On SURE-Type Double Shrinkage Estimation

Jing, Bing-Yi; Li, Zhouping; Pan, Guangming; Zhou, Wang

2016

Jing, B.-Y., Li, Z., Pan, G., & Zhou, W. (2016). On SURE-Type Double Shrinkage Estimation. *Journal of the American Statistical Association*, 111(516), 1696-1704.

<https://hdl.handle.net/10356/85599>

<https://doi.org/10.1080/01621459.2015.1110032>

© 2016 American Statistical Association. This is the author created version of a work that has been peer reviewed and accepted for publication in *Journal of the American Statistical Association*, published by Taylor & Francis on behalf of American Statistical Association. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [<http://dx.doi.org/10.1080/01621459.2015.1110032>].

Downloaded on 26 Apr 2023 18:54:12 SGT

On SURE-Type Shrinkage Estimation

Bing-Yi JING, Zhouping LI, Guangming PAN and Wang ZHOU
HKUST, LZU, NTU and NUS

Abstract

The paper is concerned with empirical Bayes shrinkage estimators for the heteroscedastic hierarchical normal model based on Stein's unbiased estimate of risk (SURE). Recently, Xie, Kou and Brown (2012) proposed a class of estimators for this type of problems and established their asymptotic optimality properties under the assumption of known but unequal variances. In this paper, we consider this problem with unequal and unknown variances, which may be more appropriate in real situations. Specifically, by putting priors for both means and variances, we **propose** novel double-shrinkage SURE type estimators that shrink both the means and the variances. Optimal properties for these estimators **are** derived under certain regularity conditions. We also **conduct** extensive simulation studies to compare the newly developed methodologies with other shrinkage techniques. Finally, the methods **are** applied to the **well known baseball data set and a gene expression data set**.

KEY WORDS: Empirical Bayes; Hierarchical model; Shrinkage both means and variances; Stein's unbiased risk estimate (SURE).

1 Introduction

Suppose that $\mathbf{X} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where $\mathbf{X} = (X_1, \dots, X_p)$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ and $\boldsymbol{\Sigma} = \text{diag}\{\sigma_1^2, \dots, \sigma_p^2\}$. That is, we assume that $\mathbf{X}$ is multivariate normal with independent components but unequal variances. We are particularly interested in the case of large p problems. Simultaneous estimation of the mean parameters θ_i 's is a fundamental yet important theoretical issue, which has become increasingly popular in modern applications of statistics. James and Stein (1961) proposed an estimator (now called James-Stein estimator), which dominates the usual maximum likelihood estimator $\hat{\boldsymbol{\theta}} = \mathbf{X}$ under squared error loss function when $p \geq 3$ and all variances σ_i^2 's are known and equal. Since then, many shrinkage estimators have been proposed and widely applied as a powerful tool to improve over traditional approaches in various problems. References include Efron and Morris (1973,1975), Berger and Strawderman (1996), Maruyama and Strawderman (2005), Tong and Wang (2007), Brown (2008) and Brown and Greenshtein (2009), Jiang and Zhang (2009), Tong, Chen and Zhao (2012), Brown, Nie and Xie (2012), among others.

Recently, borrowing the idea of the Stein's unbiased risk estimate (SURE), Xie, Kou and Brown (2012) proposed a class of SURE estimators to simultaneously estimate mean parameters for the hierarchical model, and established the asymptotic optimality property for these estimators. Note that they assume heteroscedastic but known variances, one may ask whether it is possible to extend Stein's idea to the case of unknown variances. For unknown variance, a usual strategy is to replace it by some unbiased estimator such as the sample variance, then treat the problem as if the variance was known. This could be used in the case of small p , however when p is large, there would be advantages in shrinking the variances as well as the means, see Hwang, Qiu and Zhao (2009) and references therein for detailed discussions.

We try to answer this question by introducing a double-shrinkage estimator in this paper. In particular, we use the hierarchical normal model to put objective priors on σ_i^2 's as well as on θ_i 's, and derive a class of estimators shrinking both the mean and variance. Our theoretical results can be summarized from the following three aspects: (i) Besides shrinking the means, our estimators also shrink the variances, which brings us additional benefit; (ii) The posterior distribution of variances generally has no explicit form, one needs

to approximate the posterior (see Zhao, 2010). By putting a prior of variances, together with SURE estimate, we can overcome this and obtain the exact expression; (iii) As observed by Xie, Kou and Brown (2012), one particular feature of the 2005 baseball data in Brown (2008) is its dependence between the means and the variances, our procedures have appeal in a flexible setup where this dependence may exist, and the improvement is noticeable. It is worth mentioning that our work is different with Brown and Han (2012), which also considered the optimal estimation of multivariate normal means with and unknown variance, but they concentrated on the admissibility criterion.

The rest of this paper is organized as follows. In Section 2, we introduce our **double SURE** shrinkage estimators for the normal-normal hierarchical model and discuss their asymptotic optimality properties. **Section 3 considers the SURE estimators that shrink the means toward the grand mean.** Section 4 develops a class of semi-parametric double-shrinkage estimators. The simulation and real data application will be presented in Sections 5 and 6. All proofs are relegated to the appendix.

2 Simultaneous estimation of means and variances

In this section, we consider simultaneous point estimation for the means and the variances of the multivariate normal distribution. Specifically, we assume the following Bayes hierarchical model for the observations (X_i, S_i) , $i = 1, \dots, p$,

$$\begin{aligned} X_i | (\theta_i, \sigma_i^2) &\stackrel{ind}{\sim} N(\theta_i, \sigma_i^2), \\ S_i | \sigma_i^2 &\stackrel{ind}{\sim} \sigma_i^2 \chi_m^2, \\ \theta_i | \sigma_i^2 &\stackrel{ind}{\sim} N(\mu, \sigma_i^2 / \kappa), \\ 1/\sigma_i^2 &\stackrel{iid}{\sim} \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}A\right), \end{aligned} \tag{1}$$

where X_i and S_i are independent **conditioning on σ_i^2** , and μ and γ denote the unknown hyper-parameters with

$$\gamma = (\kappa, \nu, A), \quad \text{with } \kappa, \nu - 2, A > 0.$$

Remark. We emphasize two elements. First, in the canonical model (1), we hope to estimate the parameter (θ_i, σ_i^2) from the single observation (X_i, S_i) . In general, from $m + 1$

potential sample $(X_{i,1}, \dots, X_{i,m+1})$, X_i and S_i can be regarded as the unbiased sample mean and sample variance estimators of θ_i and $m\sigma_i^2$, respectively. Notice that $m+1$ plays the role as the group size and is small in practice, see for example, the baseball data in Section 5, thus m is fixed throughout the paper. Second, there has been a great deal of work in the high dimensional problems, for example, shrinkage and variable selection (Fan and Li, 2001; Efron et al., 2004; Zou, 2006), covariance/precision matrix estimation (Bickel and Levina, 2008(a,b); Lam and Fan, 2009; Cai, Zhang and Zhou, 2010), large-scale multiple testing (Cai and Jin, 2010; Bogdan et al., 2011), among others. In addition to these literatures, the hierarchical model is a classic one which has been extensively studied and widely used in both statistics and applied areas. Our interest lies in the hierarchical normal model (1), and the estimation of the high dimensional mean vector $(\theta_1, \dots, \theta_p)$.

The following conditions will be made throughout the paper.

- **(C1)** $\limsup_{p \rightarrow \infty} p^{-1} \sum_{i=1}^p \sigma_i^8 < \infty$,
- **(C2)** $\limsup_{p \rightarrow \infty} p^{-1} \sum_{i=1}^p \theta_i^2 \sigma_i^2 < \infty$,
- **(C3)** $\limsup_{p \rightarrow \infty} p^{-1} \sum_{i=1}^p |\theta_i|^{2+\delta} < \infty$, for some $\delta > 0$.

Under this model, a routine calculation gives that

$$\theta_i | (\sigma_i^2, X_i) \sim N \left(\frac{1}{1+\kappa} X_i + \frac{\kappa}{1+\kappa} \mu, \frac{\sigma_i^2}{\kappa+1} \right), \quad (2)$$

$$1/\sigma_i^2 | S_i \sim \Gamma \left(\frac{m+\nu}{2}, \frac{S_i + \nu A}{2} \right). \quad (3)$$

It follows that the Bayes estimators of θ_i 's and σ_i^2 's are

$$\hat{\theta}_i = \frac{1}{1+\kappa} X_i + \frac{\kappa}{1+\kappa} \mu, \quad (4)$$

$$\widehat{\sigma_i^2} = \frac{m}{m+\nu-2} \cdot \frac{S_i}{m} + \frac{\nu-2}{m+\nu-2} \cdot \frac{\nu A}{\nu-2}. \quad (5)$$

Since the hyper-parameters γ and μ in (4)-(5) are typically unknown, they have to be estimated. One possibility is via the empirical Bayes approach, giving $(\hat{\gamma}^{ML}, \hat{\mu}^{ML}) =$

$\arg \max_{\gamma, \mu} f(\mathbf{x})$, where $f(\mathbf{x})$ is the marginal density of $\mathbf{X} = (X_1, X_2, \dots, X_p)$:

$$f(\mathbf{x}) = \prod_i \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \sqrt{\frac{\kappa/(1+\kappa)}{A}} \left[1 + \frac{1}{\nu} \left(\frac{x_i - \mu}{\sqrt{\frac{A}{\kappa/(1+\kappa)}}} \right)^2 \right]^{-\frac{\nu+1}{2}}. \quad (6)$$

In this paper, we consider some alternative approach to the above empirical Bayes method, based on Stein's unbiased risk estimation (SURE) theory (Stein, 1973, 1981). Given an estimator $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\sigma}}^2)$ of $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$, the sum of squared loss is given by

$$\begin{aligned} l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) &= \frac{1}{p} \sum_{i=1}^p [(\hat{\theta}_i - \theta_i)^2 + (\hat{\sigma}_i^2 - \sigma_i^2)^2] \\ &:= l_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) + l_{2,p}(\boldsymbol{\sigma}^2, \hat{\boldsymbol{\sigma}}^2), \end{aligned} \quad (7)$$

and the risk of $\hat{\boldsymbol{\eta}}$ is given by

$$\begin{aligned} R_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) &= \frac{1}{p} \sum_{i=1}^p E[(\hat{\theta}_i - \theta_i)^2] + \frac{1}{p} \sum_{i=1}^p E[(\hat{\sigma}_i^2 - \sigma_i^2)^2] \\ &:= R_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) + R_{2,p}(\boldsymbol{\sigma}^2, \hat{\boldsymbol{\sigma}}^2). \end{aligned} \quad (8)$$

Thus, for the Bayes estimators (4) and (5), we have

$$\begin{aligned} R_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) &= p^{-1} \sum_{i=1}^p E \left[\left(\frac{1}{1+\kappa} X_i + \frac{\kappa}{1+\kappa} \mu - \theta_i \right)^2 + \left(\frac{1}{m+\nu-2} (S_i + \nu A) - \sigma_i^2 \right)^2 \right] \\ &= p^{-1} \sum_{i=1}^p \left[\frac{1}{(1+\kappa)^2} \sigma_i^2 + \frac{\kappa^2}{(1+\kappa)^2} (\mu^2 + \theta_i^2 - 2\mu\theta_i) \right] \\ &\quad + p^{-1} \sum_{i=1}^p \left[\frac{2m}{(m+\nu-2)^2} \sigma_i^4 + \left(\frac{\nu-2}{m+\nu-2} \right)^2 \left(\frac{\nu A}{\nu-2} - \sigma_i^2 \right)^2 \right]. \end{aligned}$$

Replacing $\theta_i, \theta_i^2, \sigma_i^2, \sigma_i^4$ with their corresponding unbiased estimators, $X_i, X_i^2 - S_i/m, S_i/m, S_i^2/(m(m+2))$, we obtain an unbiased estimator for this risk

$$\begin{aligned} \text{SURE}(\gamma, \mu) &= \frac{1}{p} \sum_{i=1}^p \left[\frac{1}{(1+\kappa)^2} \frac{S_i}{m} + \frac{\kappa^2}{(1+\kappa)^2} \left(\mu^2 + X_i^2 - \frac{S_i}{m} - 2\mu X_i \right) \right] \\ &\quad + \frac{1}{p} \sum_{i=1}^p \left[\frac{1}{(m+\nu-2)^2} \cdot \frac{2}{m+2} \cdot S_i^2 \right. \\ &\quad \left. + \frac{(\nu-2)^2}{(m+\nu-2)^2} \left(\left(\frac{\nu A}{\nu-2} \right)^2 + \frac{S_i^2}{m^2 + 2m} - \frac{2\nu A}{\nu-2} \frac{S_i}{m} \right) \right] \\ &:= \text{SURE}_1(\kappa, \mu) + \text{SURE}_2(\nu, A). \end{aligned} \quad (9)$$

Then (γ, μ) can be estimated by minimizing $\text{SURE}(\gamma, \mu)$, that is,

$$(\hat{\gamma}^S, \hat{\mu}^S) := \arg \min_{\gamma > 0, \mu} \text{SURE}(\gamma, \mu). \quad (10)$$

Plugging (10) back into (4)-(5) leads to the SURE estimator for $\hat{\boldsymbol{\eta}}^S = (\hat{\boldsymbol{\theta}}^S, \widehat{\boldsymbol{\sigma}^2}^S)$, where

$$\hat{\theta}_i^S = \frac{1}{1 + \hat{\kappa}_S} X_i + \frac{\hat{\kappa}_S}{1 + \hat{\kappa}_S} \hat{\mu}_S, \quad (11)$$

$$\widehat{\sigma_i^2}^S = \frac{m}{m + \hat{\nu}_S - 2} \cdot \frac{S_i}{m} + \frac{\hat{\nu}_S - 2}{m + \hat{\nu}_S - 2} \cdot \frac{\hat{\nu}_S \hat{A}_S}{\hat{\nu}_S - 2}. \quad (12)$$

We now discuss the optimality properties of the SURE estimator $\hat{\boldsymbol{\eta}}^S = (\hat{\boldsymbol{\theta}}^S, \widehat{\boldsymbol{\sigma}^2}^S)$. We show that the proposed SURE estimator is optimal asymptotically. Our first result shows that $\text{SURE}(\gamma, \mu)$ gives very good approximates to $l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})$.

Theorem 1. For $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\sigma}^2})$ given in (4)-(5), we have

$$\sup_{\gamma > 0, |\mu| \leq \max_i |X_i|} |\text{SURE}(\gamma, \mu) - l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})| \xrightarrow{p} 0, \quad \text{as } p \rightarrow \infty.$$

In Theorem 1, the shrinkage location μ is restricted to $|\mu| < \max_i |X_i|$, which is a technical restriction for ease of presentation, and can be easily satisfied, since in practical situation, a shrinkage estimator would not tend to shrink toward a location outside of the range of the data. We refer to Xie, Kou and Brown (2012) for more discussions.

The next theorem shows that our SURE estimator is asymptotically nearly as good as the oracle loss estimator.

Theorem 2. For any estimator $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\sigma}^2})$ given by (4)-(5), where $\gamma > 0$, we have

$$\lim_{p \rightarrow \infty} P(l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^S) \geq l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + \epsilon) = 0, \quad \text{for any } \epsilon > 0$$

and

$$\lim_{p \rightarrow \infty} \sup [R_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^S) - R_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})] \leq 0.$$

3 SURE estimate by shrinking means toward the grand mean

The previous section states the SURE approach that shrinking the mean estimators toward a data driven location μ (see Equation (11)). At a practical level, one common technique is

to center the data by subtracting the grand mean $\bar{X} = p^{-1} \sum_{i=1}^p X_i$ from each observation X_i . Thus, we consider the mean estimators that shrink toward the grand mean, that is,

$$\hat{\theta}_i^G = \frac{1}{1+\kappa} X_i + \frac{\kappa}{1+\kappa} \bar{X}, \quad i = 1, \dots, p,$$

while keeping the estimator $\widehat{\sigma}_i^2$ the same as in (5). Simple algebras yield that

$$p^{-1} \sum_{i=1}^p E(\hat{\theta}_i^G - \theta_i)^2 = p^{-1} \sum_{i=1}^p \left(\left(\frac{1}{(1+\kappa)^2} + \frac{2\kappa}{(1+\kappa)^2} \frac{1}{p} \right) \sigma_i^2 + \frac{\kappa^2}{(1+\kappa)^2} E(\bar{X} - \theta_i)^2 \right).$$

Thus, an unbiased estimator of the risk for $\hat{\theta}_i^G$ is given by

$$\begin{aligned} & \text{SURE}_1^G(\gamma, \bar{X}) \\ &= p^{-1} \sum_{i=1}^p \left[\left(\frac{1}{(1+\kappa)^2} + \frac{2\kappa}{(1+\kappa)^2} \frac{1}{p} \right) \frac{S_i}{m} + \frac{\kappa^2}{(1+\kappa)^2} \left(\bar{X}^2 + X_i^2 - \frac{S_i}{m} - 2X_i\bar{X} + \frac{2}{p} \frac{S_i}{m} \right) \right], \end{aligned}$$

Consequently, an unbiased estimator for the risk is given by

$$\text{SURE}^G(\gamma, \bar{X}) = \text{SURE}_1^G(\kappa, \bar{X}) + \text{SURE}_2(\nu, A),$$

where $\text{SURE}_2(\nu, A)$ is given in (9). Then $\gamma = (\kappa, \nu, A)$ can be estimated by minimizing $\text{SURE}^G(\gamma, \bar{X})$, i.e.,

$$\hat{\gamma}^G := (\hat{\kappa}^G, \hat{\nu}^G, \hat{A}^G) = \arg \min_{\gamma > \mathbf{0}} \text{SURE}^G(\gamma, \bar{X}).$$

Plugging this back into (4)-(5) leads to the SURE estimator for $\hat{\eta}^G = (\hat{\theta}^G, \widehat{\sigma^2}^G)$:

$$\hat{\theta}_i^G = \frac{1}{1+\hat{\kappa}^G} X_i + \frac{\hat{\kappa}^G}{1+\hat{\kappa}^G} \bar{X}, \quad (13)$$

$$\widehat{\sigma}_i^2 = \frac{m}{m+\hat{\nu}_G-2} \cdot \frac{S_i}{m} + \frac{\hat{\nu}_G-2}{m+\hat{\nu}_G-2} \cdot \frac{\hat{\nu}_G \hat{A}_G}{\hat{\nu}_G-2}. \quad (14)$$

We have similar results to those in Theorems 1-2 above, as given below.

Theorem 3. For $\hat{\eta}^G = (\hat{\theta}^G, \widehat{\sigma^2}^G)$ we have

$$\sup_{\gamma > \mathbf{0}} |\text{SURE}^G(\gamma, \bar{X}) - l_p(\eta, \hat{\eta}^G)| \xrightarrow{p} 0, \quad \text{as } p \rightarrow \infty.$$

Theorem 4. For $\hat{\eta}^G = (\hat{\theta}^G, \widehat{\sigma^2}^G)$ and $\hat{\eta}^G$, we have

$$\lim_{p \rightarrow \infty} P(l_p(\eta, \hat{\eta}^G) \geq l_p(\eta, \hat{\eta}^G) + \epsilon) = 0, \quad \forall \epsilon > 0, \quad (15)$$

and

$$\limsup_{p \rightarrow \infty} [R_p(\eta, \hat{\eta}^G) - R_p(\eta, \hat{\eta}^G)] \leq 0.$$

4 Semi-parametric SURE shrinkage estimation

It is worthwhile to have some intuition about the form of the parametric empirical Bayes estimators in Sections 2-3: the SURE shrinkage estimators have linear form and the shrinkage factor, say, $1/(1 + \kappa)$, is decreasing in the hyper-parameter κ . This means that as κ increases, conditional on the population variance σ_i^2 , the uncertainty about θ_i is reduced, and we are more certain where the posterior θ_i is. In this section, we extend the proposed SURE estimators to a larger setting, let us consider the following hierarchical model

$$\begin{aligned}\kappa &\sim \mathcal{P}(\kappa), \\ 1/\sigma_i^2 &\stackrel{iid}{\sim} \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}A\right), \quad \theta_i|\sigma_i^2 \stackrel{ind}{\sim} N(\mu, \sigma_i^2/\kappa), \\ X_i|(\theta_i, \sigma_i^2) &\stackrel{ind}{\sim} N(\theta_i, \sigma_i^2), \quad S_i|\sigma_i^2 \stackrel{ind}{\sim} \sigma_i^2 \chi_m^2,\end{aligned}\tag{16}$$

where $\mathcal{P}$ is an unspecified hyperprior on κ . The posterior means of θ_i and σ_i^2 (if exist) are given by

$$\begin{aligned}E(\theta_i|\mathbf{X}) &= E\left(\frac{1}{1+\kappa}|\mathbf{X}\right) X_i + E\left(\frac{\kappa}{1+\kappa}|\mathbf{X}\right) \mu. \\ E(\sigma_i^2|\mathbf{X}) &= \frac{m}{m+\nu-2} \cdot \frac{S_i}{m} + \frac{\nu-2}{m+\nu-2} \cdot \frac{\nu A}{\nu-2}\end{aligned}$$

Note that $E(\frac{1}{1+\kappa}|\mathbf{X})$ plays the same role as the shrinkage factor for mean of the i th component, this motivates us to consider the general shrinkage estimators (denoted by $\hat{\boldsymbol{\eta}}^{sm}$) that have the following form

$$\begin{aligned}\hat{\theta}_i^{sm} &= (1 - b_1)X_i + b_1\mu, \\ \widehat{\sigma}_i^{sm} &= (1 - b_2)\frac{S_i}{m} + b_2\sigma_0^2,\end{aligned}\tag{17}$$

where $b_i \in [0, 1]$. We shall keep the same terminology of Xie, Kou and Brown (2012), and called the estimators (17) as the semi-parametric SURE shrinkage estimators in the sense that we do not assume parametric form for b_1 and b_2 . After some calculations, an unbiased estimator of the risk is

$$\begin{aligned}&\text{SURE}^{SM}(\mathbf{b}, \mu, \sigma_0^2) \\ &= p^{-1} \sum_{i=1}^p \left\{ \left[(1 - 2b_1)\frac{S_i}{m} + b_1^2(X_i - \mu)^2 \right] + \left[(1 - 2b_2)\frac{2}{m} \frac{S_i^2}{m^2 + 2m} + b_2^2 \left(\sigma_0^2 - \frac{S_i}{m} \right)^2 \right] \right\} \\ &:= \text{SURE}_1^{SM}(b_1, \mu) + \text{SURE}_2^{SM}(b_2, \sigma_0^2).\end{aligned}$$

Then

$$\hat{\theta}_i^{SM} = (1 - \hat{b}_1^{SM})X_i + \hat{b}_1^{SM}\hat{\mu}^{SM}, \quad (18)$$

$$\widehat{\sigma}_i^{2SM} = (1 - \hat{b}_2^{SM})\frac{S_i}{m} + \hat{b}_2^{SM}\widehat{\sigma}_0^{2SM}, \quad (19)$$

where

$$\begin{aligned} (\hat{\mathbf{b}}^{SM}, \hat{\mu}^{SM}, \widehat{\sigma}_0^{2SM}) &= \arg \min \text{SURE}^{SM}(\mathbf{b}, \mu, \sigma_0^2), \\ &\text{subject to } b_1, b_2 \in [0, 1], \mu \leq \max_{1 \leq i \leq p} |X_i|. \end{aligned}$$

Likewise, we can investigate the estimators **by shrinking the mean estimators** toward the grand mean,

$$\begin{aligned} \hat{\theta}_i^{sg} &= (1 - b_1)X_i + b_1\bar{X}, \\ \widehat{\sigma}_i^{2sg} &= (1 - b_2)\frac{S_i}{m} + b_2\sigma_0^2, \end{aligned}$$

where $b_i \in [0, 1]$. An unbiased estimator of the risk is

$$\begin{aligned} &\text{SURE}^{SG}(\mathbf{b}, \sigma_0^2) \\ &= p^{-1} \sum_{i=1}^p \left\{ \left[(1 - 2(1 - \frac{1}{p})b_1)\frac{S_i}{m} + b_1^2(X_i - \bar{X})^2 \right] + \left[(1 - 2b_2)\frac{2}{m}\frac{S_i^2}{m^2 + 2m} + b_2^2 \left(\sigma_0^2 - \frac{S_i}{m} \right)^2 \right] \right\}. \end{aligned}$$

Then

$$\hat{\theta}_i^{SG} = (1 - \hat{b}_1^{SG})X_i + \hat{b}_1^{SG}\bar{X}, \quad (20)$$

$$\widehat{\sigma}_i^{2SG} = (1 - \hat{b}_2^{SG})\frac{S_i}{m} + \hat{b}_2^{SG}\widehat{\sigma}_0^{2SG}, \quad (21)$$

where

$$\begin{aligned} (\hat{\mathbf{b}}^{SG}, \widehat{\sigma}_0^{2SG}) &= \arg \min \text{SURE}^{SG}(\mathbf{b}, \sigma_0^2), \\ &\text{subject to } b_1, b_2 \in [0, 1], \mu \leq \max_i |X_i|. \end{aligned}$$

The theoretical properties of the semi-parametric SURE shrinkage estimators are summarized as follows. We only present the results for $\hat{\boldsymbol{\eta}}^{SM}$. The asymptotic properties of $\hat{\boldsymbol{\eta}}^{SG}$ can be shown in the same manner.

Theorem 5. For $\hat{\boldsymbol{\eta}}^{sm}$ give by (17), we have

$$\sup_{b_i \in [0, 1], \sigma_0^2 > 0, \mu \leq \max_i |X_i|} |\text{SURE}^{SM}(\mathbf{b}, \mu, \sigma_0^2) - l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^{sm})| \xrightarrow{p} 0.$$

Theorem 6. For any shrinkage estimator $\hat{\boldsymbol{\eta}}^{b,\mu,\sigma_0^2}$ with

$$\begin{aligned}\hat{\theta}_i^{b,\mu,\sigma_0^2} &= (1 - \hat{b}_1)X_i + \hat{b}_1\hat{\mu} \\ \widehat{\sigma}_i^{2,b,\mu,\sigma_0^2} &= (1 - \hat{b}_2)\frac{S_i}{m} + \hat{b}_2\hat{\sigma}_0^2,\end{aligned}$$

where $b_i \in [0, 1]$, $|\mu| \leq \max_i |X_i|$ and $\sigma_0^2 > 0$, we have for any $\epsilon > 0$,

$$\lim_{p \rightarrow \infty} P(l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^{SM}) \geq l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^{b,\mu,\sigma_0^2}) + \epsilon) = 0$$

and

$$\limsup_{p \rightarrow \infty} \left[R(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^{SM}) - R(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^{b,\mu,\sigma_0^2}) \right] \leq 0.$$

5 Simulation

In this section, we demonstrate the performance of our proposed shrinkage estimators. We refer to these estimators as SUREG.New, SUREM.New (see Equations (11)-(14)), SURESM.New, SURESG.New (see Equations (18)-(21)). EBMLE denotes the MLE estimators, SURE0.XKB, SUREM.XKB and SUREG.XKB denotes the SURE estimators proposed in Xie, Kou and Brown (2012), which shrink to preset zero, general μ , and grand mean, respectively. The EBMLE estimator used here are defined by

$$\hat{\theta}_i^{MLE} = \frac{1}{1 + \hat{\kappa}_{MLE}}X_i + \frac{\hat{\kappa}_{MLE}}{1 + \hat{\kappa}_{MLE}}\hat{\mu}_{MLE}, \quad (22)$$

$$\widehat{\sigma}_i^{2,MLE} = \frac{m}{m + \hat{\nu}_{MLE} - 2} \cdot \frac{S_i}{m} + \frac{\hat{\nu}_{MLE} - 2}{m + \hat{\nu}_{MLE} - 2} \cdot \frac{\hat{\nu}_{MLE} \cdot \hat{A}_{MLE}}{\hat{\nu}_{MLE} - 2}, \quad (23)$$

where $(\hat{\kappa}_{MLE}, \hat{\nu}_{MLE}, \hat{A}_{MLE})$ can be obtained by maximizing the marginal density (6).

The SUREM.XKB is given by

$$\hat{\theta}_i^{XKB,M} = \frac{\hat{\lambda}_{XKB,M}}{A_i + \hat{\lambda}_{XKB,M}}X_i + \frac{A_i}{A_i + \hat{\lambda}_{XKB,M}}\hat{\mu}_{XKB,M}, \quad (24)$$

where

$$(\hat{\lambda}_{XKB,M}, \hat{\mu}_{XKB,M}) = \arg \min_{\lambda \geq 0, \mu} \frac{1}{p} \sum_i \frac{A_i}{(A_i + \lambda)^2} (A_i(X_i - \mu)^2 + \lambda^2 - A_i^2).$$

SURE0.XKB can be obtained by putting $\mu = 0$ in (24), that is

$$\hat{\theta}_i^{XKB,0} = \frac{\hat{\lambda}_{XKB,0}}{A_i + \hat{\lambda}_{XKB,0}}X_i, \quad (25)$$

where

$$\hat{\lambda}_{XKB,0} = \arg \min_{\lambda \geq 0} \frac{1}{p} \sum_i \frac{A_i}{(A_i + \lambda)^2} (A_i X_i^2 + \lambda^2 - A_i^2).$$

SUREG.XKB is given by

$$\hat{\theta}_i^{XKB,G} = \frac{\hat{\lambda}_{XKB,G}}{A_i + \hat{\lambda}_{XKB,G}} X_i + \frac{A_i}{A_i + \hat{\lambda}_{XKB,G}} \bar{X}, \quad (26)$$

where

$$\hat{\lambda}_{XKB,G} = \arg \min_{\lambda \geq 0} \frac{1}{p} \sum_{i=1}^p \left[\frac{A_i^2}{(A_i + \lambda)^2} (X_i - \bar{X})^2 + \frac{A_i}{A_i + \lambda} (\lambda - A_i + \frac{2}{p} A_i) \right].$$

The variances A_i 's are estimated from the sample variance of each $\mathcal{X}_i^0 = \{X_{1,i}^0, \dots, X_{m+1,i}^0\}$, i.e, $\hat{A}_i = S_i/m$ (see below).

For each case, the sample is generated as follows. For each $i = 1, \dots, p$, we first draw σ_i^2 , independently from the inverse gamma distribution: $\text{Inverse-}\Gamma(\nu/2, \nu A/2)$, and then draw θ_i from $N(0, \sigma_i^2/\kappa)$. Finally, we draw $\mathcal{X}_i^0 = \{X_{1,i}^0, \dots, X_{m+1,i}^0\}$ from $N(\theta_i, (m+1)\sigma_i^2)$. That is, the sample is

$$\mathcal{X}^0 = \begin{pmatrix} X_{1,1}^0 & X_{1,2}^0 & \cdots & X_{1,p}^0 \\ X_{2,1}^0 & X_{2,2}^0 & \cdots & X_{2,p}^0 \\ \vdots & \vdots & \ddots & \vdots \\ X_{m+1,1}^0 & X_{m+1,2}^0 & \cdots & X_{m+1,p}^0 \end{pmatrix}.$$

Thus, we have our observations

$$X_i = \frac{1}{m+1} \sum_{j=1}^{m+1} X_{j,i}^0, \quad S_i = \frac{1}{m+1} \sum_{j=1}^{m+1} (X_{j,i}^0 - X_i)^2, \quad i = 1, \dots, p.$$

We repeat this process 5000 times to obtain an average risk for each estimator, and the sample size p is chosen to increase from 20 to 500 at an step of length 20. For varied p , we investigate several cases with different (m, κ, ν, A) combinations, and evaluate the impact of these hyperparameters to the risk properties. Figure 1-3 reports the average risks of different shrinkage estimators for the mean vector θ , i.e., $R(\theta, \hat{\theta})$. make From our simulation experiment, we have the following findings.

- (i) The SURE shrinkage estimators that shrink both the means and the variances substantially improve the existing estimators in terms of the average risk;
- (ii) When κ and ν become large, our SURE shrinkage estimators result in lower risk than all the other estimators, and approach to the oracle risk. When κ become small, the risks of our SURE shrinkage estimators could be smaller than the oracle risk;
- (iii) For large κ (e.g., $\kappa = 5$), SURE type estimators that shrink only the means does not always outperform its competitors, EBMLE, but remains competitive. This is because θ_i and σ_i are dependent under this setup, large κ increases the certainty of θ_i 's.

Figure 1: *The average risk plots of different shrinkage estimators for various (m, κ, ν, A) .*

Figure 2: *The average risk plots of different shrinkage estimators for various (m, κ, ν, A) .*

Figure 3: The average risk plots of different shrinkage estimators for various (m, κ, ν, A) .

6 Real Data Analysis

6.1 Application to Baseball Data

In this section, we apply the proposed shrinkage estimators to the baseball data in Brown (2008), i.e., the batting records for all the Major League Baseball players in the season of 2005. First, we carry out the necessary preprocessing steps as in Brown (2008) and Xie, Kou and Brown (2012), say, removing the players whose numbers of at-bats are less than 11. To study the performance of the proposed estimator, we split the data set into training and test sets. Particularly, we let the training set range from month 4 to 6, and the others serve as the test set. Write N and H as the number of at-bats and the successful number of batting for each player, respectively. Then $H_{ij} \sim \text{Binomial}(N_{ij}, p_j)$, where $i = 1, 2$ indicate the training set and test set, respectively, $j = 1, \dots, p$ indicates the j th player. Before proceeding, we apply the following variance-stablizing transformation

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

So, $X_{ij} \sim N(\theta_j, \frac{1}{4N_{ij}})$, $\theta_j = \arcsin(p_j)$. Let $\bar{X}_{1j}$, $\bar{X}_{2j}$, $j = 1, \dots, p$ denote the sample means of the training and test sets, respectively. We use TSE, introduced in Brown (2008) as the error measurement for comparison, that is

$$\text{TSE}(\hat{\theta}) = \sum_j (\bar{X}_{2j} - \hat{\theta}_j)^2 - \sum_j \text{Var}(\bar{X}_{2j}). \quad (27)$$

We apply the proposed shrinkage estimators to all the baseball players, the pitchers only, and the non-pitchers only, respectively. Table 1 reports the ratios of the error of a given estimator to that of the benchmark naive estimator, which is sample mean of the training set. In the table, SUREM.XKB and SUREG.XKB denote the estimators (24) and (26), respectively, SURESG.XKB, and SURESM.XKB represents respectively, the semiparametric SURE shrinkage estimators proposed by Xie, Kou and Brown (2012), the rest are the estimators corresponding to (11)-(14), (18)-(21). For comparison, we compute the TSE of the mean vector θ , $\text{TSE}(\hat{\theta})$ (see (27)) for each estimators. Note that for SUREM.XKB, SUREG.XKB, SURESM.XKB and SURESG.XKB,

$$\text{TSE}(\hat{\theta}) = \sum_j (\bar{X}_{2j} - \hat{\theta}_j)^2 - \sum_j \frac{1}{4N_{2j}},$$

while for SUREM.New, SUREG.New, SURESM.New, and SURESG.New,

$$\text{TSE}(\hat{\theta}) = \sum_j (\bar{X}_{2j} - \hat{\theta}_j)^2 - \sum_j \hat{\sigma}_j^2,$$

where $\hat{\sigma}_j^2$ is the estimate of variance corresponding to the above four estimators. It can be seen, for the groups “All” and “non-pitchers”, according to the “TSE” standard, the SURE estimators that shrink both the means and the variances outperform the SURE estimators that only shrink the means.

Table 1. *Prediction errors of batting averages.*

	All	Pitchers	NonPitchers
Naive	1	1	1
SUREG.XKB	0.505	0.123	0.278
SUREM.XKB	0.422	0.123	0.282
SURESG.XKB	0.409	0.081	0.261
SURESM.XKB	0.419	0.077	0.278
SUREG.New	0.263	0.131	0.079
SUREM.New	0.263	0.133	0.067
SURESG.New	0.298	0.147	0.184
SURESM.New	0.298	0.134	0.184

6.2 Application to Gene Expression Data

We apply the proposed double SURE shrinkage estimators, along with other estimators to an leukemia data which includes the expression level of 5000 genes of 27 acute lymphoblastic leukemia (ALL) patients. The data can be downloaded from the Broad Institute (<http://www.broadinstitute.org/cgi-cgibin/cancer/datasets.cgi>). After taking the logarithm transformation to better approximate the normality assumption, we standardize the expression levels of 27 observations for each gene, and then calculate the t-statistics for each gene. We further delete those genes with sample variance approximating to 0. Our ultimate

data set consists of $p = 4730$, $m + 1 = 27$ adjusted gene expression, i.e.,

$$\mathcal{X}^* = \begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m+1,1} & X_{m+1,2} & \cdots & X_{m+1,p} \end{pmatrix}.$$

Thus, we have the new sample

$$X_{li} = \frac{1}{m_l + 1} \sum_{j=1}^{m_l+1} X_{j,i}, \quad S_{li} = \frac{1}{m_l + 1} \sum_{j=1}^{m_l+1} (X_{j,i} - X_i)^2, \quad i = 1, \dots, 4730,$$

where $l = 1, 2$, m_l is the survey sample size, and $m_1 + m_2 + 1 = m$.

First, we are interested in estimating the average gene expression level for each genes, we will calculate the the average risks of the estimators: the naive estimator (the sample mean of each gene), SUREM.XKB, SUREG.XKB, SUREM.New, SUREG.New. For each gene, we randomly draw a 10%, 20% and 50% of the entire data, and then estimate the true mean by different shrinkage estimators, we run the simulation 2000 times and report the average risk in Table 2. For SUREM.XKB and SUREG.XKB, the variances A_i 's are estimated from the sample variance of each gene.

Second, we confine our interest to the prediction the average of the gene expression level. Like the baseball dataset, the ALL sample is randomly splitted into two parts, one is the training set X_1 with 19 ALL sample, the other is the test set X_2 with the remaining 8 ALL sample. Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means of the training and test sets, respectively. We try to predict the average gene expression level $\bar{X}_2$ using $\bar{X}_1$. The results are shown in Tables 3. The prediction performance for the five methods is measured by the median of absolute prediction errors $\{|\hat{\theta}_i - \bar{X}_{2i}|\}$ (MPE) and the median of squared prediction errors $\{(\hat{\theta}_i - \bar{X}_{2i})^2\}$ (MSPE), where $\hat{\theta}_i$, $i = 1, \dots, 4730$. Table 3 shows that the proposed double shrinkage SURE methods stand out among its competitors in terms of MPE and MSPE.

Table 2. *Estimation errors of average of the gene expression level.*

	15%	30%	50%
Naive	0.213	0.088	0.040
SUREG.XKB	0.087	0.030	0.013
SUREM.XKB	0.108	0.037	0.016
SUREG.New	0.010	0.004	0.002
SUREM.New	0.022	0.021	0.017

Table 3. *Comparisons of median prediction errors with different methods.*

	Naive	SUREM.XKB	SUREG.XKB	SUREM.New	SUREG.New
MPE	0.300	0.255	0.255	0.217	0.219
MSPE	0.091	0.066	0.066	0.048	0.048

7 Proofs

For ease of presentation, we list some notations and facts below.

$$\begin{aligned} \bar{\theta}_p &= \frac{1}{p} \sum_{i=1}^p \theta_i, \quad \bar{\sigma}^2 = \frac{1}{p} \sum_{i=1}^p \sigma_i^2, \quad \bar{S} = \frac{1}{p} \sum_{i=1}^p \frac{S_i}{m}. \\ \left\{ \begin{array}{l} E(S_i) = m\sigma_i^2, \quad E(S_i^2) = m(m+2)\sigma_i^4 \\ E(S_i^3) = m(m+2)(m+4)\sigma_i^6 \\ E(S_i^4) = m(m+2)(m+4)(m+6)\sigma_i^8. \end{array} \right. \end{aligned} \quad (28)$$

Lemma 1. (Xie, Kou and Brown, 2012) Under Conditions (C1)-(C3), we have $E(\max_{1 \leq i \leq p} X_i^2) = O(p^{2/(2+\delta^*)})$, where $\delta^* = \min(1, \delta)$.

Proof of Theorem 1 Since

$$\begin{aligned} & \sup_{\gamma > 0, |\mu| \leq \max_i |X_i|} |\text{SURE}(\gamma, \mu) - l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})| \\ & \leq \sup_{\kappa > 0, |\mu| \leq \max_i |X_i|} |\text{SURE}_1(\kappa, \mu) - l_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})| + \sup_{\nu, A > 0} |\text{SURE}_2(\nu, A) - l_{2,p}(\boldsymbol{\sigma}^2, \hat{\boldsymbol{\sigma}}^2)|, \end{aligned} \quad (29)$$

we just need to prove the L^2 convergence of the two terms separately. For the first term,

$$\begin{aligned} & |\text{SURE}_1(\kappa, \mu) - l_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})| \\ &= \left| \frac{1}{p} \sum_{i=1}^p \left[\left(\frac{1}{(1+\kappa)^2} \frac{S_i}{m} + \frac{\kappa^2}{(1+\kappa)^2} (\mu^2 + X_i^2 - \frac{S_i}{m} - 2\mu X_i) \right) \right. \right. \\ & \quad \left. \left. - \left(\frac{1}{(1+\kappa)^2} (X_i - \theta_i)^2 + \frac{\kappa^2}{(1+\kappa)^2} (\mu - \theta_i)^2 + \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i)(\mu - \theta_i) \right) \right] \right| \\ & \leq \left| \frac{1}{p} \sum_{i=1}^p \frac{1}{(1+\kappa)^2} \left(\frac{S_i}{m} - (X_i - \theta_i)^2 \right) \right| + \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i)(\mu - \theta_i) \right| \\ & \quad + \left| \frac{1}{p} \sum_{i=1}^p \frac{\kappa^2}{(1+\kappa)^2} \left(\mu^2 + X_i^2 - \frac{S_i}{m} - 2\mu X_i - (\mu - \theta_i)^2 \right) \right|. \end{aligned} \quad (30)$$

From (C1), we have

$$\begin{aligned} & E \left[\frac{1}{p} \sum_{i=1}^p \frac{1}{(1+\kappa)^2} \left(\frac{S_i}{m} - (X_i - \theta_i)^2 \right) \right]^2 = \frac{1}{p(1+\kappa)^4} \frac{1}{p} \sum_{i=1}^p E \left[\left(\frac{S_i}{m} - (X_i - \theta_i)^2 \right)^2 \right] \\ &= \frac{1}{p(1+\kappa)^4} \frac{1}{p} \sum_{i=1}^p E \left[\frac{S_i^2}{m^2} + (X_i - \theta_i)^4 - \frac{2S_i}{m} (X_i - \theta_i)^2 \right] = \frac{2}{p(1+\kappa)^4} \frac{1}{p} \sum_{i=1}^p \frac{m+1}{m} \sigma_i^4 \rightarrow 0, \end{aligned}$$

which together with Kolmogorov's inequality yields

$$\sup_{\kappa > 0} \left| \frac{1}{p} \sum_{i=1}^p \frac{1}{(1+\kappa)^2} \left[\frac{S_i}{m} - (X_i - \theta_i)^2 \right] \right| \xrightarrow{p} 0, \text{ as } p \rightarrow \infty. \quad (31)$$

Next,

$$\begin{aligned}
& \sup_{\kappa>0, |\mu| \leq \max_i |X_i|} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i)(\mu - \theta_i) \right| \\
& \leq \sup_{\kappa>0, |\mu| \leq \max_i |X_i|} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} \mu (X_i - \theta_i) \right| + \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i) \theta_i \right| \\
& \leq \max_{1 \leq i \leq p} |X_i| \cdot \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i) \right| + \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i) \theta_i \right|.
\end{aligned}$$

Now, it follows from Lemma 1 that $E(\max_{1 \leq i \leq p} X_i^2) = O(p^{2/(2+\delta^*)})$, where $\delta^* = \min(1, \delta)$, and by (C1), we have

$$E \left[\sup_{\kappa>0} \left(\sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i) \right)^2 \right] = \frac{1}{4} \sum_{i=1}^p E(X_i - \theta_i)^2 = O(p).$$

So the Cauchy-Schwartz inequality implies that

$$\begin{aligned}
& \frac{1}{p} E \left(\max_{1 \leq i \leq p} |X_i| \sup_{\kappa>0} \left| \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i) \right| \right) \\
& \leq \frac{1}{p} \left(E(\max_{1 \leq i \leq p} X_i^2) \cdot E \left\{ \sup_{\kappa>0} \left(\sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i) \right)^2 \right\} \right)^{1/2} \\
& = O(p^{-\frac{\delta^*}{2(2+\delta^*)}}) = o(1).
\end{aligned}$$

From (C2), we have

$$E \left[\left(\frac{1}{p} \sum_{i=1}^p (X_i - \theta_i) \theta_i \right)^2 \right] = \frac{1}{p^2} \sum_{i=1}^p \sigma_i^2 \theta_i^2 \rightarrow 0, \quad \text{as } p \rightarrow \infty.$$

Thus, we have

$$\sup_{\kappa>0, |\mu| \leq \max_i |X_i|} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i)(\mu - \theta_i) \right| \xrightarrow{p} 0, \quad \text{as } p \rightarrow 0. \quad (32)$$

For the last term, notice that

$$\begin{aligned}
& \sup_{\kappa>0, |\mu| \leq \max_i |X_i|} \left| \frac{1}{p} \sum_{i=1}^p \frac{\kappa^2}{(1+\kappa)^2} \left[\mu^2 + X_i^2 - \frac{S_i}{m} - 2\mu X_i - (\mu - \theta_i)^2 \right] \right| \\
& \leq \left\{ \frac{1}{p} \left| \sum_{i=1}^p (X_i^2 - \frac{S_i}{m} - \theta_i^2) \right| + \frac{2}{p} \max_{1 \leq i \leq n} |X_i| \cdot \left| \sum_{i=1}^p (X_i - \theta_i) \right| \right\}.
\end{aligned}$$

From Lemma 1, (C1) and (C2),

$$E \left[\left(\frac{1}{p} \sum_{i=1}^p (X_i^2 - \frac{S_i}{m} - \theta_i^2) \right)^2 \right] = \frac{1}{p^2} \sum_{i=1}^p \left(4\theta_i^2 \sigma_i^2 + \frac{2m+2}{m} \sigma_i^4 \right) \rightarrow 0,$$

and

$$\frac{2}{p} \max_{1 \leq i \leq n} |X_i| \cdot \left| \sum_{i=1}^p (X_i - \theta_i) \right| \xrightarrow{p} 0,$$

which in turn lead to

$$\sup_{\kappa > 0, |\mu| \leq \max_{1 \leq i \leq p} |X_i|} \left| \frac{1}{p} \sum_{i=1}^p \frac{\kappa^2}{(1 + \kappa)^2} \left(\mu^2 + X_i^2 - \frac{S_i}{m} - 2\mu X_i - (\mu - \theta_i)^2 \right) \right| \xrightarrow{p} 0, \quad \text{as } p \rightarrow 0. \quad (33)$$

Therefore, from (30), (31), (32) and (33), we have

$$\sup_{\kappa > 0, |\mu| \leq \max_{1 \leq i \leq p} |X_i|} |\text{SURE}_1(\kappa, \mu) - l_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})| \xrightarrow{p} 0, \quad \text{as } p \rightarrow 0. \quad (34)$$

Now, let's look at the second part of (29). Note that

$$\begin{aligned} & |\text{SURE}_2(\nu, A) - l_{2,p}(\boldsymbol{\sigma}^2, \hat{\boldsymbol{\sigma}}^2)| \\ &= \frac{1}{p} \sum_{i=1}^p \left\{ \left[\frac{1}{(m+\nu-2)^2} \frac{2}{m+2} S_i^2 + \left(\frac{\nu-2}{m+\nu-2} \right)^2 \left(\left(\frac{\nu A}{\nu-2} \right)^2 + \frac{S_i^2}{m^2+2m} - \frac{2\nu A}{\nu-2} \frac{S_i}{m} \right) \right] \right. \\ & \quad \left. - \left[\left(\frac{S_i}{m+\nu-2} + \frac{\nu-2}{m+\nu-2} \frac{\nu A}{\nu-2} - \sigma_i^2 \right)^2 \right] \right\} \\ &\leq \left| \frac{1}{p} \sum_{i=1}^p \frac{1}{(m+\nu-2)^2} [(S_i - m\sigma_i^2)^2 - \frac{2}{m+2} S_i^2] \right| \\ & \quad + \left| \frac{1}{p} \sum_{i=1}^p \left\{ \left(\frac{\nu-2}{m+\nu-2} \right)^2 \left[\left(\frac{\nu A}{\nu-2} - \sigma_i^2 \right)^2 - \left(\frac{\nu A}{\nu-2} \right)^2 - \frac{S_i^2}{m^2+2m} + \frac{2\nu A}{\nu-2} \frac{S_i}{m} \right] \right\} \right| \\ & \quad + \left| \frac{1}{p} \sum_{i=1}^p \frac{2(\nu-2)}{(m+\nu-2)^2} (S_i - m\sigma_i^2) \left(\frac{\nu A}{\nu-2} - \sigma_i^2 \right) \right|. \end{aligned} \quad (35)$$

We deal with each summand in (35) using (28). First,

$$\begin{aligned} & E \left[\sup_{\nu > 2} \left(\frac{1}{p} \sum_{i=1}^p \frac{1}{(m+\nu-2)^2} \left\{ (S_i - m\sigma_i^2)^2 - \frac{2}{m+2} S_i^2 \right\} \right)^2 \right] \\ &= \sup_{\nu > 2} \frac{1}{(m+\nu-2)^4} \frac{1}{p^2} \sum_{i=1}^p E \left[(S_i - m\sigma_i^2)^2 - \frac{2}{m+2} S_i^2 \right]^2 \\ &= \sup_{\nu > 2} \frac{1}{(m+\nu-2)^4} \frac{1}{p^2} \sum_{i=1}^p E \left[\frac{m}{m+2} S_i^2 + m^2 \sigma_i^4 - 2m\sigma_i^2 S_i \right]^2 \\ &= \sup_{\nu > 2} \frac{1}{(m+\nu-2)^4} \left[\frac{m^3}{m+2} (m+4)(m+6) - m^4 - 8m^3 \right] \frac{1}{p^2} \sum_{i=1}^p \sigma_i^8 \rightarrow 0, \end{aligned}$$

where the last convergence follows from (C1). Using Kolmogorov's inequality again, we have

$$\sup_{\nu > 2} \left| \frac{1}{p} \sum_{i=1}^p \frac{1}{(m+\nu-2)^2} \left[(S_i - m\sigma_i^2)^2 - \frac{2}{m+2} S_i^2 \right] \right| \xrightarrow{p} 0, \quad \text{as } p \rightarrow 0. \quad (36)$$

For the second term, note that

$$\begin{aligned} & \left| \frac{1}{p} \sum_{i=1}^p \left\{ \left(\frac{\nu-2}{m+\nu-2} \right)^2 \left[\left(\frac{\nu A}{\nu-2} - \sigma_i^2 \right)^2 - \left(\frac{\nu A}{\nu-2} \right)^2 - \frac{S_i^2}{m^2+2m} + \frac{2\nu A}{\nu-2} \frac{S_i}{m} \right] \right\} \right| \\ &\leq \left| \frac{1}{p} \sum_{i=1}^p \left(\frac{\nu-2}{m+\nu-2} \right)^2 \cdot \left(\sigma_i^4 - \frac{S_i^2}{m^2+2m} \right) \right| + \left| \frac{1}{p} \sum_{i=1}^p \left(\frac{\nu-2}{m+\nu-2} \right)^2 \frac{2\nu A}{\nu-2} \left(\frac{S_i}{m} - \sigma_i^2 \right) \right|, \end{aligned}$$

From (C1),

$$E \left[\left(\frac{1}{p} \sum_{i=1}^p \left(\sigma_i^4 - \frac{S_i^2}{m^2 + 2m} \right) \right)^2 \right] = \frac{1}{p^2} \sum_{i=1}^p \left(\frac{8m + 24}{m(m+2)} \right) \sigma_i^8 \rightarrow 0,$$

$$E \left[\left(\frac{1}{p} \sum_{i=1}^p \left(\frac{S_i}{m} - \sigma_i^2 \right) \right)^2 \right] = \frac{1}{p^2} \sum_{i=1}^p \frac{2\sigma_i^4}{m} \rightarrow 0.$$

which along with Kolmogorov's inequality implies that

$$\sup_{\nu > 2, A > 0} \left| \frac{1}{p} \sum_{i=1}^p \left\{ \left(\frac{\nu - 2}{m + \nu - 2} \right)^2 \left[\left(\frac{\nu A}{\nu - 2} - \sigma_i^2 \right)^2 - \left(\frac{\nu A}{\nu - 2} \right)^2 - \frac{S_i^2}{m^2 + 2m} + \frac{2\nu A}{\nu - 2} \frac{S_i}{m} \right] \right\} \right| \xrightarrow{p} 0. \quad (37)$$

The convergence of the third term can be proved in a similar way, that is,

$$\sup_{\nu > 2, A > 0} \left| \frac{1}{p} \sum_{i=1}^p \frac{2(\nu - 2)}{(m + \nu - 2)^2} (S_i - m\sigma_i^2) \left(\frac{\nu A}{\nu - 2} - \sigma_i^2 \right) \right| \xrightarrow{p} 0, \quad \text{as } p \rightarrow \infty. \quad (38)$$

Now, by (35), (36), (37) and (38), we have

$$\sup_{\nu > 2, A > 0} |\text{SURE}_2(\nu, A) - l_{2,p}(\boldsymbol{\sigma}^2, \hat{\boldsymbol{\sigma}}^2)| \xrightarrow{p} 0, \quad \text{as } p \rightarrow \infty. \quad (39)$$

Thus, we complete the proof from (29), (34) and (39). $\square$

Proof of Theorem 2. Observing that

$$\text{SURE}(\hat{\boldsymbol{\gamma}}^S, \hat{\boldsymbol{\mu}}^S) \leq \text{SURE}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\mu}}),$$

and

$$\sup_{\boldsymbol{\gamma} > \mathbf{0}, |\boldsymbol{\mu}| \leq \max_{1 \leq i \leq p} |X_i|} |\text{SURE}(\boldsymbol{\gamma}, \boldsymbol{\mu}) - l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})| \xrightarrow{p} 0, \quad \text{as } p \rightarrow \infty,$$

we obtain that for any $\epsilon > 0$,

$$P(l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^S) \geq l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + \epsilon) \leq P(l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^S) - \text{SURE}(\hat{\boldsymbol{\gamma}}^S, \hat{\boldsymbol{\mu}}^S) \geq l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) - \text{SURE}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\mu}}) + \epsilon) \rightarrow 0. \quad (40)$$

Hence,

$$\lim_{p \rightarrow \infty} P(l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^S) \geq l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + \epsilon) = 0.$$

For the other inequality, note that

$$l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^S) - l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) \leq 2 \sup_{\boldsymbol{\gamma} > \mathbf{0}, \boldsymbol{\mu} \leq \max_i |X_i|} |\text{SURE}(\boldsymbol{\gamma}, \boldsymbol{\mu}) - l_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})|,$$

which together with Theorem 1 leads to

$$\limsup_{p \rightarrow \infty} [R_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}^S) - R_p(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})] \leq 0.$$

□

Proof of Theorem 3. Note that

$$\begin{aligned} R_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^G) &= \frac{1}{p} \sum_{i=1}^p E[(\hat{\theta}_i^G - \theta_i)^2] \\ &= \frac{1}{p} \sum_{i=1}^p E \left[\left(\frac{1}{1+\kappa} (X_i - \theta_i) + \frac{\kappa}{1+\kappa} (\bar{X} - \bar{\theta}_p) + \frac{\kappa}{1+\kappa} (\bar{\theta}_p - \theta_i) \right)^2 \right] \\ &= \frac{1}{p} \sum_{i=1}^p E \left[\frac{1}{(1+\kappa)^2} (X_i - \theta_i)^2 + \frac{\kappa^2}{(1+\kappa)^2} (\bar{X} - \bar{\theta}_p)^2 + \frac{\kappa^2}{(1+\kappa)^2} (\bar{\theta}_p - \theta_i)^2 \right. \\ &\quad \left. + \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i)(\bar{X} - \bar{\theta}_p) \right] \\ &= \frac{1}{p} \sum_{i=1}^p \left[\frac{1}{(1+\kappa)^2} \sigma_i^2 + \frac{\kappa^2}{(1+\kappa)^2} \frac{1}{p} \bar{\sigma}^2 + \frac{\kappa^2}{(1+\kappa)^2} (\bar{\theta}_p - \theta_i)^2 + \frac{2\kappa}{(1+\kappa)^2} \frac{1}{p} \sigma_i^2 \right]. \end{aligned} \quad (41)$$

An unbiased estimator of $R_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^G)$ is

$$\begin{aligned} \text{SURE}_1^G(\kappa, \bar{X}) &= \frac{1}{p} \sum_{i=1}^p \left[\left(\frac{1}{(1+\kappa)^2} + \frac{2\kappa}{(1+\kappa)^2} \frac{1}{p} \right) \frac{S_i}{m} + \frac{\kappa^2}{(1+\kappa)^2} \frac{1}{p} \bar{S} + \frac{\kappa^2}{(1+\kappa)^2} \left(\bar{X}^2 - \frac{\bar{S}}{p} + X_i^2 - \frac{S_i}{m} - 2X_i \bar{X} + \frac{2}{p} \frac{S_i}{m} \right) \right]. \end{aligned}$$

Consider

$$\begin{aligned} &|\text{SURE}_1^G(\kappa, \bar{X}) - l_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^G)| \\ &\leq \left| \frac{1}{p} \sum_{i=1}^p \frac{1}{(1+\kappa)^2} \left(\frac{S_i}{m} - (X_i - \theta_i)^2 \right) \right| + \left| \frac{1}{p} \sum_{i=1}^p \frac{\kappa^2}{(1+\kappa)^2} \left(\frac{\bar{S}}{p} - (\bar{X} - \bar{\theta}_p)^2 \right) \right| \\ &\quad + \left| \frac{1}{p} \sum_{i=1}^p \frac{\kappa^2}{(1+\kappa)^2} \left(\bar{X}^2 - \frac{\bar{S}}{p} + X_i^2 - \frac{S_i}{m} - 2X_i \bar{X} + \frac{2}{p} \frac{S_i}{m} - (\bar{\theta}_p - \theta_i)^2 \right) \right| \\ &\quad + \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} \left(\frac{1}{p} \frac{S_i}{m} - (X_i - \theta_i)(\bar{X} - \bar{\theta}_p) \right) \right| + \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i)(\bar{\theta}_p - \theta_i) \right| \\ &\quad + \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa^2}{(1+\kappa)^2} (\bar{X} - \bar{\theta}_p)(\bar{\theta}_p - \theta_i) \right| \end{aligned} \quad (42)$$

Following the same technique in the proof of Theorem 1, we have

$$\begin{aligned} &E \left[\sup_{\kappa > 0} \left(\frac{1}{p} \sum_{i=1}^p \frac{1}{(1+\kappa)^2} \left(\frac{S_i}{m} - (X_i - \theta_i)^2 \right) \right)^2 \right] \\ &= \sup_{\kappa > 0} \frac{1}{p^2} \frac{1}{(1+\kappa)^4} \sum_{i=1}^p E \left[\frac{S_i^2}{m} + (X_i - \theta_i)^4 - 2 \frac{S_i}{m} (X_i - \theta_i)^2 \right] \\ &= \sup_{\kappa > 0} \frac{1}{p(1+\kappa)^4} \frac{2m+2}{m} \frac{1}{p} \sum_{i=1}^p \sigma_i^4 \rightarrow 0, \end{aligned}$$

by (C1). Thus, for the first absolute value of the right hand side of the inequality in (42),

$$\sup_{\kappa > 0} \left| \frac{1}{p} \sum_{i=1}^p \frac{1}{(1+\kappa)^2} \left(\frac{S_i}{m} - (X_i - \theta_i)^2 \right) \right| \xrightarrow{p} 0, \quad \text{as } p \rightarrow \infty. \quad (43)$$

In view of the fact that

$$\bar{X} \sim N(\bar{\theta}_p, \bar{\sigma}^2/p), \quad E(\bar{X}^2) = \frac{\bar{\sigma}^2}{p} + \bar{\theta}_p^2, \quad E\bar{S} = \bar{\sigma}^2,$$

and (C1)-(C3), the convergence to zero in probability of the other terms in (42) can be shown in a similar way, that is,

$$\left\{ \begin{array}{l} \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \left(\frac{\kappa}{1+\kappa} \right)^2 \left[\frac{1}{p} \bar{S} - (\bar{X} - \bar{\theta}_p)^2 \right] \right| \\ \leq \sup_{\kappa>0} \frac{1}{p} \frac{\kappa^2}{(1+\kappa)^2} \left| \frac{1}{p} \sum_{i=1}^p \left(\frac{S_i}{m} - \sigma_i^2 \right) \right| + \sup_{\kappa>0} \frac{\kappa^2}{(1+\kappa)^2} |(\bar{X} - \bar{\theta}_p)^2 - \frac{\bar{\sigma}^2}{p}| \xrightarrow{p} 0 \\ \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \frac{\kappa^2}{(1+\kappa)^2} \left(\bar{X}^2 - \frac{\bar{S}}{p} + X_i^2 - \frac{S_i}{m} - 2X_i\bar{X} + \frac{2S_i}{pm} - (\bar{\theta}_p - \theta_i)^2 \right) \right| \\ \leq \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \frac{\kappa^2}{(1+\kappa)^2} \left(\bar{X}^2 - \frac{\bar{S}}{p} - \bar{\theta}_p^2 \right) \right| + \sup_{\kappa \geq 0} \left| \frac{1}{p} \sum_{i=1}^p \frac{\kappa^2}{(1+\kappa)^2} \left(X_i^2 - \frac{S_i}{m} - \theta_i^2 \right) \right| \\ \quad + \sup_{\kappa>0} \left| \frac{2}{p} \sum_{i=1}^p \frac{\kappa^2}{(1+\kappa)^2} (X_i\bar{X} - \frac{1}{p} \frac{S_i}{m} - \theta_i\bar{\theta}_p) \right| \xrightarrow{p} 0 \\ \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} \left(\frac{1}{p} \frac{S_i}{m} - (X_i - \theta_i)(\bar{X} - \bar{\theta}_p) \right) \right| \xrightarrow{p} 0, \\ \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa}{(1+\kappa)^2} (X_i - \theta_i)(\bar{\theta}_p - \theta_i) \right| \xrightarrow{p} 0, \\ \sup_{\kappa>0} \left| \frac{1}{p} \sum_{i=1}^p \frac{2\kappa^2}{(1+\kappa)^2} (\bar{X} - \bar{\theta}_p)(\bar{\theta}_p - \theta_i) \right| \xrightarrow{p} 0. \end{array} \right. \quad (44)$$

From (42), (43) and (44), we have

$$\sup_{\kappa>0} |\text{SURE}_1^G(\gamma, \bar{X}) - l_{1,p}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^G)| \xrightarrow{p} 0, \quad \text{as } p \rightarrow 0.$$

The other part

$$\sup_{\nu>2, A>0} |\text{SURE}_2(\nu, A) - l_{2,p}(\boldsymbol{\sigma}^2, \widehat{\boldsymbol{\sigma}^2})| \xrightarrow{p} 0, \quad \text{as } p \rightarrow 0,$$

has been proved in Theorem 1. Therefore, we complete the proof. $\square$

Proof of Theorem 4. The proof is similar to that of Theorem 2 with Theorem 1 established. $\square$

Proof of Theorem 5. The proof follows the same line as that of Theorem 1. $\square$

Proof of Theorem 6. The proof follows the same line as that of Theorem 2. $\square$

REFERENCES

- Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics*, **36**, 199-227.
- Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *The Annals of Statistics*, **36**, 2577-2604.
- Brown, L. D. (2008). In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Annals of Applied Statistics*, **2**, 113-152.
- Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, **39**, 1551-1579.
- 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**(4), 1685-1704.
- Brown, L. D., Nie, H. and Xie, X. (2012). Ensemble minimax estimation for multivariate normal means. *The Annals of Statistics*. (To appear).
- Brown, L.D. and Han, X. (2012). Optimal estimation of multidimensional normal means with an unknown variance. *Submitted*.
- Cai, T. T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, **38**, 100-145.
- Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, **38**, 2118-2144.
- Chtelat, D. and Wells, M.T. (2012). Improved multivariate normal mean estimation with unknown covariance when p is greater than n . *The Annals of Statistics*, **40**, 3137-3160.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**, 407-499.
- 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.

- Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70**, 311-319.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.
- Hwang, J.T., Qiu, J., and Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society, Series B*, **71**, 265-285.
- James, W. and Stein, C. M. (1961). Estimation with quadratic loss. *Proceedings of 4th Berkeley Symposium on Probability and Statistics*, I, 367-379.
- Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, **37**, 1647-1684.
- Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, **37**, 4254-4278.
- Maruyama, Y. and Strawderman, W. E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *Annals of Statistics*, **4**, 1753-1770.
- Morris, C. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, **78**, 47-55.
- Stein, C. M. (1973). Estimation of the mean of a multivariate distribution. *Proceedings Prague Symposium on Asymptotic Statistics*, 345-381.
- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, **9** (6), 1135-1151.
- Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association*, **102**, 113-122.
- Tong, T., Chen, L. and Zhao, H. (2012). Improved mean estimation and its application to diagonal discriminant analysis. *Bioinformatics*, **28**, 531-537.
- Xie, X. C., Kou, S. C. and Brown, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, **107**, 1465-1479.

Zhao, Z. (2007). Double shrinkage empirical Bayesian estimation for unknown and unequal variances. *Statistics and Its Interface*, 1-8.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.