# SURE Estimates for a Heteroscedastic Hierarchical Model

Xianchao XIE, S. C. KOU, and Lawrence D. BROWN

Hierarchical models are extensively studied and widely used in statistics and many other scientific areas. They provide an effective tool for combining information from similar resources and achieving partial pooling of inference. Since the seminal work by James and Stein (1961) and Stein (1962), shrinkage estimation has become one major focus for hierarchical models. For the homoscedastic normal model, it is well known that shrinkage estimators, especially the James-Stein estimator, have good risk properties. The heteroscedastic model, though more appropriate for practical applications, is less well studied, and it is unclear what types of shrinkage estimators are superior in terms of the risk. We propose in this article a class of shrinkage estimators based on Stein's unbiased estimate of risk (SURE). We study asymptotic properties of various common estimators as the number of means to be estimated grows ($p \to \infty$). We establish the asymptotic optimality property for the SURE estimators. We then extend our construction to create a class of semiparametric shrinkage estimators and establish corresponding asymptotic optimality results. We emphasize that though the form of our SURE estimators is partially obtained through a normal model at the sampling level, their optimality properties do not heavily depend on such distributional assumptions. We apply the methods to two real datasets and obtain encouraging results.

KEY WORDS: Asymptotic optimality; Heteroscedasticity; Shrinkage estimator; Stein's unbiased risk estimate (SURE).



## 1. INTRODUCTION

Hierarchical modeling has become an increasingly important statistical method in many scientific and engineering applications. It provides an effective tool to combine information and achieve partial pooling of inference. The application of hierarchical models usually involves simultaneous inference of some quantities of interest for different yet similar groups of populations. The earliest study of such problems in statistics is perhaps the simultaneous estimation of several normal means. Since the seminal work by James and Stein (1961), shrinkage estimation has been influential in the development of hierarchical normal models. Stein (1962) described a hierarchical, empirical Bayes interpretation for this estimator (see also Lindley 1962). Efron and Morris (1973) further developed this empirical Bayes interpretation and proposed several competing parametric empirical Bayes estimators. A full Bayesian treatment of this problem can be found in the article by Berger and Strawderman (1996). Recently, Brown and Greenshtein (2009) proposed a nonparametric empirical Bayes method.

There has been substantial research toward understanding the risk properties of shrinkage estimators for the homoscedastic hierarchical normal models (i.e., all the variances in the subpopulations are equal). Baranchik (1970) gave a general form of admissible minimax estimators. Strawderman (1971) studied a class of proper Bayes minimax estimators. Brown (1971) gave a sufficient condition for admissibility of generalized Bayes estimators. The use of loss other than the usual quadratic one is discussed by Brown (1975) and Berger (1976). The heteroscedastic case (i.e., the unequal variance case), on the other hand, is less well addressed, though it is more practical for real applications. Typical minimax estimators, like the one given by Hudson (1974) and Berger (1976), usually shrink the coordinates with lower variances more than those with higher ones, as opposed to the common intuition that more shrinkage should be applied to components with higher variance. The estimators considered in this article do not exhibit this counter-intuitive behavior.

For real-world applications, parametric empirical Bayes estimators (Efron and Morris 1975; Morris 1983) are widely adopted. The application of parametric empirical Bayes models usually involves the specification of a second-level model and the estimation of the corresponding hyper-parameters. For example, for the normal case, the common practice is to choose the normal-normal hierarchical structure and estimate the hyperparameters through empirical Bayes maximum likelihood estimator (EBMLE) or empirical Bayes method of moments (EBMOM). There has also been substantial study on the application of hierarchical Bayes models and nonparametric empirical Bayes methods. Brown (2008) evaluated the performance of various shrinkage estimators using the data on batting average for Major League Baseball players over a single season. It was noted that the parametric empirical Bayes maximum likelihood and the hierarchical Bayes method tend to have a poor performance due to their heavy reliance on the parametric assumptions of the second-level model. Other methods like the EBMOM and nonparametric empirical Bayes method were shown to achieve a better performance. Motivated from such an empirical study, it is hence interesting to know whether it is possible to formally compare those different shrinkage estimators and identify the "optimal" shrinkage estimator.

For this purpose, we propose a class of shrinkage estimators that can be readily applied in the heteroscedastic hierarchical normal models. We name our shrinkage estimators SURE shrinkage estimators, since the method is inspired by Stein's unbiased risk estimate (SURE; Stein 1973, 1981). We first focus

Xianchao Xie is at Harvard University, Cambridge, MA 02138 (E-mail: *xie1981@gmail.com*). S. C. Kou is Professor of Statistics, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *kou@stat.harvard.edu*). Lawrence D. Brown is Professor of Statistics at the University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *lbrown@wharton.upenn.edu*). S. C. Kous' research is supported in part by NIH/NIGMS grant R01GM090202 and NSF grant DMS-0449204. L. Brown's research is supported in part by NSF grant DMS-1007657. The authors thank Professor Philippe Rigollet at Princeton University for helpful discussion.


Journal of the American Statistical Association
December 2012, Vol. 107, No. 500, Theory and Methods
DOI: 10.1080/01621459.2012.728154

on shrinkage estimators whose forms are derived from the classic normal–normal hierarchical model and show that our SURE shrinkage estimators possess asymptotic optimality properties within this (sub)class. The results are then generalized to a class of semiparametric shrinkage estimators that only require the shrinkage factors to satisfy the intuitive condition that the amount of shrinkage is monotone in the component variance, that is, more shrinkage is applied to a component with higher variance. It is emphasized that this asymptotic optimality property neither depends on the specific distributional assumptions nor requires that the sequence of group means be independent of the group variance, an assumption that is implicit in many of the classical empirical Bayes methods like EBMLE and EBMOM. Therefore, there are scenarios where the SURE estimators *strictly dominate* the classical methods. Simulation studies are presented to compare the performance of the proposed estimators with several other shrinkage estimators. We apply our method to the baseball data analyzed by Brown (2008) and report encouraging results. We also use our method to analyze a housing dataset and note some interesting phenomena when applying these methods.

The remainder of the article is organized as follows: In Section 2, we introduce the basic setup and define the parametric SURE estimators along with a brief discussion of some other competing shrinkage estimators. The case of shrinking toward the origin and toward the grand mean is discussed in detail in Sections 3 and 4. Section 5 considers general parametric SURE estimators, where, in addition to the shrinkage factor, the shrinkage location is also determined by the data. Section 6 introduces a class of semiparametric shrinkage estimators and discusses their optimality properties. We conduct a comprehensive simulation study in Section 7 and apply our method to analyzing two real datasets in Section 8. A brief summary is given in Section 9. The technical proofs are relegated to the Appendix.

## 2. BASIC SETUP

Consider the estimation problem

$$X_i|\theta_i \sim N(\theta_i, A_i), \quad i = 1, 2, \ldots, p, \tag{2.1}$$

where the $X_i$ are independently distributed with known (potentially) distinct variances $A_i$. The classical conjugate hierarchical model puts a prior on $\theta_i$

$$\theta_i \sim N(\mu, \lambda), \quad \text{independently for } i = 1, 2, \ldots, p,$$

where $\lambda$ is an unknown hyperparameter.

In this section and Section 3, we first assume the value of the prior mean as $\mu = 0$. The case of unknown prior mean will be the focus of later sections.

Application of Bayes formula (when $\mu = 0$) gives us

$$\theta_i|X_i \sim N\left(\frac{\lambda}{\lambda + A_i}X_i, \frac{\lambda A_i}{\lambda + A_i}\right), \quad X_i \sim N(0, \lambda + A_i),$$

which leads to the Bayes shrinkage estimator

$$\hat{\theta}_i^{\lambda} = \frac{\lambda}{\lambda + A_i}X_i.$$

The empirical Bayes method tries to estimate the unknown hyperparameter $\lambda$ using the marginal distribution of $\mathbf{X}$

$$f(\mathbf{X}|\lambda, \mathbf{A}) \propto \prod_i (\lambda + A_i)^{-1/2} \exp\left\{-X_i^2/(2(\lambda + A_i))\right\}. \tag{2.2}$$

The EBMLE $\hat{\lambda}_{ML}$, which uniquely maximizes the above marginal MLE, can be obtained as the solution of

$$\sum_i \left[\frac{X_i^2}{(\lambda + A_i)^2} - \frac{1}{\lambda + A_i}\right] = 0, \tag{2.3}$$

whenever this equation has a solution. If Equation (2.3) does not have a solution, that is, it is negative when $\lambda = 0$, $\hat{\lambda}_{ML}$ is then zero. The corresponding EBMLE for $\boldsymbol{\theta}$ is

$$\hat{\theta}_i^{ML} := \hat{\theta}_i^{\hat{\lambda}_{ML}} = \frac{\hat{\lambda}_{ML}}{\hat{\lambda}_{ML} + A_i}X_i.$$

Another estimate based on the marginal distribution (Equation (2.2)) is the moment estimate

$$\hat{\lambda}_{MM} = \frac{1}{p}\sum_{i=1}^{p}\left(X_i^2 - A_i\right),$$

or its positive part

$$\hat{\lambda}_{MM}^{+} = \left(\frac{1}{p}\sum_{i=1}^{p}\left(X_i^2 - A_i\right)\right)^{+}.$$

In the homoscedastic case, where $A_i = A$ for $i = 1, \ldots, p$, we have $\hat{\lambda}_{ML} = \hat{\lambda}_{MM}^{+}$ and

$$\hat{\boldsymbol{\theta}}^{ML} = \hat{\boldsymbol{\theta}}^{MM+} = \left(1 - \frac{pA}{\sum_{i=1}^{p} X_i^2}\right)^{+}\boldsymbol{X}.$$

Hence, in this case these two estimators are closely related to the positive-part James-Stein estimator

$$\hat{\theta}_i^{JS+} = \left(1 - \frac{(p-2)A}{\sum_{i=1}^{p} X_i^2}\right)^{+}X_i.$$

In this article, instead of relying on the marginal distribution of $\mathbf{X}$ to estimate $\lambda$, we consider an alternative perspective. The motivation of our methods comes from Stein's unbiased risk estimate (SURE): under the sum of squared-error loss $l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{p}\sum_i(\hat{\theta}_i - \theta_i)^2$, if one uses the shrinkage estimator $\hat{\theta}_i^{\lambda} = \frac{\lambda}{\lambda + A_i}X_i$ to estimate $\boldsymbol{\theta}$ with a fixed $\lambda$, then an unbiased estimate for its risk

$$R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda}) = E[l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})] = \frac{1}{p}\sum_i \frac{A_i}{(A_i + \lambda)^2}\left(A_i\theta_i^2 + \lambda^2\right) \tag{2.4}$$

is

$$\mathrm{SURE}(\lambda) = \frac{1}{p}\sum_i \left[\left(\frac{A_i}{A_i + \lambda}\right)^2 X_i^2 + \frac{A_i(\lambda - A_i)}{A_i + \lambda}\right]. \tag{2.5}$$

Note that Equation (2.4) is just the usual bias-squared plus variance description of the risk; Equation (2.5) can be derived from Stein's unbiased estimate of the risk or directly from Equation (2.4) since $\theta_i^2 = E(X_i^2) - A_i$. This relationship suggests that we

can estimate $\lambda$ from the data as the minimizer of SURE($\lambda$):

$$\hat{\lambda}_{\text{SURE}} = \arg\min_{\lambda \geq 0} \text{SURE}(\lambda)$$
$$= \arg\min_{\lambda \geq 0} \sum_i \left[ \left( \frac{A_i}{A_i + \lambda} \right)^2 X_i^2 + \frac{A_i(\lambda - A_i)}{A_i + \lambda} \right]. \quad (2.6)$$

Setting $\text{SURE}'(\lambda) = 0$ yields an easily solved expression for $\hat{\lambda}_{\text{SURE}}$ as the solution to

$$\sum_i \left[ \frac{A_i^2}{(A_i + \lambda)^3} X_i^2 - \frac{A_i^2}{(A_i + \lambda)^2} \right] = 0. \quad (2.7)$$

If Equation (2.7) does not have a solution, $\hat{\lambda}_{\text{SURE}}$ is then zero. The corresponding SURE estimate for $\boldsymbol{\theta}$ is

$$\hat{\theta}_i^{\text{SURE}} := \hat{\theta}_i^{\hat{\lambda}_{\text{SURE}}} = \frac{\hat{\lambda}_{\text{SURE}}}{\hat{\lambda}_{\text{SURE}} + A_i} X_i.$$

Again, it is worth pointing out that in the homoscedastic case, the three estimators $\hat{\lambda}_{ML}$, $\hat{\lambda}^+_{MM}$, and $\hat{\lambda}_{\text{SURE}}$ are identical and are closely related to the famous positive-part James-Stein estimator. But once the $A_i$ are not all equal, $\hat{\lambda}_{ML}$, $\hat{\lambda}^+_{MM}$, and $\hat{\lambda}_{\text{SURE}}$ give distinct results.

The idea of minimizing the unbiased estimate of risk to obtain the estimate of tuning parameters has a considerable history in statistics. Li (1985, 1986, 1987) discussed the asymptotic properties of the SURE method and its connection to generalized cross-validation in various scenarios. From a slightly different perspective, Johnstone (1987) discussed the admissibility properties of SURE and some alternative estimates of the risk. Kneip (1994) studied the property of SURE in a class of ordered linear smoothers. Donoho and Johnstone (1995) applied SURE to choose the threshold in their SureShrink method. Cavalier et al. (2002) established a nonasymptotic oracle inequality and used it to study the minimax adaptive results of SURE in some inverse problems. We emphasize that our results differ from the previous ones in that the model under consideration is heteroscedastic and our asymptotic results allow us to directly compare our SURE estimators with other shrinkage estimators. Numerical comparisons in Sections 7 and 8 indicate that our estimators have desirable risk properties relative to a number of other shrinkage estimators.

## 3. RISK PROPERTIES OF THE SURE ESTIMATOR

In this section, we consider the risk properties of the SURE estimator. We show that in the heteroscedastic case the SURE estimator $\hat{\boldsymbol{\theta}}^{\text{SURE}}$ is optimal in an asymptotic sense, whereas it is not necessarily so for the other estimators, including the empirical Bayes ML and MOM estimators.

Our first result concerns how well SURE($\lambda$) approximates $l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})$.

*Theorem 3.1.* Assuming two conditions

(A) $\limsup_{p\to\infty} \frac{1}{p} \sum_{i=1}^p A_i^2 < \infty$,

(B) $\limsup_{p\to\infty} \frac{1}{p} \sum_{i=1}^p A_i \theta_i^2 < \infty$,

we have

$$\sup_{0 \leq \lambda \leq \infty} |\text{SURE}(\lambda) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})| \to 0 \text{ in } L^2 \text{ and in probability,}$$
as $p \to \infty$.

Conditions (A) and (B) are required mainly to facilitate a short proof of the above result. Though it is likely that conditions (A) and (B) can be further relaxed, they do not seem to be particularly restrictive and we thus do not seek the full generality here. Theorem 3.1 shows that the risk estimate SURE($\lambda$) is not only unbiased for $R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})$, but, more importantly, is also *uniformly* close to the actual loss $l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})$. We thus expect that minimizing SURE($\lambda$) would lead to an estimate with competitive performance. To facilitate our discussion of the risk properties of our SURE shrinkage estimator, we next introduce the oracle loss (OL) hyperparameter:

$$\tilde{\lambda}^{OL} = \tilde{\lambda}^{OL}(\boldsymbol{\theta}; X_1, \ldots, X_p) = \arg\min_{\lambda \geq 0} l_p(\theta, \hat{\theta}^{\lambda})$$
$$= \arg\min_{\lambda \geq 0} \frac{1}{p} \sum_{i=1}^p \left( \frac{\lambda}{\lambda + A_i} X_i - \theta_i \right)^2.$$

Correspondingly, we define the OL "estimator" $\tilde{\boldsymbol{\theta}}^{OL}$ as

$$\tilde{\boldsymbol{\theta}}^{OL} = \frac{\tilde{\lambda}^{OL}}{\tilde{\lambda}^{OL} + \boldsymbol{A}} \boldsymbol{X}.$$

Of course, $\tilde{\boldsymbol{\theta}}^{OL}$ is not really an estimator since it depends on the unknown $\boldsymbol{\theta}$ (hence, we use the notation $\tilde{\boldsymbol{\theta}}^{OL}$ rather than $\hat{\boldsymbol{\theta}}^{OL}$). Although not obtainable in practice, $\tilde{\boldsymbol{\theta}}^{OL}$ lays down the theoretical limit that one can ever hope to reach: no estimator within the class of estimators of the form $\hat{\boldsymbol{\theta}}^{\hat{\lambda}} = \frac{\hat{\lambda}}{\hat{\lambda} + \mathbf{A}} \mathbf{X}$ can have smaller achieved loss or risk. The performance of the SURE estimator, interestingly, comes close to the oracle one. The following theorem shows under very mild assumptions that our SURE estimator is asymptotically nearly as good as the oracle loss (OL) estimator.

*Theorem 3.2.* Assume conditions (A) and (B). Then

$$\lim_{p\to\infty} P(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\text{SURE}}) \geq l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL}) + \varepsilon) = 0 \quad \text{for any fixed } \varepsilon > 0.$$

The results in the above theorem and all subsequent ones are for given $A_i$'s and $\theta_i$'s; that is, the probabilities and expectations are evaluated given the sequence of $(\theta_i, A_i)$. We require in Theorem 3.2 that $\varepsilon$ is fixed. As one referee kindly pointed out, the result can be enhanced by letting $\varepsilon$ approach zero at some rate that depends on the sequence of $A_i$'s and $\theta_i$'s. A direct consequence of the preceding theorem is that the SURE estimator has a loss that is asymptotically no larger than that of any other estimator in the general class.

*Corollary 3.1.* Assume conditions (A) and (B). Then for any estimator $\hat{\lambda}_p \geq \mathbf{0}$ and the corresponding $\hat{\boldsymbol{\theta}}^{\hat{\lambda}_p} = \frac{\hat{\lambda}_p}{\hat{\lambda}_p + \mathbf{A}} \mathbf{X}$, we always have

$$\lim_{p\to\infty} P(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\text{SURE}}) \geq l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p}) + \varepsilon) = 0 \quad \text{for any fixed } \varepsilon > 0.$$

Theorem 3.2 shows that the loss of $\hat{\boldsymbol{\theta}}^{\text{SURE}}$ converges in probability to the optimum oracle value $l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL})$. We can actually show that under the same conditions $\hat{\boldsymbol{\theta}}^{\text{SURE}}$ is asymptotically as good as $\tilde{\boldsymbol{\theta}}^{OL}$ in terms of expected loss.

*Theorem 3.3.* Assume conditions (A) and (B). Then

$$\lim_{p\to\infty} [R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\text{SURE}}) - E(l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL}))] = 0.$$

It follows from this theorem that $\hat{\boldsymbol{\theta}}^{\mathrm{SURE}}$ has an asymptotically oracle risk: its risk is asymptotically smaller than (at least no larger than) any other estimator in the class.

*Corollary 3.2.* Assume conditions (A) and (B). Then for any estimator $\hat{\lambda}_p \geq \mathbf{0}$ and the corresponding $\hat{\boldsymbol{\theta}}^{\hat{\lambda}_p} = \frac{\hat{\lambda}_p}{\hat{\lambda}_p + \mathbf{A}}\mathbf{X}$, we always have

$$\limsup_{p\to\infty}[R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) - R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p})] \leq 0.$$

Corollaries 3.1 and 3.2 suggest why $\hat{\boldsymbol{\theta}}^{\mathrm{SURE}}$ is generally better than either $\hat{\boldsymbol{\theta}}^{ML}$ or $\hat{\boldsymbol{\theta}}^{MM+}$ for heteroscedastic problems. Note that $\hat{\boldsymbol{\theta}}^{\mathrm{SURE}}$ is asymptotically as good as the OL estimator. Any other asymptotically optimal estimator must have this same property. Theorem 3.1 indicates that this requires such an oracle estimator to asymptotically agree with $\hat{\boldsymbol{\theta}}^{\mathrm{SURE}}$. But in the heteroscedastic case $\hat{\boldsymbol{\theta}}^{ML}$ and $\hat{\boldsymbol{\theta}}^{MM+}$ satisfy different estimating equations from that of $\hat{\boldsymbol{\theta}}^{\mathrm{SURE}}$, as described in Section 2. Hence, for heteroscedastic problems, neither $\hat{\boldsymbol{\theta}}^{ML}$ nor $\hat{\boldsymbol{\theta}}^{MM+}$ can generally be asymptotically optimal in the class of estimators of the form $\hat{\boldsymbol{\theta}}^{\hat{\lambda}} = \frac{\hat{\lambda}}{\hat{\lambda}+\mathbf{A}}\mathbf{X}$. Sections 7 and 8 will illustrate this point through numerical examples.

## 4. SHRINKAGE TOWARD THE GRAND MEAN

The results in the previous section focus on the shrinkage estimators that shrink toward a preset value (taken to be zero above). In practice, it is often the case that, instead of a preset value, we want to shrink toward the grand mean $\bar{X}$. To use the previous result in this case, one might first center the data by subtracting the grand mean from each sample $X_i$, and then pretend that the resulting $X_i - \bar{X}$ are "independent" with "variance" $A_i$, and, following Equation (2.6), one could minimize

$$\sum_i \left[\left(\frac{A_i}{A_i+\lambda}\right)^2 (X_i - \bar{X})^2 + \frac{A_i(\lambda - A_i)}{A_i+\lambda}\right]$$

to obtain the estimate $\hat{\lambda}'$. The estimate of $\theta_i$ then becomes

$$\hat{\theta}_i' = \frac{\hat{\lambda}'}{A_i+\hat{\lambda}'}X_i + \frac{A_i}{A_i+\hat{\lambda}'}\bar{X}, \tag{4.1}$$

which can be used in practice. However, our previous theoretical results are no longer directly applicable. In particular, the optimality property of the resulting estimator is no longer established, since neither $X_i - \bar{X}$ are independent nor the variances are exactly $A_i$.

Fortunately, similar ideas of using the unbiased risk estimate can still be applied. Consider the shrinkage estimator in the following form

$$\hat{\theta}_i^{\lambda,\bar{X}} = \frac{\lambda}{A_i+\lambda}X_i + \frac{A_i}{A_i+\lambda}\bar{X}.$$

Its risk is given by

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\bar{X}}) = E[l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\bar{X}})] = \frac{1}{p}\sum_{i=1}^p \frac{A_i^2}{(A_i+\lambda)^2}(\theta_i - \bar{\theta}_p)^2 + \frac{1}{p}\sum_{i=1}^p \frac{1}{(A_i+\lambda)^2}\left(\lambda^2 A_i + \frac{1}{p}A_i^2(\bar{A}_p + 2\lambda)\right),$$

where

$$\bar{A}_p = \frac{1}{p}\sum_{i=1}^p A_i, \qquad \bar{\theta}_p = \frac{1}{p}\sum_{i=1}^p \theta_i.$$

An unbiased risk estimate is

$$\mathrm{SURE}^G(\lambda) = \frac{1}{p}\sum_{i=1}^p \frac{A_i^2}{(A_i+\lambda)^2}(X_i - \bar{X})^2 + \frac{1}{p}\sum_{i=1}^p \frac{A_i}{A_i+\lambda}\left(\lambda - A_i + \frac{2}{p}A_i\right),$$

that is,

$$E[\mathrm{SURE}^G(\lambda)] = R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\bar{X}}).$$

Minimizing $\mathrm{SURE}^G(\lambda)$ then leads to the grand-mean shrinkage estimator

$$\hat{\theta}_i^G = \frac{\hat{\lambda}_G}{A_i+\hat{\lambda}_G}X_i + \frac{A_i}{A_i+\hat{\lambda}_G}\bar{X}, \tag{4.2}$$

where

$$\hat{\lambda}_G = \arg\min_{\lambda\geq 0} \mathrm{SURE}^G(\lambda).$$

Since this estimate is inspired by the Stein's risk identity, we still call it the SURE estimate. Parallel to the results in the previous section, the grand-mean SURE estimator also possesses asymptotic optimality properties. First, we have the following theorem, which tells us that $\mathrm{SURE}^G(\lambda)$ is uniformly close to the achieved loss $l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\bar{X}})$. Thus, one expects that minimizing $\mathrm{SURE}^G(\lambda)$ would lead to a competitive estimate.

*Theorem 4.1.* Assume conditions (A), (B), and (C) $\limsup_{p\to\infty} \frac{1}{p}\sum_{i=1}^p \theta_i < \infty$.
Then

$$\sup_{0\leq\lambda\leq\infty} |\mathrm{SURE}^G(\lambda) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\bar{X}})| \to 0$$

in $L^1$ and in probability, as $p \to \infty$.

To establish the asymptotic optimality of our SURE estimator, similar to Section 3, we define the grand-mean OL "estimator" $\tilde{\boldsymbol{\theta}}^{GOL}$ as

$$\tilde{\boldsymbol{\theta}}^{GOL} = \frac{\tilde{\lambda}^{GOL}}{\tilde{\lambda}^{GOL}+\boldsymbol{A}}\boldsymbol{X} + \frac{\boldsymbol{A}}{\tilde{\lambda}^{GOL}+\boldsymbol{A}}\bar{X},$$

where

$$\tilde{\lambda}^{GOL} = \arg\min_{\lambda\geq 0} l_p(\theta, \hat{\theta}^{\lambda,\bar{X}}) = \arg\min_{\lambda\geq 0} \frac{1}{p}\sum_{i=1}^p \left(\frac{\lambda}{\lambda+A_i}X_i + \frac{A_i}{\lambda+A_i}\bar{X} - \theta_i\right)^2.$$

No estimator within the class of estimators of the form $\hat{\boldsymbol{\theta}}^{\hat{\lambda},\bar{X}}$, $\hat{\lambda} = \hat{\lambda}(X_1, \ldots, X_p)$, can have smaller achieved loss or risk than $\tilde{\boldsymbol{\theta}}^{GOL}$. However, the performance of the SURE estimator comes close: under very mild assumptions our SURE estimator $\hat{\boldsymbol{\theta}}^G$ is asymptotically nearly as good as the grand-mean OL estimator, as shown in the next theorem.

*Theorem 4.2.* Assume conditions (A)–(C). Then

$$\lim_{p\to\infty} P(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^G) \geq l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{GOL}) + \varepsilon) = 0 \quad \text{for any fixed } \varepsilon > 0.$$

Theorem 4.2 implies that the SURE estimator is asymptotically optimal:

*Corollary 4.1.* Assume conditions (A)–(C). Then for any estimator $\hat{\lambda}_p \geq 0$ and the corresponding $\hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\bar{X}} = \frac{\hat{\lambda}_p}{\hat{\lambda}_p+\mathbf{A}}\mathbf{X} + \frac{\mathbf{A}}{\hat{\lambda}_p+\mathbf{A}}\bar{X}$, we always have

$$\lim_{p\to\infty} P(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^G) \geq l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\bar{X}}) + \varepsilon) = 0 \quad \text{for any fixed } \varepsilon > 0.$$

Theorem 4.2 and Corollary 4.1 compare the estimators in term of the loss. Under the same mild assumptions we can show that the comparison can be extended to the expected loss.

*Theorem 4.3.* Assume conditions (A)–(C). Then

$$\lim_{p\to\infty} \left[ R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^G) - E\left(l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{GOL})\right) \right] = 0.$$

*Corollary 4.2.* Assume conditions (A)–(C). Then for any estimator $\hat{\lambda}_p \geq 0$ and the corresponding $\hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\bar{X}} = \frac{\hat{\lambda}_p}{\hat{\lambda}_p+\mathbf{A}}\mathbf{X} + \frac{\mathbf{A}}{\hat{\lambda}_p+\mathbf{A}}\bar{X}$, we always have

$$\limsup_{p\to\infty} \left[ R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^G) - R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\bar{X}}) \right] \leq 0.$$

Therefore, in general, the SURE estimator is asymptotically better than (or at least as good as) any estimator, including the empirical Bayes ones, for heteroscedastic problems.

## 5. SHRINKAGE TOWARD A GENERAL DATA DRIVEN LOCATION

Instead of shrinking toward the origin or the grand mean, one might let the data determine where to shrink to. Specifically, we can consider the estimator in the form of

$$\hat{\theta}_i^{\lambda,\mu} = \frac{\lambda}{A_i+\lambda} X_i + \frac{A_i}{A_i+\lambda}\mu.$$

Its risk is

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu}) = \frac{1}{p}\sum_i \frac{A_i}{(A_i+\lambda)^2}\left(A_i(\theta_i-\mu)^2 + \lambda^2\right),$$

for which an unbiased estimate is

$$\mathrm{SURE}^M(\lambda,\mu) = \frac{1}{p}\sum_i \frac{A_i}{(A_i+\lambda)^2}\left(A_i(X_i-\mu)^2 + \lambda^2 - A_i^2\right).$$

We can then estimate both $\mu$ and $\lambda$ by minimizing $\mathrm{SURE}^M(\lambda,\mu)$ to obtain

$$\hat{\theta}_i^M = \frac{\hat{\lambda}_M}{A_i+\hat{\lambda}_M} X_i + \frac{A_i}{A_i+\hat{\lambda}_M}\hat{\mu}_M, \tag{5.1}$$

where

$$(\hat{\lambda}_M, \hat{\mu}_M) = \arg\min_{\lambda\geq 0,\,\mu} \mathrm{SURE}^M(\lambda,\mu).$$

As before, we expect the SURE estimator $\hat{\boldsymbol{\theta}}^M$ to possess asymptotic optimality properties. The following theorem, parallel to Theorems 3.1 and 4.1, tells us that $\mathrm{SURE}^M(\lambda,\mu)$ closely approximates $l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu})$ in a uniform fashion. Thus, one expects that minimizing $\mathrm{SURE}^M(\lambda,\mu)$ would again lead to a competitive estimate.

*Theorem 5.1.* Assume conditions (A), (B), and (C′) $\limsup_{p\to\infty} \frac{1}{p}\sum_{i=1}^p |\theta_i|^{2+\delta} < \infty$ for some $\delta > 0$.

Then we have

$$\sup_{0\leq\lambda\leq\infty,\ |\mu|\leq\max_i|X_i|} \left|\mathrm{SURE}^M(\lambda,\mu) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu})\right| \to 0$$

in $L^1$ and in probability, as $p \to \infty$.

Note that condition (C′) assumes that the $(2+\delta)$th moment of $\theta$ is bounded; it is slightly stronger than condition (C). Note also that Theorem 5.1 restricts the shrinkage location $\mu$ to be within $[-\max_i|X_i|, \max_i|X_i|]$. This assumption is included for technical reasons to ease the proof in the Appendix. In practice, it is harmless since no sensible shrinkage estimator would attempt to shrink toward a location that lies outside the range of the data.

Next, parallel to the development of Sections 3 and 4, we define the general-mean OL "estimator" $\tilde{\boldsymbol{\theta}}^{MOL}$ as

$$\tilde{\boldsymbol{\theta}}^{MOL} = \frac{\tilde{\lambda}^{MOL}}{\tilde{\lambda}^{MOL}+\boldsymbol{A}}\boldsymbol{X} + \frac{\boldsymbol{A}}{\tilde{\lambda}^{MOL}+\boldsymbol{A}}\tilde{\mu}^{MOL}$$

where

$$\begin{aligned}\left[\tilde{\lambda}^{MOL}, \tilde{\mu}^{MOL}\right] &= \arg\min_{\lambda\geq 0,\ |\mu|\leq\max_i|X_i|} l_p(\theta, \hat{\theta}^{\lambda,\mu}) \\ &= \arg\min_{\lambda\geq 0, |\mu|\leq\max_i|X_i|} \left\| \frac{\lambda}{\lambda+\boldsymbol{A}}\boldsymbol{X} + \frac{\boldsymbol{A}}{\lambda+\boldsymbol{A}}\mu - \boldsymbol{\theta}\right\|^2.\end{aligned}$$

The next theorem and corollary show that the SURE estimator $\hat{\boldsymbol{\theta}}^M$ is asymptotically nearly as good as the general-mean OL estimator, and, consequently, it is asymptotically better than (or at least as good as) any other shrinkage estimator in terms of the achieved loss.

*Theorem 5.2.* Assume conditions (A), (B), and (C′). Then

$$\lim_{p\to\infty} P\left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^M) \geq l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{MOL}) + \varepsilon\right) = 0 \quad \text{for any fixed } \varepsilon > 0.$$

*Corollary 5.1.* Assume conditions (A), (B), and (C′). Then for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\hat{\mu}_p} = \frac{\hat{\lambda}_p}{\hat{\lambda}_p+\mathbf{A}}\mathbf{X} + \frac{\mathbf{A}}{\hat{\lambda}_p+\mathbf{A}}\hat{\mu}_p$ with $\hat{\lambda}_p \geq 0$ and $|\hat{\mu}_p| \leq \max_i|X_i|$, we have

$$\lim_{p\to\infty} P\left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^M) \geq l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\hat{\mu}_p}) + \varepsilon\right) = 0 \quad \text{for any fixed } \varepsilon > 0.$$

Under the same mild assumptions, the comparison of the estimators can be extended to the expected loss as well.

*Theorem 5.3.* Assume conditions (A), (B), and (C′). Then

$$\lim_{p\to\infty} \left[ R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^M) - E(l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{MOL})) \right] = 0.$$

*Corollary 5.2.* Assume conditions (A), (B), and (C′). Then for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\hat{\mu}_p} = \frac{\hat{\lambda}_p}{\hat{\lambda}_p+\mathbf{A}}\mathbf{X} + \frac{\mathbf{A}}{\hat{\lambda}_p+\mathbf{A}}\hat{\mu}_p$ with $\hat{\lambda}_p \geq 0$ and $|\hat{\mu}_p| \leq \max_i|X_i|$, we have

$$\limsup_{p\to\infty} \left[ R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^M) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\lambda}_p,\hat{\mu}_p}) \right] \leq 0.$$

Theorems 5.2 and 5.3 tell us that the SURE estimator is asymptotically optimal: it has the smallest loss and risk among all shrinkage estimators of the form $\hat{\boldsymbol{\theta}}^{\hat{\lambda},\hat{\mu}} = \frac{\hat{\lambda}}{\hat{\lambda}+\mathbf{A}}\mathbf{X} + \frac{\mathbf{A}}{\hat{\lambda}+\mathbf{A}}\hat{\mu}$. A

special case is the comparison between the general-mean SURE estimator $\hat{\boldsymbol{\theta}}^M$ and the grand-mean shrinkage estimator $\hat{\boldsymbol{\theta}}^G$.

*Corollary 5.3.* Assume conditions (A), (B), and (C′). Then

$$\limsup_{p\to\infty}\left[R(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^M)-R(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^G)\right]\le 0.$$

In other words, $\hat{\boldsymbol{\theta}}^M$ asymptotically outperforms $\hat{\boldsymbol{\theta}}^G$. This result provides a theoretical underpinning of the empirical result of Section 7, where we shall see that $\hat{\boldsymbol{\theta}}^M$ encountered a smaller loss than $\hat{\boldsymbol{\theta}}^G$.

Another possible variation of the SURE estimate is to consider the weighted loss function $l_w(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})=\sum_i w_i(\hat{\theta}_i-\theta_i)^2$. An unbiased risk estimate for $\hat{\boldsymbol{\theta}}^{\lambda,\mu}$ in this case is

$$\mathrm{SURE}^W(\mu,\lambda)=\frac{1}{p}\sum_i\frac{w_iA_i}{(A_i+\lambda)^2}\left(A_i(X_i-\mu)^2+\lambda^2-A_i^2\right).$$

The theoretical properties (such as the optimality) of the resulting estimator would be an interesting question worth further investigation.

Note that when we take $w_i\propto 1/A_i$, the $\mathrm{SURE}^W(\mu,\lambda)$ can be rewritten as

$$\begin{aligned}\mathrm{SURE}^W(\mu,\lambda)&=\frac{1}{p}\sum_{i=1}^p\left(\frac{\left(\hat{\theta}_i^{\mu,\lambda}-X_i\right)^2}{A_i}+2\cdot df_i-1\right)\\&=\frac{1}{p}\sum_{i=1}^p\left[\left(\hat{\theta}_i^{\mu,\lambda}-X_i\right)^2/A_i+\frac{2}{A_i}\hat{\mathrm{cov}}\left(\theta_i^{\mu,\lambda},X_i\right)-1\right],\end{aligned}$$

where

$$\hat{\mathrm{cov}}\left(\hat{\theta}_i^{\mu,\lambda},X_i\right)=\frac{\lambda A_i}{A_i+\lambda}$$

is the unbiased estimate of the covariance penalty (Efron 1986, 2004). The SURE criterion in this case coincides with Mallows' $C_p$ (Mallows 1973), or equivalently the AIC (Akaike 1973), where the number of parameters is taken to be the generalized degree of freedom (Ye 1998). The above results thus serve as a rigorous confirmation of the belief that AIC-type criteria usually lead to models that enjoy good risk properties.

*Remark.* In the discussion above we have assumed that at the sampling level, the model is normal: $X_i|\theta_i\sim N(\theta_i,A_i)$. It is noted here that such a distributional assumption is actually not necessary. With some minimum regularity conditions (such as the tail of the distribution does not decay too slowly), all the theorems and corollaries will remain valid. One assumption that we do make is that the variances are known or can be estimated independently.

## 6. SEMIPARAMETRIC SURE SHRINKAGE ESTIMATION

As we noted in the previous sections, the optimality properties of the SURE estimators do not depend on the hypothetical normal prior. However, the general form $\hat{\theta}_i=\frac{\lambda}{A_i+\lambda}X_i+\frac{A_i}{A_i+\lambda}\mu$ of the shrinkage estimators studied in the preceding section is indeed motivated from the normal prior. In this section, we consider a larger class of shrinkage estimators, generalize the SURE estimator in this larger setting, and study its asymptotic optimality properties. This new class of shrinkage estimators enjoys a more flexible form. We shall see that the generalized SURE estimator is optimal among this larger class of shrinkage estimators. Because it is optimal within a larger class of estimators, it automatically performs asymptotically at least as well as the SURE estimators in previous sections. There are circumstances in which it can strictly outperform those estimators, as explored in Sections 7 and 8.

To motivate this larger class of shrinkage estimators, let us consider the hierarchical setting of

$$\begin{aligned}\lambda&\sim\pi(\lambda)\\\theta_i|\lambda;\mu&\overset{\text{iid}}{\sim}N(\mu,\lambda)\\X_i|\theta_i;A_i&\overset{\text{ind}}{\sim}N(\theta_i,A_i),\end{aligned}$$

where $\pi$ is an unspecified hyperprior on $\lambda$. The posterior mean of $\theta_i$ (assuming existence) is

$$E(\theta_i|\mathbf{X})=E\left(\frac{\lambda}{A_i+\lambda}|\mathbf{X}\right)X_i+E\left(\frac{A_i}{A_i+\lambda}|\mathbf{X}\right)\mu. \quad (6.1)$$

We can interpret $E(\frac{\lambda}{A_i+\lambda}|\mathbf{X})$, which is monotonically decreasing in $A_i$, as the shrinkage factor for the $i$th component. This suggests us to consider general shrinkage estimators of the form

$$\hat{\theta}_i^{b_i,\mu}=(1-b_i)\cdot X_i+b_i\cdot\mu,$$

where $b_i\in[0,1]$, and in this general form we no longer require $b_i$ to assume any parametric form: there is no hyperparameter $\lambda$. Clearly, without putting any constraint on the $b_i$'s, one expects that the resulting SURE shrinkage estimator may suffer from problems such as overfitting. One natural way to prevent this from happening is to require the following condition on the shrinkage factors

$$\text{Requirement (MON)}: b_i\le b_j \quad \text{for any } i \text{ and } j \text{ such that } A_i\le A_j,$$

or equivalently $b_i$ is nondecreasing in $A_i$. In other words, the larger the variance is, the stronger is the shrinkage. This requirement is quite intuitive, especially in light of Equation (6.1). Note that this requirement is satisfied by all the previous parametric SURE estimators.

To derive our semiparametric shrinkage estimator, we first observe that an unbiased risk estimate of $\hat{\boldsymbol{\theta}}^{\mathbf{b},\mu}$ is

$$\mathrm{SURE}^M(\mathbf{b},\mu)=\frac{1}{p}\sum_{i=1}^p\left[b_i^2(X_i-\mu)^2+(1-2b_i)A_i\right].$$

Minimizing the SURE with respect to $(\mathbf{b},\mu)$ then leads to our semiparametric SURE shrinkage estimator

$$\hat{\theta}_i^{SM}=\left(1-\hat{b}_i^{SM}\right)\cdot X_i+\hat{b}_i^{SM}\cdot\hat{\mu}^{SM}, \quad (6.2)$$

where

$$\begin{aligned}(\hat{\mathbf{b}}^{SM},\hat{\mu}^{SM})=&\text{ minimizer of } \mathrm{SURE}^M(\mathbf{b},\mu)\\&\text{subject to } b_i\in[0,1] \text{ and Requirement (MON).}\end{aligned}$$

Parallel to the parametric case, we can also consider the estimator that shrinks toward the grand mean, that is,

$$\hat{\theta}_i^{b_i,\bar{X}} = (1 - b_i) \cdot X_i + b_i \cdot \bar{X}.$$

An unbiased estimate of its risk is

$$\text{SURE}^G(\mathbf{b}) = \frac{1}{p} \sum_{i=1}^{p} \left[ b_i^2 (X_i - \bar{X})^2 + \left(1 - 2\left(1 - \frac{1}{p}\right) b_i\right) A_i \right].$$

Minimizing the $\text{SURE}^G$ with respect to $\mathbf{b}$ then leads to our semiparametric SURE grand-mean shrinkage estimator

$$\hat{\theta}_i^{SG} = \left(1 - \hat{b}_i^{SG}\right) \cdot X_i + \hat{b}_i^{SG} \cdot \bar{X}, \tag{6.3}$$

where

$$\hat{\mathbf{b}}^{SG} = \text{minimizer of } \text{SURE}^G(\mathbf{b}) \quad \text{subject to } b_i \in [0, 1] \text{ and Requirement (MON).}$$

It is emphasized that even though we used Equation (6.1) to motivate our methods, we do not actually impose any particular parametric form on our estimates of the shrinkage factor $b_i$ other than the range and the monotonicity requirement. This is the reason we term our methods "semiparametric." The theoretical properties of the semiparametric SURE shrinkage estimators are summarized as follows. To save space, we only discuss the asymptotic optimality of the general-mean SURE estimator $\hat{\theta}_i^{SM}$ below; the asymptotic property of $\hat{\theta}_i^{SG}$ can be similarly studied.

*Theorem 6.1.* Assuming conditions (A), (B), and (C′), we have

$$\sup \left| \text{SURE}^M(\boldsymbol{b}, \mu) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\mu}) \right| \to 0$$

in $L^1$ and in probability, as $p \to \infty$, where the supremum is taken over $b_i \in [0, 1]$, $|\mu| \le \max_i |X_i|$ and Requirement (MON).

*Theorem 6.2.* Assume conditions (A), (B), and (C′). Then for any shrinkage estimator $\hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}}_p,\hat{\mu}_p} = (1 - \hat{\mathbf{b}}_p) \cdot \mathbf{X} + \hat{\mathbf{b}}_p \cdot \hat{\mu}_p$, where $\hat{\mathbf{b}}_p \in [0, 1]$ satisfies Requirement (MON) and $|\hat{\mu}_p| \le \max_i |X_i|$, we have

$$\lim_{p\to\infty} P\left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{SM}) \ge l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}}_p,\hat{\mu}_p}) + \varepsilon\right) = 0 \quad \text{for any fixed } \varepsilon > 0,$$

and

$$\limsup_{p\to\infty} \left[ R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{SM}) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\mathbf{b}}_p,\hat{\mu}_p}) \right] \le 0.$$

Theorem 6.2 shows that our semiparametric SURE shrinkage estimator is optimal among the class of shrinkage estimators whose shrinkage factor is a nondecreasing function of the variance. In particular, the semiparametric SURE shrinkage estimator is asymptotically superior than (at least no worse than) any hierarchical empirical Bayes estimator.

## 7. SIMULATION STUDY

In this section, we conduct a number of simulations to study the performance of the SURE estimators. We consider $\hat{\boldsymbol{\theta}}^G$, $\hat{\boldsymbol{\theta}}^M$ (see Equations (4.2) and (5.1)) and the two semiparametric shrinkage estimators $\hat{\boldsymbol{\theta}}^{SG}$, $\hat{\boldsymbol{\theta}}^{SM}$ (Equations (6.3) and (6.2)) and compare their performance with that of the EBMLE estimator $\hat{\boldsymbol{\theta}}^{ML}$, the EBMOM estimator $\hat{\boldsymbol{\theta}}^{MM}$, and an extension of the James-Stein estimator $\hat{\theta}_i^{JS+}$. The EBMLE estimator used here is given by

$$\hat{\theta}_i^{ML} := \hat{\theta}_i^{\hat{\lambda}_{ML}} = \frac{\hat{\lambda}_{ML}}{\hat{\lambda}_{ML} + A_i} X_i + \frac{A_i}{\hat{\lambda}_{ML} + A_i} \hat{\mu}_{ML}, \tag{7.1}$$

where $\hat{\lambda}_{ML}$ and $\hat{\mu}_{ML}$ are obtained by maximizing the marginal density

$$f(\mathbf{X}|\lambda, \mathbf{A}) \propto \prod_i (\lambda + A_i)^{-1/2} \exp\{-(X_i - \mu)^2/(2(\lambda + A_i))\},$$

and the EBMOM estimator is given by

$$\hat{\theta}_i^{MM} := \hat{\theta}_i^{\hat{\lambda}_{MM}} = \frac{\hat{\lambda}_{MM}}{\hat{\lambda}_{MM} + A_i} X_i + \frac{A_i}{\hat{\lambda}_{MM} + A_i} \hat{\mu}_{MM}, \tag{7.2}$$

where $\hat{\lambda}_{MM}$ and $\hat{\mu}_{MM}$ are obtained as the root of the following equations

$$\mu = \frac{\sum_i X_i/(A_i + \lambda)}{\sum_i 1/(A_i + \lambda)},$$

$$\lambda = \frac{1}{p - 1} \left( \sum_i (X_i - \mu)^2 - (p - 1)/p \sum_i A_i \right)^+.$$

The extended James-Stein estimator is

$$\hat{\theta}_i^{JS+} := \hat{\mu}_{JS+} + \left(1 - \frac{p - 3}{\sum_i (X_i - \hat{\mu}_{JS})^2/A_i}\right)^+ (X_i - \hat{\mu}_{JS+}),$$

$$\hat{\mu}_{JS+} = \frac{\sum_i X_i/A_i}{\sum_i 1/A_i}, \tag{7.3}$$

which has been discussed by Brown (2008).

For each simulation, we first draw $(A_i, \theta_i)$ $(i = 1, \ldots, p)$ independently from a distribution $\pi(A, \theta)$ and then draw $X_i$ given $(A_i, \theta_i)$. The shrinkage estimators are then found via the formulas described above. This process is repeated a large number of times ($N = 100{,}000$) to obtain an accurate estimate of the average risk for each estimator. The sample size $p$ is chosen to vary from 20 to 500 at an interval of length 20.

In each example, we also calculate the oracle *risk* "estimator" $\tilde{\boldsymbol{\theta}}^{OR}$, defined as

$$\tilde{\boldsymbol{\theta}}^{OR} = \frac{\tilde{\lambda}^{OR}}{\tilde{\lambda}^{OR} + \boldsymbol{A}} \boldsymbol{X} + \frac{\boldsymbol{A}}{\tilde{\lambda}^{OR} + \boldsymbol{A}} \tilde{\mu}^{OR},$$

where

$$(\tilde{\lambda}^{OR}, \tilde{\mu}^{OR}) = \arg \min_{\lambda \ge 0,\, \mu} R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu})$$

$$= \arg \min_{\lambda \ge 0,\, \mu} \sum_{i=1}^{p} \frac{1}{p} E\left[ \left( \frac{\lambda}{\lambda + A_i} X_i + \frac{A_i}{\lambda + A_i} \mu - \theta_i \right)^2 \right].$$

Similar to the OL estimators, the oracle risk estimator $\tilde{\boldsymbol{\theta}}^{OR}$ cannot be used without the knowledge of $\boldsymbol{\theta}$, but it does provide a sensible lower bound of the risk achievable by any shrinkage

Table 1. The limiting risk $\lim_{p\to\infty} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ of different shrinkage estimators. The six columns (1)–(6) correspond to the six simulation examples

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| EBMLE | 0.3357 | 0.0697 | 0.0775 | 0.0057 | 0.2470 | 0.0775 |
| EBMOM | 0.3357 | 0.0697 | 0.0755 | 0.0058 | 0.2434 | 0.0755 |
| J-S | 0.3632 | 0.0737 | 0.0797 | 0.0056 | 0.2594 | 0.0797 |
| Oracle | 0.3357 | 0.0697 | 0.0540 | 0.0051 | 0.1947 | 0.0540 |
| SURE $\hat{\theta}^G$ | 0.3357 | 0.0697 | 0.0553 | 0.0051 | 0.2337 | 0.0553 |
| SURE $\hat{\theta}^M$ | 0.3357 | 0.0697 | 0.0540 | 0.0051 | 0.1947 | 0.0540 |
| SURE $\hat{\theta}^{SG}$ | 0.3357 | 0.0697 | 0.0523 | 0.0050 | 0.2335 | 0.0523 |
| SURE $\hat{\theta}^{SM}$ | 0.3357 | 0.0697 | 0.0491 | 0.0050 | 0.1739 | 0.0491 |

estimator with the given parametric form. An alternative oracle estimator, which we do not pursue here, is the hierarchical Bayes estimator where the correct hyper-prior is used.

Note that since we have a large number ($N = 100{,}000$) of repetitions in our simulation, the averaged risk of the oracle risk estimator plotted against $p$ will be essentially a flat line. For each shrinkage estimator considered here, the risk $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ will converge to a limit as $p \to \infty$. This limit can be calculated numerically. Table 1 shows these limiting risks for each simulation example.

*Example 7.1.* We draw $(A, \theta, X)$ such that $A \sim \text{Unif}(0.1, 1)$ and $\theta \sim N(0, 1)$ independently, and $X \sim N(\theta, A)$. Note that we draw $A$ from Unif(0.1, 1) instead of from Unif(0, 1) to make sure that the variances $A_i$ are bounded away from 0. The oracle risk estimator $\tilde{\boldsymbol{\theta}}^{OR}$ is found to have $\lambda_0 = 1$ and $\mu_0 = 0$. The corresponding risk for $\tilde{\boldsymbol{\theta}}^{OR}$ is $R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OR}) = 1 - \ln(2/1.1)/0.9 \approx 0.3357$. The plot in Figure 1(a) shows the risks of the seven shrinkage estimators as the sample size $p$ varies. Clearly, the performance of all shrinkage estimators except the extended James-Stein estimator eventually approaches that of the oracle risk estimator. Table 1 confirms the picture. Note that when the sample size is relatively small, the four SURE estimators incur slightly larger risks compared with the two empirical Bayes estimators. This is because the hierarchical distribution on $A$ and $\theta$ is exactly the one assumed by the empirical Bayes estimators; in particular, the EBMLE relies on the parametric normal form of the prior, and the EBMOM estimator assumes independence between $A$ and $\theta$, both of which are satisfied here. The SURE estimators require neither of these conditions but still achieve rather competitive performance. When the sample size is moderately large, all six estimators well approach the limit given by the oracle risk estimator. The extended James-Stein estimator behaves far worse than the others.

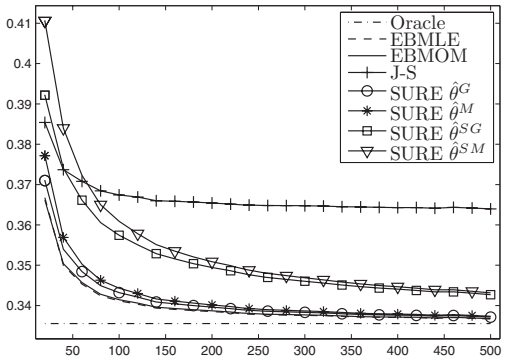
*Example 7.2.* We draw $(A, \theta, X)$ such that $A \sim \text{Unif}(0.1, 1)$ and $\theta \sim \text{Unif}(0, 1)$ independently, and $X \sim N(\theta, A)$. In this example, $\theta$ no longer comes from a normal distribution, but $\theta$ and $A$ are still independent. The oracle risk estimator is found to have $\lambda_0 \approx 0.0834$ and $\mu_0 = 0.5$. The corresponding risk for $\tilde{\boldsymbol{\theta}}^{OR}$ is $R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OR}) \approx 0.0697$. The plot in Figure 1(b) shows the risks of the seven shrinkage estimators as the sample size $p$ varies. Again, as $p$ gets large, the performance of all shrinkage estimators except the extended James-Stein estimator eventually approaches that of the oracle risk estimator, as confirmed by Table 1. This observation indicates that the parametric form of the prior on $\theta$ is not crucial as long as $A$ and $\theta$ are independent.

*Example 7.3.* $(A, \theta, X)$ are drawn such that $A \sim \text{Unif}(0.1, 1), \theta = A$, and $X \sim N(\theta, A)$. In this example, $A$ and $\theta$ are no longer independent of each other. The oracle risk estimator is found to have $\lambda_0 \approx 0.0781$ and $\mu_0 \approx 0.5949$ numerically. The corresponding risk for $\tilde{\boldsymbol{\theta}}^{OR}$ is $R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OR}) \approx 0.0540$. The plot in Figure 1(c) shows the risks of the seven shrinkage estimators as functions of $p$, the sample size. As our theoretical result in Section 5 indicates, the performance of the SURE estimator $\hat{\boldsymbol{\theta}}^M$ approaches that of the oracle risk estimator, which is seen in Figure 1(c). The limiting risks of the SURE grand-mean shrinkage estimator $\hat{\boldsymbol{\theta}}^G$, the two empirical Bayes estimators, and the extended James-Stein estimator, on the other hand, are strictly greater than the risk of the oracle estimator, as shown in Table 1. The main reason for the difference is that $A$ and $\theta$ are no longer independent. It is quite interesting to note from Table 1 that the limiting risks of the two semiparametric shrinkage estimators $\hat{\boldsymbol{\theta}}^{SG}$ and $\hat{\boldsymbol{\theta}}^{SM}$ are actually strictly smaller than the oracle risk (although due to the scale of the plot, it is not easy to spot). The reason for this "better-than-oracle" performance is that the semiparametric estimators are not restricted to the specific parametric family that the oracle estimator assumes.

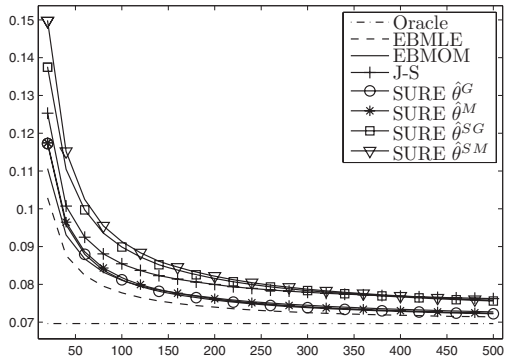
*Example 7.4.* In this example, $(A, \theta, X)$ are drawn as $A \sim \text{Inv-}\chi^2_{10}, \theta = A$ and $X \sim N(\theta, A)$. The inverse chi-square distribution is used here as it is the conjugate distribution for normal variance. The oracle risk estimator is found to have $\lambda_0 \approx 0.0032$ and $\mu_0 \approx 0.1266$ numerically. The corresponding risk for the oracle risk estimator is $R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OR}) \approx 0.0051$. The plot in Figure 1(d) shows the risks of the seven shrinkage estimators as functions of the sample size $p$. We see from this figure and Table 1 that the risks of the SURE estimators $\hat{\boldsymbol{\theta}}^M$ and $\hat{\boldsymbol{\theta}}^G$ approach that of the oracle risk estimator, whereas the limiting risks of the James-Stein and empirical Bayes estimators $\hat{\boldsymbol{\theta}}^{JS+}$, $\hat{\boldsymbol{\theta}}^{ML}$, and $\hat{\boldsymbol{\theta}}^{MM}$ are strictly greater than the oracle risk. Note that the limiting risks of the two semiparametric shrinkage estimators $\hat{\boldsymbol{\theta}}^{SG}$ and $\hat{\boldsymbol{\theta}}^{SM}$ are in fact strictly smaller than the oracle risk (although due to the scale of the plot, it is not easy to spot).

*Example 7.5.* The setting in this example is chosen in such a way that it reflects grouping in the data. We draw $(A, \theta, X)$ as $A \sim \frac{1}{2} \cdot 1_{\{A=0.1\}} + \frac{1}{2} \cdot 1_{\{A=0.5\}}$, (i.e., $A$ is 0.1 or 0.5 with 50% probability each), $\theta | A = 0.1 \sim N(2, 0.1), \theta | A = 0.5 \sim N(0, 0.5)$, and $X | \theta, A \sim N(\theta, A)$ so that there exist two groups in the data. The oracle risk estimator is found to have $\lambda_0 \approx 0.8347$ and $\mu_0 \approx 0.1506$. The corresponding risk for the oracle risk estimator is $R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OR}) \approx 0.1947$. Figure 1(e) plots the risks of the seven shrinkage estimators versus the sample size $p$. We see clearly that the risk of the SURE estimator $\hat{\boldsymbol{\theta}}^M$ approaches that of the oracle risk estimator, whereas the risks of the other four parametric shrinkage estimators ($\hat{\boldsymbol{\theta}}^G$, $\hat{\boldsymbol{\theta}}^{ML}$, $\hat{\boldsymbol{\theta}}^{MM}$, and $\hat{\boldsymbol{\theta}}^{JS+}$) are notably greater than the oracle risk. The semiparametric shrinkage estimator $\hat{\boldsymbol{\theta}}^{SM}$ is seen to achieve an even significant improvement over the oracle one, confirming the results of Section 6.

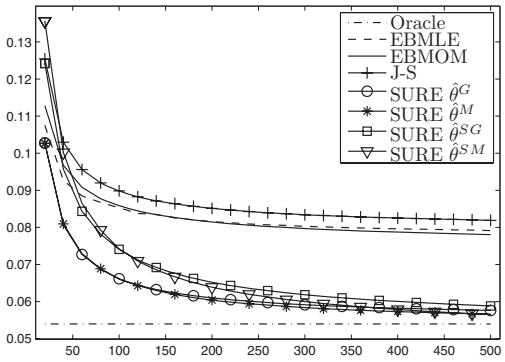
*Example 7.6.* In this example, we allow $X$ to depart from the normal model, that is, $X \nsim N(\theta, A)$, to assess the sensitivity in performance of the estimators to the normality



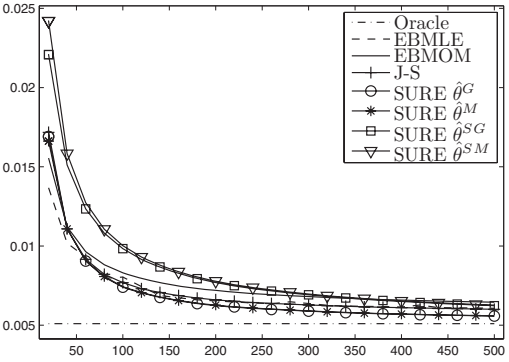

(a) $A \sim \text{Unif}(0.1, 1)$, $\theta \sim N(0, 1)$ independently; $X \sim N(\theta, A)$.



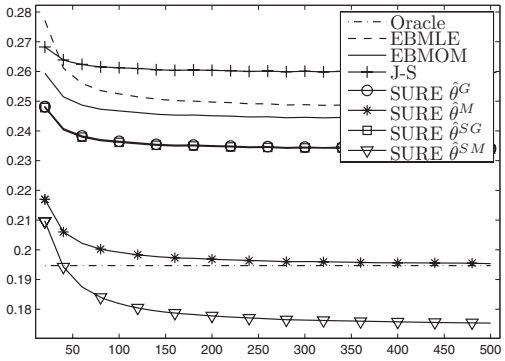

(b) $A \sim \text{Unif}(0.1, 1)$, $\theta \sim \text{Unif}(0, 1)$ independently; $X \sim N(\theta, A)$.



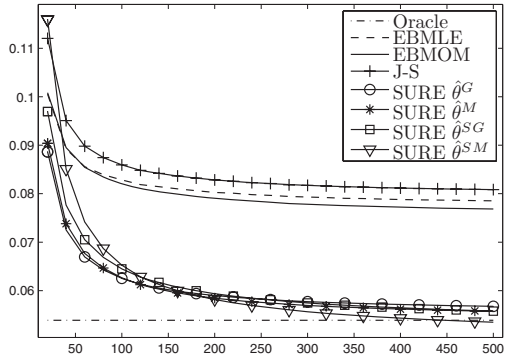

(c) $A \sim \text{Unif}(0.1, 1)$, $\theta = A$; $X \sim N(\theta, A)$.




(d) $A \sim \text{Inv-}\chi^2_{10}$, $\theta = A$; $X \sim N(\theta, A)$.




(e) $A \sim \frac{1}{2} \cdot 1_{\{A=0.1\}} + \frac{1}{2} \cdot 1_{\{A=0.5\}}$, $\theta | A = 0.1 \sim N(2, 0.1)$, $\theta | A = 0.5 \sim N(0, 0.5)$; $X \sim N(\theta, A)$.




(f) $A \sim \text{Unif}(0.1, 1)$, $\theta = A$; $X \sim \text{Unif}[\theta - \sqrt{3A}, \theta + \sqrt{3A}]$.

Figure 1. Comparing the risks of different shrinkage estimators. (a)–(f) correspond to the six simulation examples.

assumption. $(A, \theta, X)$ are drawn as $A \sim \text{Unif}(0.1, 1)$, $\theta = A$, and $X \sim \text{Unif}[\theta - \sqrt{3A}, \theta + \sqrt{3A}]$. The oracle risk estimator is found to have $\lambda_0 \approx 0.0781$ and $\mu_0 \approx 0.5949$ numerically. The corresponding risk for the oracle risk estimator $\tilde{\boldsymbol{\theta}}^{OR}$ is $R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OR}) \approx 0.0540$. Figure 1(f) plots the risks of the seven shrinkage estimators versus the sample size $p$. We see that the performance of SURE estimator $\hat{\boldsymbol{\theta}}^M$ approaches that of the oracle risk estimator, whereas the empirical Bayes estimators $\hat{\boldsymbol{\theta}}^{ML}$and $\hat{\boldsymbol{\theta}}^{MM}$ and the extended James-Stein estimator $\hat{\boldsymbol{\theta}}^{JS+}$ do notably worse. Table 1 shows that the limiting risks of the two semiparametric shrinkage estimators $\hat{\boldsymbol{\theta}}^{SG}$ and $\hat{\boldsymbol{\theta}}^{SM}$ are strictly smaller than the oracle risk (though the gap is not big enough to be easily seen on the plot).

## 8. APPLICATION TO REAL DATA

### 8.1 Prediction of Batting Average

In this section, we apply the SURE estimators to the baseball data by Brown (2008) to assess their effectiveness. The data analyzed here are the batting records for all the Major League

Baseball players in the season of 2005. Like in the article by Brown (2008), we divide the dataset into two half seasons and try to predict the batting average of each player for the second half using the data from the first half. We also carried out the necessary preprocessing steps proposed there. For example, we removed from the analysis the players whose number of at-bats is less than 11. For each player, let the number of at-bats be $N$ and the successful number of batting be $H$; we have,

$$H_{ij} \sim \text{Binomial}(N_{ij}, p_j),$$

where $i = 1, 2$ is the season indicator and $j = 1, \ldots, p$ is the player indicator. As in the article by Brown (2008), the following variance-stablizing transformation is used before applying the shrinkage estimators

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

resulting in

$$X_{ij} \sim N\left(\theta_j, \frac{1}{4N_{ij}}\right), \quad \theta_j = \arcsin(p_j).$$

One error measurement, denoted as TSE, introduced by Brown (2008), is adopted here as the basis of comparison. TSE measures the sum of squared errors in terms of $\theta$ and $X$, the transformed values:

$$\text{TSE}(\hat{\boldsymbol{\theta}}) = \sum_j (X_{2j} - \hat{\theta}_j)^2 - \sum_j \frac{1}{4N_{2j}}.$$

Table 2 summarizes the result, where the shrinkage estimators are applied three times—to all the baseball players, the pitchers only, and the nonpitchers only. The values reported are the ratios of the error of a given estimator to that of the benchmark naive estimator, which simply uses the first half season $X_{1j}$ to predict the second half $X_{2j}$. In the table, EB-MM is the empirical Bayes method-of-moment estimator (7.2). EB-ML is the empirical Bayes maximum likelihood estimator (7.1). James-Stein corresponds to the extended James-Stein estimator (7.3). Since this particular dataset has been widely studied, we also compare our methods with a number of more recently developed methods, including the nonparametric shrinkage methods by Brown and Greenshtein (2009), the binomial mixture model by Muralidharan (2010), and the weighted least squares and general maximum likelihood estimators (with/without the covariate at bats effect) by Jiang and Zhang (2009, 2010). Results for those methods are from Brown (2008), Muralidharan (2010), and Jiang and Zhang (2009, 2010). The last group shows the results for our SURE estimators. SURE $\hat{\boldsymbol{\theta}}^G$ is the SURE grand-mean shrinkage estimator (Equation (4.2)). SURE $\hat{\boldsymbol{\theta}}^M$ is the SURE estimator (Equation (5.1)), where the shrinkage location $\hat{\mu}$ is also determined from the data. The last two estimators are the semiparametric SURE shrinkage estimators. SPSURE $\hat{\boldsymbol{\theta}}^{SG}$ is the semiparametric grand-mean shrinkage estimator (Equation (6.3)); SPSURE $\hat{\boldsymbol{\theta}}^{SM}$ is the semiparametric SURE general-mean estimator (Equation (6.2)).

Table 2. Prediction errors of batting averages

| | ALL | Pitchers | Nonpitchers |
|---|---|---|---|
| Naive | 1 | 1 | 1 |
| Grand mean | 0.852 | 0.127 | 0.378 |
| Parametric EB-MM | 0.593 | 0.129 | 0.387 |
| Parametric EB-ML | 0.902 | 0.117 | 0.398 |
| James-Stein | 0.525 | 0.164 | 0.359 |
| Nonparametric EB | 0.508 | 0.212 | 0.372 |
| Binomial mixture | 0.588 | 0.156 | 0.314 |
| Weighted least square (Null) | 1.074 | 0.127 | 0.468 |
| Weighted generalized MLE (Null) | 0.306 | 0.173 | 0.326 |
| Weighted least square (AB) | 0.537 | 0.087 | 0.290 |
| Weighted generalized MLE (AB) | 0.301 | 0.141 | 0.261 |
| SURE $\hat{\boldsymbol{\theta}}^G$ | 0.505 | 0.123 | 0.278 |
| SURE $\hat{\boldsymbol{\theta}}^M$ | 0.422 | 0.123 | 0.282 |
| SPSURE $\hat{\boldsymbol{\theta}}^{SG}$ | 0.409 | 0.081 | 0.261 |
| SPSURE $\hat{\boldsymbol{\theta}}^{SM}$ | 0.419 | 0.077 | 0.278 |

The numerical results demonstrate that the SURE estimators have quite appealing performance. The total squared errors of the SURE estimators are significantly smaller than almost all of their competitors, with the only exception being that the weighted general maximum likelihood methods achieve a better performance in the all players' case. The main reason, we believe, is that the baseball data contain features that may degrade the performance of classical empirical Bayes methods, as discussed by Brown (2008). For example, substantial evidence against the normal prior assumption is observed, and, furthermore, ignoring the correlation between the mean $\theta$ and the variance $A$ is not justifiable here (a player with large $p$ tends to play more games, resulting in large $N$). Both of these features can invalidate the use of empirical Bayes methods. On the other hand, our SURE shrinkage estimators, especially the semiparametric ones, are shown to be robust and optimal in much more general circumstances, resulting in the superior numerical outcome.

Figure 2 plots the shrinkage factor for four of the estimators we have considered for the "all batters" data—the EBMOM, EBMLE, the parametric SURE estimator $\hat{\boldsymbol{\theta}}^M$, and the semiparametric SURE estimator $\hat{\boldsymbol{\theta}}^{SM}$. In the parametric case, the

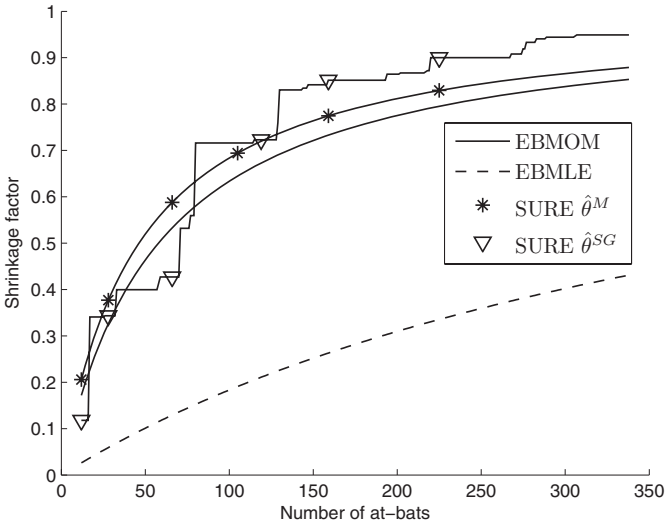



Figure 2. Plot of the shrinkage factors $\hat{\lambda}/(\hat{\lambda} + A)$ or $1 - \hat{b}$ for all-batters. Four estimators are compared: EB-MM, EB-MLE, the parametric SURE $\hat{\boldsymbol{\theta}}^M$, and the semiparametric SURE estimator $\hat{\boldsymbol{\theta}}^{SM}$.

shrinkage factor is $\hat{\lambda}/(\hat{\lambda}+A)$; in the semiparametric case, it is $1-\hat{b}^{SM}$ as in Equation (6.2). Note that in each case the shrinkage factor ($\hat{\lambda}/(\hat{\lambda}+A)$ or $1-\hat{b}^{SM}$) increases with $N$, the number of at-bats, as they should. This corresponds to a decrease in terms of $A=1/4N$. Note that the shrinkage factor for the EB-ML estimator is much smaller than those for the other estimators, which corresponds to greater shrinkage to the central location, and this is intimately related to the relatively poor performance of this estimator for the current dataset. Note also that $1-\hat{b}^{SM}$ increases with $N$ in a stepwise fashion. The fact that it is nondecreasing is a direct consequence of its definition. The stepwise property is an indirect consequence of its definition—monotone solutions to the minimization problem in Equation (6.2) or (6.3) will always be stepwise monotone. Finally, note that for large values of $N$ (approximately $N \geq 130$), $1-\hat{b}^{SM}$ has the largest value among the four shrinkage factors. Thus, for this dataset, the SURE estimator $\hat{\theta}^{SM}$ shrinks somewhat less than the EB-MM estimator or the parametric SURE estimator $\hat{\theta}^{M}$ when $N$ is large, but shrinks comparably to these estimators for smaller $N$. It is also true that the estimates of central tendency differ for these estimators, but the differences are small to moderate. The corresponding values of $\hat{\mu}$ are $\hat{\mu}=0.528, 0.538, 0.456, 0.529$, respectively.

### 8.2 Estimation of Housing Price

In this subsection, we apply the SURE estimators to a housing dataset. The goal is to estimate the average housing price in each town of Scheffield, England, from a small fraction of the data, as would be the case of a survey sampling. The data was produced by the Land Registry of the United Kingdom. It contains the information about all houses sold in Scheffield, England, from 2000 to mid-2008. The sale price, the sale time, the postcode that identifies the location of the house, and other relevant statistics about the sales are available for each house that has been sold during this time period. Nagaraja et al. (2011) discussed various analysis of similar, larger datasets from the United States. We here confine our interest mainly to the estimation of average housing prices for each town in Scheffield, which has a distinct postcode. As conventional, the logarithm of the housing prices are used throughout the study to better approximate the normality assumption. Our analysis starts by removing the inflation from year to year by subtracting from each sample the overall year effect. (A more sophisticated method might build a two-way model with the year effect treated as a fixed effect and the area effect as a random effect.) We then randomly draw a small fraction of the data. This small fraction serves as a survey sample, from which we want to estimate the average housing price of each town of the entire dataset. One particularly interesting feature of the dataset is that the number of towns is small (around 20), while the number of house sales in most towns is large (above 500). To have a clear picture, we let the survey sample size range from 10% to 20% of the entire data. We exclude the towns with less than 20 house sales so that we would at least have three data points in the sample for each town.

To compare the performance of different shrinkage estimators, we run the simulation $N=10{,}000$ times and report the average results in Table 3. The variances $A_i$ are estimated from the sample variance of each town. As in the previous example, we use the naive estimator, the sample mean of each town, as the benchmark. Each number in the table refers to the ratio of the squared error of a particular estimator to that of the naive estimator. TSE stands for the total squared estimation error on the logarithm scale, while TSEP corresponds to that on the original scale. Note that unlike the baseball data, the parameter of interest $\theta_j$ (the average housing prices) can be directly obtained here. We can therefore evaluate the actual TSE and TSEP instead of estimating them through adjusting the prediction errors. There has also been discussion on alternatives other than squared error loss in the study of housing price (see Varian 1975, for one such example).

Table 3. Estimation errors for housing prices under different sample sizes

| | Sampling 10% | | Sampling 15% | | Sampling 20% | |
|---|---|---|---|---|---|---|
| | TSE | TSEP | TSE | TSEP | TSE | TSEP |
| Naive | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| EB(MM) | 0.735 | 0.748 | 0.779 | 0.788 | 0.816 | 0.824 |
| EB(ML) | 0.734 | 0.747 | 0.776 | 0.785 | 0.813 | 0.821 |
| J-S | 0.991 | 0.992 | 0.994 | 0.994 | 0.996 | 0.996 |
| SURE $\hat{\boldsymbol{\theta}}^{G}$ | 0.746 | 0.793 | 0.772 | 0.820 | 0.819 | 0.872 |
| SURE $\hat{\boldsymbol{\theta}}^{M}$ | 0.997 | 1.074 | 1.050 | 1.131 | 1.100 | 1.184 |
| SPSURE $\hat{\boldsymbol{\theta}}^{SG}$ | 0.556 | 0.574 | 0.518 | 0.534 | 0.522 | 0.536 |
| SPSURE $\hat{\boldsymbol{\theta}}^{SM}$ | 0.879 | 0.881 | 0.863 | 0.863 | 0.865 | 0.865 |

There are several interesting observations. First, the improvement of shrinkage estimators over the naive estimator as a group is not as impressive as in the baseball data case, though significant error reduction is still achieved. This is because the number of groups here is significantly smaller (20 here compared to around 500 in the baseball data). Second, as the sample size increases, the relative performance of shrinkage estimators decreases. This is because the variance of each sample mean decreases, resulting in smaller shrinkage. Third, overall speaking, the SURE shrinkage estimators achieve better performance compared with the other shrinkage estimators. The good performance of the semiparametric SURE estimator $\hat{\boldsymbol{\theta}}^{SG}$ is particularly noteworthy. Fourth, when the number of groups $p$ is small (around 20 here), it is not necessarily always beneficial to simultaneously estimate $\mu$, the shrinkage location, and the shrinkage factors, since the asymptotic result is yet to take effect. Shrinking the estimates toward a predetermined location such as the grand mean could give better results.

## 9. SUMMARY

Inspired by Stein's unbiased risk estimate (SURE), we propose in this article a class of shrinkage estimators for the heteroscedastic hierarchical model, which is arguably more realistic in practical applications. We show that each SURE shrinkage estimator is asymptotically optimal in its own class. This includes the parametric SURE estimators, whose forms are derived from the classical parametric hierarchical model, as well as semiparametric SURE estimators, which only assume that the individual shrinkage factor is monotone in the variance. We note that the asymptotic optimality of the SURE shrinkage estimators do not depend on the specific distributional assumptions, such as the normal assumption. We test the SURE estimators in

comprehensive simulation studies and two real datasets, observing encouraging results: the SURE estimators offer numerically superior performance compared to the classical empirical Bayes and James-Stein estimators. The semiparametric SURE estimators appear to be particularly competitive. We recommend the use of the semiparametric SURE estimator $\hat{\boldsymbol{\theta}}^{SM}$ (where the shrinkage location is simultaneously estimated), when the number of groups are large. For data with small number of groups, we recommend the semiparametric SURE estimator $\hat{\boldsymbol{\theta}}^{SG}$, which shrinks toward the grand mean.

There are several relevant research questions not fully addressed in this article. For example, the sparse normal means problem (Johnstone and Silverman 2004) has become increasingly important in statistics. It therefore would be of interest to study the performance of the proposed methods under this setting. It could also be of interest to study the extent to which the proposed estimators are minimax by using the techniques discussed by Maruyama and Strawderman (2005). It would also be of interest to study whether these estimators are ensemble minimax in the sense of Efron and Morris (1973) and Brown, Nie, and Xie (submitted). The peculiar features in the baseball data suggest that models that explicitly consider the dependence between $\theta_i$'s and $A_i$'s might be more appropriate. For example, we can consider a hierarchical Bayes model where $\theta_i$ explicitly depends on $A_i$. It is interesting to see how the performance of such estimators is compared with the SURE estimators proposed in this article.

## APPENDIX: PROOFS

*Proof of Theorem 3.1.* We only need to show the $L^2$ convergence. Since

$$\begin{aligned}&\mathrm{SURE}(\lambda) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})\\ &\quad= \frac{1}{p}\sum_i \left(X_i^2 - A_i - \theta_i^2 - \frac{2\lambda}{A_i+\lambda}\left(X_i^2 - X_i\theta_i - A_i\right)\right),\end{aligned}$$

we know

$$\begin{aligned}&\sup_{0\le\lambda\le\infty}\left|\mathrm{SURE}(\lambda) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})\right|\\ &\quad\le \sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_i \frac{2\lambda}{A_i+\lambda}\left(X_i^2 - X_i\theta_i - A_i\right)\right| + \left|\frac{1}{p}\sum_i\left(X_i^2 - A_i - \theta_i^2\right)\right|.\end{aligned}$$

We consider the two terms separately. For the first term, without loss of generality, let us assume $A_1 \le A_2 \le \cdots \le A_p$. Then we know

$$\begin{aligned}&\sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_i \frac{2\lambda}{A_i+\lambda}\left(X_i^2 - X_i\theta_i - A_i\right)\right|\\ &\qquad\le \sup_{1\ge c_1\ge\cdots\ge c_p\ge 0}\left|\frac{2}{p}\sum_{i=1}^{p} c_i\left(X_i^2 - X_i\theta_i - A_i\right)\right|.\end{aligned}$$

As in Lemma 2.1 by Li (1986), observe that

$$\begin{aligned}&\sup_{1\ge c_1\ge\cdots\ge c_p\ge 0}\left|\frac{2}{p}\sum_{i=1}^{p} c_i\left(X_i^2 - X_i\theta_i - A_i\right)\right|\\ &\qquad= \max_{1\le j\le p}\left|\frac{2}{p}\sum_{i=1}^{j}\left(X_i^2 - X_i\theta_i - A_i\right)\right|.\end{aligned}$$

Let $M_j = \sum_{i=1}^{j}(X_i^2 - X_i\theta_i - A_i)$. Then $\{M_j; j = 1, 2, \ldots\}$ forms a martingale. The $L^p$ maximum inequality implies

$$E\left(\max_{1\le j\le p} M_j^2\right) \le 4E\left(M_p^2\right) = 4\sum_{i=1}^{p}\left(2A_i^2 + A_i\theta_i^2\right).$$

Regularity conditions (A) and (B) thus guarantee that $E(\max_j (\frac{2}{p}M_j)^2) \to 0$, which yields

$$\sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_i \frac{2\lambda}{A_i+\lambda}\left(X_i^2 - X_i\theta_i - A_i\right)\right| \to 0 \text{ in } L^2, \text{ as } p\to\infty.$$

For the second term $\frac{1}{p}\sum_i(X_i^2 - A_i - \theta_i^2)$, a direct calculation gives

$$E\left[\left(\frac{1}{p}\sum_i\left(X_i^2 - A_i - \theta_i^2\right)\right)^2\right] = \frac{2}{p^2}\sum_{i=1}^{p}\left(A_i^2 + 2A_i\theta_i^2\right) \to 0,$$

by conditions (A) and (B). This completes the proof. □

*Proof of Theorem 3.2.* Since $\mathrm{SURE}(\hat{\lambda}_{\mathrm{SURE}}) \le \mathrm{SURE}(\tilde{\lambda}^{OL})$ by definition, and we know from the preceding theorem that $\sup_\lambda |\mathrm{SURE}(\lambda) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})| \to 0$ in probability, it follows that for any $\varepsilon > 0$

$$\begin{aligned}&P(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) \ge l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL}) + \varepsilon)\\ &\quad\le P\left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) - \mathrm{SURE}(\hat{\lambda}_{\mathrm{SURE}}) \ge l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL}) - \mathrm{SURE}(\tilde{\lambda}^{OL}) + \varepsilon\right)\\ &\quad\to 0,\end{aligned}$$

which completes the proof. □

*Proof of Corollary 3.1.* This is a direct consequence of the definition of $\tilde{\boldsymbol{\theta}}^{OL}$ and Theorem 3.2. □

*Proof of Theorem 3.3.* Since

$$\begin{aligned}&l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) - l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL})\\ &\quad= (l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) - \mathrm{SURE}(\hat{\lambda}_{\mathrm{SURE}})) + (\mathrm{SURE}(\hat{\lambda}_{\mathrm{SURE}}) - \mathrm{SURE}(\tilde{\lambda}^{OL}))\\ &\qquad+ (\mathrm{SURE}(\tilde{\lambda}^{OL}) - l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL}))\\ &\quad\le 2\sup_{0\le\lambda\le\infty}|\mathrm{SURE}(\lambda) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda})|,\end{aligned}$$

we know from Theorem 3.1 that

$$l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) - l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL}) \to 0 \text{ in } L^2 \text{ and in } L^1.$$

Therefore,

$$\lim_{p\to\infty}\left[R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{SURE}}) - R_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{OL})\right] = 0.$$ □

*Proof of Corollary 3.2.* This is a direct consequence of the definition of $\tilde{\boldsymbol{\theta}}^{OL}$ and Theorem 3.3. □

*Proof of Theorem 4.1.* Since

$$\begin{aligned}&\mathrm{SURE}^G(\lambda) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\bar{X}})\\ &\quad= \frac{1}{p}\sum_i\left(\left(X_i^2 - A_i - \theta_i^2\right) - \frac{2\lambda}{A_i+\lambda}\left(X_i^2 - X_i\theta_i - A_i\right)\right.\\ &\qquad\left.- \frac{2A_i}{A_i+\lambda}\left(\bar{X}(X_i - \theta_i) - \frac{A_i}{p}\right)\right),\end{aligned}$$

we have

$$\sup_{0\le\lambda\le\infty}\left|\mathrm{SURE}^G(\lambda)-l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\lambda,\bar{X}})\right|$$
$$\le\left|\frac{1}{p}\sum_i\left(X_i^2-A_i-\theta_i^2\right)\right|+\sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_i\frac{2\lambda}{A_i+\lambda}\left(X_i^2-X_i\theta_i-A_i\right)\right|$$
$$+\sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_i\frac{2A_i}{A_i+\lambda}\left(\bar{X}(X_i-\theta_i)-\frac{A_i}{p}\right)\right|. \qquad \square$$

The convergence of the first two terms in $L^2$ has already been established in the proof of Theorem 3.1. We only need to show that the last term converges to 0 in $L^1$. But

$$\sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_i\frac{2A_i\left(\bar{X}(X_i-\theta_i)-A_i/p\right)}{A_i+\lambda}\right|$$
$$\le\sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_i\frac{2A_i(X_i-\theta_i)}{A_i+\lambda}\right|\cdot\left|\bar{X}\right|+\frac{2}{p^2}\sum_i A_i.$$

Following the technique in the proof of Theorems 3.1, it can be shown that

$$\sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_{i=1}^{p}\frac{2A_i(X_i-\theta_i)}{A_i+\lambda}\right|\to 0 \text{ in } L^2.$$

We also know that $E\bar{X}^2=\frac{1}{p^2}\sum_i A_i+(\frac{1}{p}\sum_i\theta_i)^2$, which is bounded by Conditions (A) and (C). Therefore, we have

$$\sup_{0\le\lambda\le\infty}\left|\frac{1}{p}\sum_i\frac{2A_i(X_i-\theta_i)}{A_i+\lambda}\right|\cdot\left|\bar{X}\right|\to 0 \text{ in } L^1,$$

by Cauchy-Schwartz inequality, and this completes the proof.

*Proof of Theorem 4.2.* With Theorem 4.1 established, the proof is almost identical to that of Theorem 3.2. $\square$

*Proof of Corollary 4.1.* This is a direct consequence of the definition of $\tilde{\boldsymbol{\theta}}^{GOL}$ and Theorem 4.2. $\square$

*Proof of Theorem 4.3.* Since

$$l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^G)-l_p(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}^{GOL})$$
$$=(l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^G)-\mathrm{SURE}^G(\hat{\lambda}_G))+(\mathrm{SURE}^G(\hat{\lambda}_G)-\mathrm{SURE}^G(\tilde{\lambda}^{GOL}))$$
$$+(\mathrm{SURE}^G(\tilde{\lambda}^{GOL})-l_p(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}^{GOL}))$$
$$\le 2\sup_{0\le\lambda\le\infty}|\mathrm{SURE}^G(\lambda)-l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\lambda,\bar{X}})|,$$

we know from Theorem 4.1 that

$$l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\mathrm{SURE}})-l_p(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}^{OL})\to 0 \text{ in } L^1.$$

Therefore,

$$\lim_{p\to\infty}[R_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\mathrm{SURE}})-R_p(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}^{OL})]=0. \qquad \square$$

*Proof of Corollary 4.2.* This is a direct consequence of the definition of $\tilde{\boldsymbol{\theta}}^{GOL}$ and Theorem 4.3. $\square$

To prove Theorem 5.1, we need the following lemma.

*Lemma A.1.* Assume conditions (A), (B), and (C′). Then we have

$$E\left(\max_{1\le i\le p}X_i^2\right)=O(p^{2/(2+\delta^*)}),$$

where $\delta^*=\min(1,\delta)$.

*Proof.* We can write $X_i=\sqrt{A_i}Z_i+\theta_i$, where $Z_i$ are iid standard normal random variables. It follows from $X_i^2=A_iZ_i^2+\theta_i^2+2\sqrt{A_i}\theta_iZ_i$ that

$$\max_{1\le i\le p}X_i^2\le\max_i A_i\cdot\max_i Z_i^2+\max_i\theta_i^2+2\max_i\sqrt{A_i}\,|\theta_i|\cdot\max_i|Z_i|. \tag{A.1}$$

Condition (A) implies that $\max_i A_i^2\le\sum_i A_i^2=O(p)$. Thus, $\max_i A_i=O(p^{1/2})$. Similarly, Condition (B) implies that $\max_i\sqrt{A_i}|\theta_i|=O(p^{1/2})$. Condition (C′) implies that $\sum_i|\theta_i|^{2+\alpha}=O(p)$ for all $0\le\alpha\le\delta$; in particular, $\sum_i|\theta_i|^{2+\delta^*}=O(p)$. Since $\max_i|\theta_i|^{2+\delta^*}\le\sum_i|\theta_i|^{2+\delta^*}=O(p)$, we know that $\max_i\theta_i^2=O(p^{2/(2+\delta^*)})$. It is well known (see, e.g., Embrechts et al. 1997, chap. 3) that

$$E\left(\max_{1\le i\le p}|Z_i|\right)=O(\sqrt{\log p}),\quad E\left(\max_{1\le i\le p}Z_i^2\right)=O(\log p).$$

Taking them into Equation (A.1), we obtain

$$E\left(\max_{1\le i\le p}X_i^2\right)=O(p^{1/2}\log p)+O(p^{2/(2+\delta^*)})+O(p^{1/2}\sqrt{\log p})$$
$$=O(p^{2/(2+\delta^*)}). \qquad \square$$

*Proof of Theorem 5.1:* Since

$$\mathrm{SURE}^M(\lambda,\mu)-l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\lambda,\mu})=\mathrm{SURE}(\lambda)-l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\lambda})$$
$$-\frac{2\mu}{p}\sum_i\frac{A_i}{A_i+\lambda}(X_i-\theta_i),$$

it follows that

$$\sup_{0\le\lambda\le\infty,\,|\mu|\le\max_i|X_i|}|\mathrm{SURE}^M(\lambda,\mu)-l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\lambda,\mu})|$$
$$\le\sup_{0\le\lambda\le\infty}|\mathrm{SURE}(\lambda)-l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\lambda})|+\frac{2}{p}\max_{1\le i\le p}|X_i|$$
$$\times\sup_{0\le\lambda\le\infty}\left|\sum_i\frac{A_i(X_i-\theta_i)}{A_i+\lambda}\right|. \tag{A.2}$$

We know from Theorem 3.1 that

$$\sup_{0\le\lambda\le\infty}|\mathrm{SURE}(\lambda)-l_p(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\lambda})|\to 0 \text{ in } L^2.$$

It remains to show that the second term in Equation (A.2) converges to zero in $L^1$.

Following the same steps as in the proof of Theorem 3.1, we can show that

$$\sup_{0\le\lambda\le\infty}\left|\sum_i\frac{A_i(X_i-\theta_i)}{A_i+\lambda}\right|\le\max_{1\le j\le p}\left|\sum_{i=j}^{p}(X_i-\theta_i)\right|.$$

Therefore, by the $L^p$ maximum inequality on martingales, we have

$$E\left\{\sup_{0\le\lambda\le\infty}\left[\sum_i\frac{A_i(X_i-\theta_i)}{A_i+\lambda}\right]^2\right\}\le E\left\{\max_{1\le j\le p}\left[\sum_{i=j}^{p}(X_i-\theta_i)\right]^2\right\}$$
$$\le 4E\left(\sum_{i=1}^{p}(X_i-\theta_i)\right)^2$$
$$=4\sum_i A_i=O(p).$$

Combining this with Lemma A.1, we obtain by Cauchy-Schwartz inequality

$$\frac{1}{p}E\left(\max_{1\le i\le p}|X_i|\cdot\sup_{0\le\lambda\le\infty}\left|\sum_i\frac{A_i(X_i-\theta_i)}{A_i+\lambda}\right|\right)$$
$$\le\frac{1}{p}\left(E(\max_{1\le i\le p}X_i^2)\cdot E\left\{\sup_{0\le\lambda\le\infty}\left[\sum_i\frac{A_i(X_i-\theta_i)}{A_i+\lambda}\right]^2\right\}\right)^{1/2}$$
$$=O(p^{-\frac{\delta^*}{2(2+\delta^*)}})=o(1),$$

which completes the proof. □

*Proof of Theorem 5.2.* With Theorem 5.1 established, the proof of

$$\lim_{p\to\infty} P(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^M) \geq l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{MOL}) + \varepsilon) = 0 \quad \text{for any } \varepsilon > 0$$

is essentially identical to the proof of Theorem 3.2. □

*Proof of Theorem 5.1.* This is a direct consequence of the definition of $\tilde{\boldsymbol{\theta}}^{MOL}$ and Theorem 5.2. □

*Proof of Theorem 5.3.* Since

$$l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^M) - l_p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{MOL}) \leq 2 \sup_{0\leq\lambda\leq\infty,\, |\mu|\leq\max_i |X_i|} |\mathrm{SURE}^M(\lambda, \mu) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\mu})|,$$

the result follows from Theorem 5.1. □

*Proof of Corollary 5.2.* This is a direct consequence of the definition of $\tilde{\boldsymbol{\theta}}^{MOL}$ and Theorem 5.3. □

*Proof of Corollary 5.3.* This is a special case of Corollary 5.2. □

*Proof of Theorem 6.1.* First, we have

$$\begin{aligned}|\mathrm{SURE}^M(\boldsymbol{b}, \mu) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\mu})| &\leq \frac{1}{p}\left|\sum_{i=1}^p \left(X_i^2 - \theta_i^2 - A_i\right)\right| + \frac{1}{p}\left|\sum_{i=1}^p 2(1-b_i)\left(X_i^2 - X_i\theta_i - A_i\right)\right| \\ &\quad + \frac{1}{p}\left|\sum_{i=1}^p 2b_i(X_i - \theta_i)\cdot\mu\right|. \end{aligned}$$ □

Note that the order of $b_i$ is determined by that of $A_i$, which is not random. The rest of the proof follows essentially the same steps as in that of Theorem 5.1 upon using Lemma A.1.

*Proof of Theorem 6.2.* With Theorem 6.1 established, the proof of

$$\lim_{p\to\infty} P(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{SM}) \geq l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{b}_p,\hat{\mu}_p}) + \varepsilon) = 0 \quad \text{for any } \varepsilon > 0$$

is the same as the proof of Theorem 3.2. Likewise, to show

$$\limsup_{p\to\infty}[R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{SM}) - R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{b}_p,\hat{\mu}_p})] \leq 0,$$

we use the inequality

$$l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{SM}) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{b}_p,\hat{\mu}_p}) \leq 2\sup|\mathrm{SURE}^M(\boldsymbol{b}, \mu) - l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\mu})|,$$

and Theorem 6.1. □

*[Received November 2011. Revised March 2012.]*

## REFERENCES


Akaike, H. (1973), "Information Theory as an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademiai Kiado, pp. 267–281. [1470]

Baranchik, A. J. (1970), "A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution," *The Annals of Mathematical Statistics*, 41, 642–645. [1465]

Berger, J. (1976), "Admissible Minimax Estimation of a Multivariate Normal Mean With Arbitrary Quadratic Loss," *The Annals of Statistics*, 4, 223–226. [1465]

Berger, J., and Strawderman, W. E. (1996), "Choice of Hierarchical Priors: Admissibility in Estimation of Normal Means," *The Annals of Statistics*, 24, 931–951. [1465]

Brown, L. D. (1971), "Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems," *The Annals of Mathematical Statistics*, 42, 855–903. [1465]

——— (1975), "Estimation With Incompletely Specified Loss Functions (the Case With Several Location Parameters)," *Journal of the American Statistical Association*, 70, 417–427. [1465]

——— (2008), "In-Season Prediction of Batting Averages: A Field Test of Empirical Bayes and Bayes Methodologies," *The Annals of Applied Statistics*, 2, 113–152. [1465,1471,1473,1474]

Brown, L. D., and Greenshtein, E. (2009), "Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-Dimensional Vector of Means," *The Annals of Statistics*, 37, 1685–1704. [1465,1474]

Brown, L. D., Nie, H., and Xie, X. (Submitted), "Ensemble Minimax Estimation for Multivariate Normal Means". [1476]

Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov A. B. (2002), "Oracle Inequalities for Inverse Problems," *The Annals of Statistics*, 30, 843–874. [1467]

Donoho, D. L., and Johnstone, I. M. (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 1200–1224. [1467]

Efron, B. (1986), "How Biased is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association*, 81, 461–470. [1470]

——— (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," *Journal of American Statistical Association*, 99, 619–642. [1470]

Efron, B., and Morris, C. (1973), "Stein's Estimation Rule and Its Competitors: An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130. [1465,1476]

——— (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of American Statistical Association*, 70, 311–319. [1465]

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), *Modelling Extreme Events*, New York: Springer. [1477]

Hudson, H. M. (1974), "Empirical Bayes Estimation," Technical Report No. 58, Department of Statistics, Stanford University. [1465]

James, W., and Stein, C. M. (1961), "Estimation With Quadratic Loss," *Proceedings of the 4th Berkeley Symposium on Probability and Statistics*, I, 367–379. [1465]

Jiang, W., and Zhang, C.-H. (2009), "General Maximum Likelihood Empirical Bayes Estimation of Normal Means," *The Annals of Statistics*, 37, 1647–1684. [1474]

——— (2010), "Empirical Bayes In-Season Prediction of Baseball Batting Averages," in *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*, eds. J. Berger, T. Cai and M. J. Iain, Beachwood, OH: Institute of Mathematical Statistics, pp. 263–273. [1474]

Johnstone, I. M. (1987), "On the Admissibility of Some Unbiased Estimates of Loss," in *Statistical Decision Theory and Related Topics IV*, eds. G. S. Gupta and J. Berger, New York: Springer-Verlag, pp. 361–380. [1467]

Johnstone, I. M., and Silverman, B. M. (2004), "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences," *The Annals of Statistics*, 32, 1594–1649. [1476]

Kneip, A. (1994), "Ordered Linear Smoothers," *The Annals of Statistics*, 22, 835–866. [1467]

Li, K.-C. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-Validation," *The Annals of Statistics*, 13, 1352–1377. [1467]

——— (1986), "Asymptotic Optimality of $C_L$ and Generalized Cross-Validation in Ridge Regression With Application to Spline Smoothing," *The Annals of Statistics*, 14, 1101–1112. [1467,1476]

——— (1987), "Asymptotic Optimality for CP, CL, Cross-Validation, and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975. [1467]

Lindley, D. V. (1962), Discussion of "Confidence Sets for the Mean of a Multivariate Normal Distribution" by C. M. Stein, *Journal of the Royal Statistical Society*, Series B, 24, 285–287. [1465]

Mallows, C. L. (1973), "Some Comments on CP," *Technometrics*, 15, 661–675. [1470]

Maruyama, Y., and Strawderman, W. (2005), "A New Class of Generalized Bayes Minmax Ridge Regression Estimators," *The Annals of Statistics*, 33, 1753–1770. [1476]

Morris, C. (1983). "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–55. [1465]

Muralidharan, O. (2010), "An Empirical Bayes Mixture Method for Effect Size and False Discovery Rate Estimation," *The Annals of Applied Statistics*, 4, 422–438. [1474]

Nagaraja, C. H., Brown, L. D., and Zhao, L. H. (2011), "An Autoregressive Approach to House Price Modeling," *The Annals of Applied Statistics*, 5, 124–149. [1475]

Stein, C. M. (1962), "Confidence Sets for the Mean of a Multivariate Normal Distribution" (with discussion), *Journal of the Royal Statistical Society*, Series B, 24, 265–296. [1465]

——— (1973), "Estimation of the Mean of a Multivariate Distribution," *Proceedings of the Prague Symposium on Asymptotic Statistics*, 345–381. [1465]

——— (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151. [1465]

Strawderman, W. (1971), "Proper Bayes Estimators of the Multivariate Normal Mean," *The Annals of Mathematical Statistics*, 42, 385–388. [1465]

Varian, H. (1975), "A Bayesian Approach to Real Estate Assessment," in *Studies in Bayesian Econometrics and Statistics*, S. Fienberg and A. Zellner, eds. Amsterdam: Elsevier North-Holland, pp. 195–208. [1475]

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120–131. [1470]