag{A.7}$$

For fixed μ, to prove that

$$E\big\{\frac{1}{rc}|(\widehat{\eta}^{\mu,\lambda_a,\lambda_b} - \eta)^\top(\widehat{\eta}^{\mu,\lambda_a,\lambda_b} - \eta)|\big\} \to 0$$

it is enough to show that the variance of each random term in (A.7) is $o((rc)^2)$. The first term in (A.7) has been just dealt with in the $\mu = 0$ case. Therefore, it remains to show that

$$\mu^2\frac{1}{r^2c^2}\text{Var}(1^\top G^\top(I-G)y) \to 0$$

as $r, c \to \infty$. Now,

$$\begin{aligned}\text{Var}(1^\top G^\top(I-G)y) = 1^\top G^\top(I-G)M(I-G)^\top G1 &\leq \lambda_1(G^\top(I-G)M(I-G)^\top G)1^\top 1 \\ &= rc \cdot \lambda_1(G^\top(I-G)M(I-G)^\top G).\end{aligned}$$

Let $L = M^{-\frac{1}{2}}(M - GM)M^{-1}(M - GM)M^{-\frac{1}{2}}$ (note: $(M - GM)$ is symmetric). Since

$M - GM \leq M$, then $M^{-\frac{1}{2}}((M - GM))M^{-\frac{1}{2}} \leq I$, and by squaring we get that $L \leq I$. Hence,

$$\begin{aligned} G^\top(I-G)M(I-G)^\top G = G^\top(M-GM)M^{-1}(M-GM)^\top G &= G^\top M^{\frac{1}{2}} L M^{\frac{1}{2}} G \\ &\leq G^\top M G \\ &\leq G^2 \end{aligned}$$

and so

$$\lambda_1(G^\top(I-G)M(I-G)^\top G) \leq \lambda_1(G^2) \leq \gamma_{r,c}.$$

In conclusion,

$$\frac{1}{r^2c^2}\mathrm{Var}(1^\top G^\top(I-G)y) \leq \frac{1}{r^2c^2}\gamma_{r,c} \cdot rc = \frac{1}{rc}\gamma_{r,c} \to 0$$

as $r, c \to \infty$.

Now, μ is bounded by assumption, while all bounds derived above are independent of λ_a and λ_b, therefore $\sup_{\lambda_a,\lambda_b \geq 0} \big\{\frac{1}{r^2c^2}\mathrm{Var}[l_{r,c}(\eta; \widehat{\eta}^{\mu,\lambda_a,\lambda_b})]\big\}$ also tends to zero as $r, c \to \infty$, and (b) is proved.

□

Proof of Theorem 2.5.3. By definition, $SURE(y; \widehat{\mu}^{SURE}, \widehat{\lambda}_a^{SURE}, \widehat{\lambda}_b^{SURE}) \leq SURE(y; \widetilde{\mu}^{OL}, \widetilde{\lambda}_a^{OL}, \widetilde{\lambda}_b^{OL})$, hence

$$P_{r,c}\big\{L(\eta, \widehat{\eta}^{SURE}) \geq L(\eta, \widetilde{\eta}^{OL}) + \epsilon\big\} \leq P_{r,c}\big\{A(y;\eta) \geq B(y;\eta) + \epsilon\big\}$$

where

$$
\begin{aligned}
A(y;\eta) &= L(\eta, \widehat{\eta}^{SURE}) - SURE(y; \widehat{\mu}^{SURE}, \widehat{\lambda}_a^{SURE}, \widehat{\lambda}_b^{SURE}) \\
B(y;\eta) &= L(\eta, \widetilde{\eta}^{OL}) - SURE(y; \widetilde{\mu}^{OL}, \widetilde{\lambda}_a^{OL}, \widetilde{\lambda}_b^{OL}).
\end{aligned}
$$

Using Markov's inequality and the fact that $A(y;\eta) - B(y;\eta) \geq 0$,

$$
\begin{aligned}
P_{r,c}\big\{A(y;\eta) \geq B(y;\eta) + \epsilon\big\} &\leq \epsilon^{-1} E_{r,c}\{A(y;\eta) - B(y;\eta)\} \\
&\leq 2\epsilon^{-1} \sup_{|\mu|\leq B;\ \lambda_1,\lambda_2\geq 0} E_{r,c}|L(\eta, \widehat{\eta}^{\mu,\lambda_a,\lambda_b}) - SURE(y;\mu,\lambda_a,\lambda_b)|.
\end{aligned}
$$

Incorporating corollary (2.5.2), the last term tends to zero as $r, c \to \infty$. □

Proof of Theorem 2.5.4. Write

$$
\begin{aligned}
L(\eta, \widetilde{\eta}^{SURE}) - L(\eta, \widetilde{\eta}^{OL}) =& \{L(\eta, \widetilde{\eta}^{SURE}) - SURE(y; \widehat{\mu}^{SURE}, \widehat{\lambda}_a^{SURE}, \widehat{\lambda}_b^{SURE})\} \\
&- \{L(\eta, \widetilde{\eta}^{OL}) - SURE(y; \widetilde{\mu}^{OL}, \widetilde{\lambda}_a^{OL}, \widetilde{\lambda}_b^{OL})\} \\
&+ \{SURE(y; \widehat{\mu}^{SURE}, \widehat{\lambda}_a^{SURE}, \widehat{\lambda}_b^{SURE}) - SURE(y; \widetilde{\mu}^{OL}, \widetilde{\lambda}_a^{OL}, \widetilde{\lambda}_b^{OL})\}.
\end{aligned}
$$

By definition of $(\widehat{\mu}^{SURE}, \widehat{\lambda}_a^{SURE}, \widehat{\lambda}_b^{SURE})$, the last term is nonpositive, therefore

$$
E_{r,c}\{L(\eta, \widetilde{\eta}^{SURE}) - L(\eta, \widetilde{\eta}^{OL})\} \leq 2 \sup_{|\mu|\leq B;\ \lambda_1,\lambda_2\geq 0} E_{r,c}|L(\eta, \widehat{\eta}^{\mu,\lambda_a,\lambda_b}) - SURE(y;\mu,\lambda_a,\lambda_b)|
$$

which tends to zero as $r, c \to \infty$ by Corollary 2.5.2. □

Proof of Corollary 2.5.5. (a) and (b) are direct consequences, respectively, of Theorems 2.5.3 and 2.5.4, since $L(\eta, \widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b}) \geq L(\eta, \widetilde{\eta}^{OL})$ and, hence, also $E_{r,c}L(\eta, \widehat{\eta}^{\widehat{\mu},\widehat{\lambda}_a,\widehat{\lambda}_b}) \geq E_{r,c}L(\eta, \widetilde{\eta}^{OL})$. □

A.2.2. Details for Section 2.2

The following facts about derivatives involving matrix expressions are used below.

(i) $\frac{\partial}{\partial x}\{x^\top Ax\} = x^\top(A + A^\top)$

(ii) $\frac{\partial}{\partial \alpha} \log |A| = \mathrm{tr}(A^{-1}\frac{\partial A}{\partial \alpha})$

(iii) $\frac{\partial}{\partial \alpha} A^{-1} = -A^{-1}\frac{\partial A}{\partial \alpha}A^{-1}$

(iv) $\frac{\partial}{\partial \alpha}\{UBV\} = \frac{\partial U}{\partial \alpha}BV + UB\frac{\partial V}{\partial \alpha}$ for matrices U, B, V where B is constant w.r.t. α

ML estimates are computed based on the likelihood of y in the hierarchical model (2.4). Our derivation is very similar to Searle and McCulloch (2001, ch. 6.3, 6.4, 6.8 and 6.12) but with slightly different notation. Since $y \sim N_{rc}(1\mu, \sigma^2 V)$,

$$f = \frac{1}{(2\pi\sigma^2)^{rc/2}|V|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - 1\mu)^\top V^{-1}(y - 1\mu)\right\} \tag{A.8}$$

and the log-likelihood is

$$l(\mu, \boldsymbol{\theta}) = -(rc)/2 \cdot \log(2\pi\sigma^2) - \frac{1}{2}\log|V| - \frac{1}{2\sigma^2}(y - 1\mu)^\top V^{-1}(y - 1\mu) \tag{A.9}$$

Using chain rule,

$$\frac{\partial l}{\partial \mu} \overset{\text{(i)}}{=} -\frac{1}{\sigma^2}(y - 1\mu)^\top V^{-1}\frac{\partial\{y - 1\mu\}}{\partial \mu} = \frac{1}{\sigma^2}(y - 1\mu)^\top V^{-1}1 \tag{A.10}$$

Also,

$$\begin{aligned}\frac{\partial l}{\partial \lambda_a^2} &\overset{\text{(ii)}}{=} -\frac{1}{2}\mathrm{tr}\left(V^{-1}\frac{\partial V}{\partial \lambda_a^2}\right) - \frac{1}{2\sigma^2}(y - 1\mu)^\top \left[\frac{\partial V^{-1}}{\partial \lambda_a^2}\right](y - 1\mu) \\ &= -\frac{1}{2}\left\{\mathrm{tr}\left(V^{-1}\frac{\partial V}{\partial \lambda_a^2}\right) + \frac{1}{\sigma^2}(y - 1\mu)^\top \left[\frac{\partial V^{-1}}{\partial \lambda_a^2}\right](y - 1\mu)\right\} \\ &\overset{\text{(iii)}}{=} -\frac{1}{2}\left\{\mathrm{tr}\left(V^{-1}\frac{\partial V}{\partial \lambda_a^2}\right) - \frac{1}{\sigma^2}(y - 1\mu)^\top V^{-1}\left[\frac{\partial V}{\partial \lambda_a^2}\right]V^{-1}(y - 1\mu)\right\} \\ &= -\frac{1}{2}\left\{\mathrm{tr}\left(V^{-1}Z_aZ_a^\top\right) - \frac{1}{\sigma^2}(y - \boldsymbol{\mu})^\top V^{-1}Z_aZ_a^\top V^{-1}(y - \mu)\right\} \qquad \text{(A.11)}\end{aligned}$$

where in the last equality we use the fact that

$$V = \lambda_a^2 Z_a Z_a^\top + \lambda_b^2 Z_b Z_b^\top + \sigma^2 M. \qquad \text{(A.12)}$$

On equating to zero, we get from (A.10)

$$\mu = \frac{1^\top V^{-1} y}{1^\top V^{-1} 1}, \qquad \text{(A.13)}$$

the GLS estimate of μ. From (A.11),

$$\mathrm{tr}\left(V^{-1}Z_aZ_a^\top\right) - \frac{1}{\sigma^2}(y - 1\mu)^\top V^{-1}Z_aZ_a^\top V^{-1}(y - 1\mu) = 0. \qquad \text{(A.14)}$$

By symmetry, taking the partial derivative w.r.t. λ_b^2 gives

$$\mathrm{tr}\left(V^{-1}Z_bZ_b^\top\right) - \frac{1}{\sigma^2}(y - 1\mu)^\top V^{-1}Z_bZ_b^\top V^{-1}(y - 1\mu) = 0. \qquad \text{(A.15)}$$

Plugging (A.13) into (A.14) and (A.15) gives the estimating equations for λ_a^2 and λ_b^2 as

$$\begin{aligned}\mathrm{tr}\left(V^{-1}Z_aZ_a^\top\right) - \frac{1}{\sigma^2}y^\top(I - P)^\top V^{-1}Z_aZ_a^\top V^{-1}(I - P)y = 0 \qquad \text{(A.16)} \\ \mathrm{tr}\left(V^{-1}Z_bZ_b^\top\right) - \frac{1}{\sigma^2}y^\top(I - P)^\top V^{-1}Z_bZ_b^\top V^{-1}(I - P)y = 0 \qquad \text{(A.17)}\end{aligned}$$

where

$$P = 1(1^\top V^{-1}1)^{-1}1^\top V^{-1} \tag{A.18}$$

is the GLS projection matrix.

2.1.SURE. Noting that in (2.16), in comparison to (A.9), $V^{-1}M^2V^{-1}$ replaces V^{-1}, hence the partial derivative w.r.t. μ vanishes for

$$\mu = \frac{1^\top[V^{-1}M^2V^{-1}]y}{1^\top[V^{-1}M^2V^{-1}]1}. \tag{A.19}$$

Furthermore,

$$\begin{aligned}
&\frac{\partial}{\partial\lambda_a^2}SURE = \\
&\overset{\text{(iv)}}{=} -2\sigma^2\text{tr}\left(\frac{\partial V^{-1}}{\partial\lambda_a^2}M^2\right) + (y-1\mu)^\top\left\{\frac{\partial V^{-1}}{\partial\lambda_a^2}M^2V^{-1} + V^{-1}M^2\frac{\partial V^{-1}}{\partial\lambda_a^2}\right\}(y-1\mu) \\
&= -2\sigma^2\text{tr}\left(\frac{\partial V^{-1}}{\partial\lambda_a^2}M^2\right) + 2(y-1\mu)^\top\left[\frac{\partial V^{-1}}{\partial\lambda_a^2}M^2V^{-1}\right](y-1\mu) \\
&\overset{\text{(iii)}}{=} 2\sigma^2\text{tr}\left(V^{-1}\frac{\partial V}{\partial\lambda_a^2}V^{-1}M^2\right) - 2(y-1\mu)^\top\left[V^{-1}\frac{\partial V}{\partial\lambda_a^2}V^{-1}M^2V^{-1}\right](y-1\mu) \\
&= 2\sigma^2\text{tr}(V^{-1}Z_aZ_a^\top V^{-1}M^2) - 2(y-1\mu)^\top[V^{-1}Z_aZ_a^\top V^{-1}M^2V^{-1}](y-1\mu)
\end{aligned} \tag{A.20}$$

Hence, on equating (A.11) to zero we obtain

$$\text{tr}(V^{-1}Z_aZ_a^\top V^{-1}M^2) - \frac{1}{\sigma^2}(y-1\mu)^\top[V^{-1}Z_aZ_a^\top V^{-1}M^2V^{-1}](y-1\mu) = 0 \tag{A.21}$$

By symmetry, equating the partial derivative w.r.t. λ_b^2 to zero gives

$$\text{tr}(V^{-1}Z_bZ_b^\top V^{-1}M^2) - \frac{1}{\sigma^2}(y-1\mu)^\top[V^{-1}Z_bZ_b^\top V^{-1}M^2V^{-1}](y-1\mu) = 0 \tag{A.22}$$

Plugging (A.19) into (A.21) and (A.22) gives the estimating equations for λ_a^2, λ_b^2 as

$$\operatorname{tr}\left(V^{-1}Z_aZ_a^\top V^{-1}M^2\right) - \frac{1}{\sigma^2}y^\top(I-\widetilde{P})^\top V^{-1}Z_aZ_a^\top V^{-1}M^2V^{-1}(I-\widetilde{P})y = 0 \tag{A.23}$$

$$\operatorname{tr}\left(V^{-1}Z_bZ_b^\top V^{-1}M^2\right) - \frac{1}{\sigma^2}y^\top(I-\widetilde{P})^\top V^{-1}Z_bZ_b^\top V^{-1}M^2V^{-1}(I-\widetilde{P})y = 0 \tag{A.24}$$

where

$$\widetilde{P} = 1(1^\top V^{-1}M^2V^{-1}1)^{-1}1^\top V^{-1}M^2V^{-1}. \tag{A.25}$$

A.2.3. Details for Section 2.3

By definition, $V = Z\Lambda\Lambda^\top Z^\top + M$. We apply the matrix inverse identity to get

$$V^{-1} = M^{-1} - M^{-1}Z\Lambda(\Lambda^\top Z^\top M^{-1}Z\Lambda + I_q)^{-1}\Lambda^\top Z^\top M^{-1}. \tag{A.26}$$

Hence, also

$$MV^{-1} = I_{rc} - Z\Lambda(\Lambda^\top Z^\top M^{-1}Z\Lambda + I_q)^{-1}\Lambda^\top Z^\top M^{-1} \tag{A.27}$$

and

$$MV^{-1}M = M - Z\Lambda(\Lambda^\top Z^\top M^{-1}Z\Lambda + I_q)^{-1}\Lambda^\top Z^\top. \tag{A.28}$$

From (A.28),

$$\begin{aligned}\operatorname{tr}(V^{-1}M^2) &= \operatorname{tr}(MV^{-1}M)\\ &= \operatorname{tr}(M) - \operatorname{tr}(Z\Lambda(\Lambda^\top Z^\top M^{-1}Z\Lambda + I_q)^{-1}\Lambda^\top Z^\top).\end{aligned} \tag{A.29}$$

Therefore, (2.16) can be written as

$$SURE = -\sigma^2 \mathrm{tr}(M) + 2\sigma^2 \mathrm{tr}\{(\Lambda^\top Z^\top M^{-1} Z\Lambda + I_q)^{-1}(\Lambda^\top Z^\top Z\Lambda)\} + \|MV^{-1}(y - 1\mu)\|^2 \quad (A.30)$$

In computing (A.30):

1. The middle term is computed as the sum of the *elementwise* product of $(\Lambda^\top Z^\top M^{-1} Z\Lambda + I_q)^{-1}$ and $\Lambda^\top Z^\top Z\Lambda$, using the property $\mathrm{tr}(A^\top B) = \sum_{i,j} A_{ij} B_{ij}$

2. $(\Lambda^\top Z^\top M^{-1} Z\Lambda + I_q)^{-1}$ is computed efficiently employing a sparse Cholesky factorization of $\Lambda^\top Z^\top M^{-1} Z\Lambda + I_q$ similarly to the implementation in the lme4 package in `R`.

3. The quantity $\min_\mu \|MV^{-1}(y - 1\mu)\|^2$ is computed by regressing $MV^{-1}y$ on $MV^{-1}1_{rc}$ using the `lm` function in `R`. In doing that, the vector $MV^{-1}x$ (for $x = y$ and $x = 1_{rc}$) is computed as (using (A.27))

$$MV^{-1}x = x - Z\Lambda(\Lambda^\top Z^\top M^{-1} Z\Lambda + I_q)^{-1}\Lambda^\top Z^\top (M^{-1}x) \quad (A.31)$$

where (A.31) is implemented proceeding ”from right to left” to always compute a product of a matrix and a *vector*, instead of two matrices: First find $M^{-1}x$, then find $(\Lambda^\top Z^\top)(M^{-1}x)$, and so on.

BIBLIOGRAPHY

K. Alam. A family of admissible minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 517–525, 1973.

K. Alam and J. R. Thompson. Estimation of the mean of a multivariate normal distribution. Technical Report BU-213-M, Indiana University, 1964.

A. J. Baranchik. *Multiple regression and estimation of the mean of a multivariate normal distribution*. PhD thesis, Dept. of Statistics, Stanford University., 1964.

A. J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, 41(2):642–645, 1970.

D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. URL `http://CRAN.R-project.org/package=lme4`. R package version 1.1-7.

D. M. Bates. lme4: Mixed-effects modeling with r. `http://lme4.r-forge.r-project.org/book`, 2010.

J. Berger. A robust generalized bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, pages 716–761, 1980.

J. Berger, M. Bock, L. Brown, G. Casella, and L. Gleser. Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *The Annals of Statistics*, pages 763–771, 1977.

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

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

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

P. Bhattacharya. Estimating the mean of a multivariate normal population with general quadratic loss function. *The Annals of Mathematical Statistics*, pages 1819–1824, 1966.

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

L. D. Brown. On the admissibility of invariant estimators of one or more location parameters. *The Annals of Mathematical Statistics*, pages 1087–1136, 1966.

L. 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.

L. 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.

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

D. Edelman. Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *The Annals of Statistics*, 16(4):1609–1622, 1988.

B. Efron and C. Morris. Empirical bayes on vector observations: An extension of stein's method. *Biometrika*, 59(2):335–347, 1972a.

B. Efron and C. Morris. Limiting the risk of bayes and empirical bayes estimatorspart ii: The empirical bayes case. *Journal of the American Statistical Association*, 67(337): 130–139, 1972b.

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

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

B. Efron and C. Morris. Families of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 11–21, 1976.

E. I. George et al. Minimax multiple shrinkage estimation. *The Annals of Statistics*, 14(1): 188–205, 1986.

D. G. Herr. On the history of anova in unbalanced, factorial designs: The first 30 years. *The American Statistician*, 40(4):265–270, 1986.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

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

W. Jiang and C.-H. 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.

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

R. Koenker and I. Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506): 674–685, 2014.

S. Kou and J. J. Yang. Optimal shrinkage estimation in heteroscedastic hierarchical linear models. *arXiv preprint arXiv:1503.06262*, 2015.

K.-C. Li and J. T. Hwang. The data-smoothing aspect of stein estimates. *The Annals of Statistics*, pages 887–897, 1984.

D. Lindley. Discussion of the paper by stein. *J. Roy. Statist. Soc. Ser. B*, 24:265–296, 1962.

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

C. N. Morris. Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

C. N. Morris, M. Lysy, et al. Shrinkage estimation in multilevel normal models. *Statistical Science*, 27(1):115–134, 2012.

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

H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems, 1951. URL `http://projecteuclid.org/euclid.bsmsp/1200500224`.

J. E. Rolph. Choosing shrinkage estimators for regression problems. *Communications in Statistics-Theory and Methods*, 5(9):789–802, 1976.

S. L. Sclove. Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, 63(322):596–606, 1968.

S. Searle. Estimable functions and testable hypotheses in linear models. Technical Report BU-213-M, Cornell University, Biometrics Unit, April 1966.

S. Searle. *Linear Models for Unbalanced Data*. Wiley Series in Probability and Statistics. Wiley, 2006. ISBN 9780470040041. URL `http://books.google.com/books?id=PmhUAAAACAAJ`.

S. R. Searle and C. E. McCulloch. *Generalized, linear and mixed models*. Wiley, 2001.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

C. Stein. An approach to the recovery of interblock information in balanced incomplete block designs. *Research paper in statistics: Festschrift for J. Neyman*, pages 351–366, 1966.

C. M. Stein. Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 265–296, 1962.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Proceedings of the Prague Symposium Asymptotic Statistics*, page 345381, 1973.

W. E. Strawderman. Proper bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, pages 385–388, 1971.

W. E. Strawderman. Minimax adaptive generalized ridge regression estimators. *Journal of the American Statistical Association*, 73(363):623–627, 1978.

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

Z. Tan. Improved minimax estimation of a multivariate normal mean under heteroscedasticity. *Bernoulli*, 21(1):574–603, 02 2015. doi: 10.3150/13-BEJ580. URL `http://dx.doi.org/10.3150/13-BEJ580`.

J. R. Thompson. Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, 63(321):113–122, 1968.

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

X. Xie, S. Kou, and L. D. Brown. Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. *Preprint*, 2015.

F. Yates. The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 29(185):51–66, 1934.

C.-H. Zhang. Empirical bayes and compound estimation of normal means. *Statistica Sinica*, 7(1):181–193, 1997.

C.-H. Zhang. Compound decision theory and empirical bayes methods: invited paper. *The Annals of Statistics*, 31(2):379–390, 2003.