Journal of Statistical Theory and Applications
Vol. **19(2)**, June 2020, *pp.* 148–161
DOI: https://doi.org/10.2991/jsta.d.200422.001; ISSN 1538-7887
https://www.atlantis-press.com/journals/jsta

Empirical Estimation for Sparse Double-Heteroscedastic Hierarchical Normal Models

Vida Shantia[1], S. K. Ghoreishi[2,*]

[1]*Department of Statistics, Science and Research branch, Islamic Azad University, Tehran, Iran*
[2]*Department of Statistics, Faculty of Sciences, University of Qom, Qom, Iran*

ARTICLE INFO

Article History
Received 05 Mar 2019
Accepted 11 Jun 2019

Keywords
Asymptotic optimality
Heteroscedasticity
Empirical estimators
Sparsity
Stein's unbiased risk estimate (SURE)

2000 Mathematics Subject Classification: 62F15, 62F30

ABSTRACT

The available heteroscedastic hierarchical models perform well for a wide range of real-world data, but for the data sets which exhibit heteroscedasticity mainly due to the lack of constant means rather than unequal variances, the existing models tend to overestimate the variance of the second level model which in turn will cause substantial bias in the parameter estimates. Therefore, in this study, we develop heteroscedastic hierarchical models, called double-heteroscedastic hierarchical models, that take into account the heterogeneity in the means for the second level of the models, in addition to considering the heterogeneity of variance for the first level of the models. In these models, we assume that the vector of means in the second level is sparse. We derive Stein's unbiased risk estimators (SURE) for the parameters in the model based on data decomposition and study their risk properties both in theory and in numerical experiments under the squared loss. The comparison between our SURE estimator and the classical estimators such as empirical Bayes maximum likelihood estimator (EBMLE) and empirical Bayes moment estimator (EBMOM) is illustrated through a simulation study. Finally, we apply our model to a Baseball data set.

1. INTRODUCTION

Hierarchical models have been extensively studied and widely used in many disciplines such as biology, climatology, ecology, medicine and engineering. A hierarchical model is a multi-level model which integrates information from different sources to achieve coherent inferences of unknowns (e.g., [1–3]). Among the many statisticians who have made significant contribution to theories and applications of hierarchical models, James and Stein [4] and Stein [5] were pioneers by studying simultaneous statistical inferences for the mean of several normal populations. The joint work of James and Stein has fundamentally increased the use of hierarchical models in recent decades. Stein's shrinkage estimators that had an interesting empirical Bayes interpretation were the basis for developing shrinkage estimation in multilevel normal models. Later, Efron and Morris [6] showed further implications of Stein's shrinkage estimator and proposed several competing parametric empirical Bayes estimators.

To date, both parametric and nonparametric empirical Bayes properties of shrinkage estimators have been extensively studied under either a homoscedastic (equal subpopulation variances) or heteroscedastic (unequal subpopulation variances) assumption. For more details, see Berger and Strawderman [7] and Brown and Greenshtein [8]. The majority of these investigations have focused on the risk properties of estimators with various loss functions. For example, the admissible minimax estimators for homoscedastic hierarchical normal models with usual quadratic loss functions were considered by Baranchik [9], a class of proper Bayes minimax estimators was studied by Strawderman [10], and a sufficient condition for admissibility of generalized Bayes estimators was investigated by Brown [11]. Comparisons between those estimators under different loss functions are discussed in Brown [12], Berger [13] and Berger and Strawderman [7].

Recently, heteroscedastic hierarchical normal models have received more attention due to the demand for real applications. Xie *et al.* [14] proposed a class of shrinkage estimators based on Stein's unbiased risk estimate (SURE) and studied the asymptotic properties of various common estimators as the number of means to be estimated increases. In particular, they established the asymptotic optimality property for the SURE estimators. They further extended their estimators to a class of semi-parametric shrinkage estimators and established corresponding asymptotic optimality results. Ghoreishi and Meshkani [15] considered the estimation of a set of normal population means by assuming heteroscedasticity in both levels of a two-level hierarchical model. They developed weighted shrinkage estimators of population means based on weighted SURE. This is achieved by first estimating the nuisance parameters of variances and then using them in

*Corresponding author. Email: *atty_ghoreishi@yahoo.com*

the derivation of the shrinkage estimators of means. Still in the context of heteroscedastic models, Xie *et al.* [16] discussed the simultaneous inference of mean parameters in a family of distributions with quadratic variance function. They studied the asymptotic optimality properties of their semi-parametric/parametric shrinkage estimators that were defined for the location-scale family and the natural exponential family. Ghoreishi [17] studied the Bayesian analysis of a fully heteroscedastic hierarchical model. He used a class of local-global shrinkage priors, Dirichlet–Laplace priors and evaluated the optimal posterior concentration of their corresponding Bayes estimators.

Compared to nonparametric Bayes estimators, parametric empirical Bayes estimators are more commonly used in real data analysis [18]. The application of parametric empirical Bayes usually involves unknown hyper-parameters that need to be estimated. The empirical Bayes maximum likelihood estimator (EBMLE), empirical Bayes moment estimator (EBMOM) and SURE are the common approaches for the hyper-parameter estimation, see Brown [19] and references therein.

Although in general the available heteroscedastic hierarchical models perform well for a wide range of real-world data, they may fit poorly for the data sets which exhibit heteroscedasticity mainly due to the lack of constant means rather than unequal variances. For such data, implementation of the existing heteroscedastic hierarchical models tends to overestimate the variance of the second-level model which will cause substantial bias in the parameter estimates. According to the above description, the present study is an attempt to consider the following points

- We develop heteroscedastic hierarchical models that take into account the heterogeneity in the means for the second-level of the model, in addition to considering the heterogeneity of variance for the first level of the models. We call these extended models as "double-heteroscedastic hierarchical models."
- Throughout the paper, we assume that the vector of the second level means is a sparse vector. So, we derive SURE for the parameters in the model based on data decomposition and study their risk properties both in theory and in numerical experiments under the squared loss. The comparison between our SURE estimator and the classical estimators such as EBMLE and EBMOM is illustrated through a simulation study.

Practically, our sparse double-heteroscedastic hierarchical models can be widely used in signal detection and gene effect studies. For example, when many gene families (a group of homologous genes which they share similar sequences and they may have identical functions) are compared for the healthy and treated groups, only a few of these sequences are detected, which are known to cause the disease. Here, the sparse double-heteroscedastic hierarchical models perform well for this comparison.

In the application section, we apply our model to the baseball data analyzed by Brown [19] and compare the batting average of two halves for each player and, finally, detect the players who have had different performances in two halves of the baseball season.

The paper is structured as follows: In section 2, we first give preliminaries including notations and definitions, then introduce a two-level sparse heteroscedastic hierarchical model. Section 3 derives estimators for all parameters in the hierarchical model, and Section 4 studies their risk properties. Extensive stimulation studies for evaluation of our parameter estimates are given in Section 5 and a data example of applying our model is presented in Section 6, followed by a brief discussion in Section 7. Technical proofs are given in the Appendix.

2. PRELIMINARIES

2.1. Notations and Definitions

Given a vector $\mathbf{x} \in \mathcal{R}^n$, define its l_p-norm as $\| \mathbf{x} \|_p^p = \sum_i |x_i|^p$, and define $\| \mathbf{x} \|_0 = \sum_i I_{x_i \neq 0}$. Given $x \in \mathcal{R}$, define $x_+ := \max(0, x)$. For $i \in \{1, 2, \dots, n\}$, let $\alpha_{(i)}$ stand for the i-th largest entry of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$ in absolute value. In other words, one has $|\alpha_{(1)}| \geq |\alpha_{(2)}| \geq \dots \geq |\alpha_{(n)}|$. Also, $[n]$ is short for the set $\{1, 2, \dots, n\}$, and $|S|$ is the cardinality of set S.

2.2. Two-Level Sparse Heteroscedastic Hierarchical Normal Model

Let $Y_1, Y_2, \dots, Y_n$ be independent normal random variables with mean $\theta_1, \theta_2, \dots, \theta_n$ and variance $A_1, A_2, \dots, A_n$, respectively. Here we assumes A_i's are possibly distinct points in $\mathcal{R}$. Then we assume θ_i's are independent and normally distributed with mean α_i's and constant variance λ. These assumptions form a two-level heteroscedastic hierarchical model as follows:

$$\begin{aligned} Y_i|\theta_i &\sim N(\theta_i, A_i) \\ \theta_i &\sim N(\alpha_i, \lambda); \qquad i = 1, 2, \dots, n. \end{aligned} \tag{2.1}$$

We further assume that the vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ is a sparse vector of order k_0, that is, $\| \theta \|_0 \leq k_0$, where $k_0 \ll n$. This sparsity assumption intrinsically guarantees the identifiability of θ_i's. Note that the sparsity of vector $\alpha = (\alpha_1, \dots, \alpha_n)$ also violates the condition that the mean of θ_i's are constant.

From model (2.1) and using Bayes' theorem, we have the posterior distribution of θ_i as the following:

$$N\left(\frac{\lambda}{\lambda+A_i}Y_i+\frac{A_i}{\lambda+A_i}\alpha_i,\ \frac{\lambda A_i}{\lambda+A_i}\right).$$

We opt for the posterior mode to be a class of shrinkage estimators for θ_i. Then, given the sparsity condition, this class of estimators become

$$\hat{\theta}_i^S=\begin{cases}\dfrac{\lambda}{\lambda+A_i}Y_i+\dfrac{A_i}{\lambda+A_i}\alpha_i & \text{if } i\in S\\ \dfrac{\lambda}{\lambda+A_i}Y_i & \text{if } i\in S^c\end{cases}. \tag{2.2}$$

where $S=\{i|\ \alpha_i\neq 0\}$ and $|S|=k_0$.

Since the usual heteroscedastic hierarchical models are derived based on model (2.1) but with conditions $\alpha_1=\alpha_2=\ldots=\alpha_n$, our model (2.1) is more flexible and thus will be applicable to a wider range of real data. Statistical inferences using Model (2.1) include the estimation of the hyper-parameter λ, as well as the estimation of the local hyper-parameters $\alpha_1,\alpha_2,\ldots,\alpha_n$. Indeed, the latter task is the biggest challenge in practice. Our primary goal is to develop a methodology for estimating the hyper-parameters in Model (2.1). In addition, we are interested in the performance of the sparse shrinkage estimators $\hat{\theta}_i^S$'s in (2.2).

3. EMPIRICAL PARAMETER ESTIMATORS OF HYPER-PARAMETERS

We introduce empirical methods for estimating the hyper-parameters in model (2.1). We will start with the empirical methods for estimating the hyper-parameter λ, followed by the estimation of sparsity parameter k_0 and nonzero set S. Then we discuss the empirical estimation of $\alpha_1,\alpha_2,\ldots,\alpha_n$.

3.1. Empirical Estimators of λ

Although empirical approaches such as EBMLE and EBMOM have been frequently used to estimate the model hyper-parameters, we aim to use SURE to estimate the hyper-parameter λ (and other hyper-parameters), because the SURE estimators possess asymptotic optimality properties within the class of heteroscedastic hierarchical normal models [14,15].

From (2.1), the marginal distribution of Y_i is

$$N(\alpha_i,\lambda+A_i). \tag{3.1}$$

Let $\mathcal{J}$ be a collection of subsets of $\{1,2,\ldots,n\}$ of size $n/2$. For simplicity, we assume n to be even. Define

$$X_{J_n}:=\sum_{i\in J_n}\frac{Y_i^2}{\lambda+A_i},\quad \forall J_n\in\mathcal{J}.$$

It is easy to verify that X_{J_n} follows a chi-square distribution with $n/2$ degrees of freedom and noncentrality parameter $\sum_{i\in J_n}\frac{\alpha_i^2}{\lambda+A_i}$. Given λ, assume J_n^* is the subset such that

$$X_{J_n^*}=\inf_{J_n\in\mathcal{J}}X_{J_n}.$$

By Lemma 1 in Carpentier and Verzelen [20], it is straightforward to see that for some constant $c>0$,

$$P\left(\frac{1}{16}<\frac{1}{|J_n^*|}X_{J_n^*}\leq 2.2\right)\geq 1-2e^{-c|J_n^*|}.$$

For more details on the proof and the c value see Boucheron *et al.* [21]. The above inequality follows by taking a union bound over all X_{J_n} for $J_n\in\mathcal{J}$. Therefore, for a relatively large sample size, one approximately has

$$\frac{1}{16}<\frac{1}{\lambda+A_{max}}\frac{1}{|J_n^*|}\sum_{i\in J_n^*}Y_i^2\leq\frac{1}{\lambda+A_{min}}\frac{1}{|J_n^*|}\sum_{i\in J_n^*}Y_i^2\leq 2.2.$$

Consequently, a range of the empirical estimate for λ is given by

$$\left[\frac{1}{2.2|J_n^*|}\sum_{i\in J_n^*}Y_i^2 - A_{min}\right]_+ < \lambda < \left[\frac{16}{|J_n^*|}\sum_{i\in J_n^*}Y_i^2 - A_{max}\right]_+, \tag{3.2}$$

where $A_{min} = \min_{i\in J_n^*} A_i$ and $A_{max} = \max_{i\in J_n^*} A_i$.

If we are interested in estimating λ by EBMOM approach, we adopt the following estimator:

$$\hat{\lambda} = \frac{\left[\sum_{i\in J_n^*}(Y_i^2 - A_i)\right]_+}{|J_n^*|}. \tag{3.3}$$

Or, if we are interested in an EBMLE estimator of λ, it will be obtained by maximizing the marginal density of Y_{J_n}, $j \in J_n^*$ with respect to λ. The EBMLE estimator satisfies the following equation whenever the root exists:

$$\sum_{i\in J_n^*}\left\{\frac{1}{\lambda + A_i} - \frac{Y_i^2}{(\lambda + A_i)^2}\right\} = 0. \tag{3.4}$$

We however focus on estimating λ using the SURE approach. This approach is based on the following squared-loss function:

$$L_1(\hat{\theta}, \theta) = \sum_{i\in J_n^*}(\hat{\theta}_i - \theta_i)^2. \tag{3.5}$$

Under this loss function and using (2.2), the shrinkage estimator

$$\hat{\theta}_i^S = \frac{\lambda}{\lambda + A_i}Y_i$$

can be used for estimating θ_i over J_n^*. Therefore, the corresponding SURE estimator of λ is obtained by minimizing the unbiased estimator of risk function

$$\begin{aligned} R_1(\hat{\theta}^S, \theta) = E\left(L_1(\hat{\theta}^S, \theta)\right) &= E\left(\sum_{i\in J_n^*}(\hat{\theta}_i^S - \theta_i)^2\right) \\ &= E\left(\sum_{i\in J_n^*}\left(\frac{\lambda}{\lambda + A_i}Y_i - \theta_i\right)^2\right) \\ &= \sum_{i\in J_n^*}E\left(\frac{\lambda}{\lambda + A_i}Y_i - \theta_i\right)^2 \\ &= \sum_{i\in J_n^*}\frac{A_i}{\lambda + A_i}(A_i{\theta_i}^2 + \lambda^2). \end{aligned} \tag{3.6}$$

If we define

$$SURE_1(\lambda) = \sum_{i\in J_n^*}\left[\left(\frac{A_i}{\lambda + A_i}\right)^2 Y_i^2 + \frac{A_i(\lambda - A_i)}{A_i + \lambda}\right], \tag{3.7}$$

then the SURE estimator of λ is obtained as the minimizer of $SURE_1(\lambda)$ which is also the solution of the following equation whenever the root exists

$$\sum_{i\in J_n^*}\left[\frac{A_i^2}{(\lambda + A_i)^3}Y_i^2 - \frac{A_i^2}{(A_i + \lambda)^2}\right] = 0. \tag{3.8}$$

For more details see Xie *et al.* [14] and Ghoreishi and Meshkani [15].

3.2. Empirical Sparsity Estimation

The estimation of sparsity $k_0 = \| \alpha \|_0$ relies on the empirical characteristic function of $\{Y_i;\ i \in J_n^{*c} = [n] - J_n^*\}$. Define the function $h: \mathcal{R} \to [0,1]$ as $h(x) = 1 - 2\frac{1-\cos(x)}{(x)^2}$. Conventionally, we define $h(0) = 0$ for $x = 0$. It is easy to see that $h(x) \geq 1 - 4/x^2$ for large $|x|$ and $h(x) = x^2/50 + o(x^2)$ around 0. Therefore, given $t > 0$, we have

$$k_0 = \| \alpha \|_0 \approx \sum_{i \in J^{*c}} h\left(\frac{t\alpha_i}{\sqrt{\lambda + A_i}} \right). \tag{3.9}$$

We then employ the above approximation to estimate $k_0 = \| \alpha \|_0$.

To derive the estimator, we first consider the empirical characteristic function of Y_i's that is defined as

$$\kappa_t(x) := \int_{-1}^{1} (1 - |u|) e^{t^2 u^2/2} \cos(txu) du.$$

Then we define a statistic $Z_{\hat{\lambda}}(t)$ by

$$Z_{\hat{\lambda}}(t) := \sum_{i \in J_n^{*c}} \left(1 - \kappa_t \left(\frac{Y_i}{\sqrt{\hat{\lambda} + A_i}} \right) \right). \tag{3.10}$$

It is straightforward to see that

$$E\left(Z_{\hat{\lambda}}(t) \mid \hat{\lambda} \right) = \sum_{i \in J_n^{*c}} \left[1 - 2 \frac{1 - \cos\left(\frac{t\alpha_i}{\sqrt{\hat{\lambda} + A_i}} \right)}{\left(\frac{t\alpha_i}{\sqrt{\hat{\lambda} + A_i}} \right)^2} \right] = \sum_{i \in J^{*c}} h\left(\frac{t\alpha_i}{\sqrt{\hat{\lambda} + A_i}} \right),$$

which implies that $Z_{\hat{\lambda}}(t)$ is an empirical estimator of $k_0 = \| \alpha \|_0$. Hence, our empirical estimator for k_0 is

$$\hat{k}_0 = \left[\lfloor Z_{\hat{\lambda}}(t) \rfloor + 1 \right]_+,$$

where $\lfloor u \rfloor$ stands for the integer part of u and $u_+ = \max(0, u)$. This estimator is a function of t which is usually determined using cross-validation with the objective function as the mean squared prediction error (MSPE) of θ estimates. Section 5 gives more details on how to numerically determine t.

In addition, we may be interested in constructing an $1 - \delta$ empirical confidence interval for k_0. For this purpose, we define

$$w(x_1, x_2, \ldots, x_k) = \sum_{i \in k} (1 - \kappa_t(x_i)).$$

It is easy to verify that for a given t and $i = 1, 2, \ldots, k$,

$$\sup_{\mathbf{x}_{-i}, x, y} |w(x_1, x_2, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_k) - w(x_1, x_2, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_k)| \leq 2e^{t^2/2},$$

where $\mathbf{x}_{-i} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$. Therefore, using McDiarmid's inequality, Boucheron *et al.* [21], we have an empirical $1 - \delta$ confidence interval for k_0 as

$$P\left(\left| Z_{\hat{\lambda}}(t) - E(Z_{\hat{\lambda}}(t)) \right| \leq e^{t^2/2} \sqrt{\ln \delta^{-1} . J^{*c}} \right) \geq 1 - \delta,$$

or equivalently, the confidence interval can be expressed into

$$k_0 : \sum_{i \in J_n^{*c}} \left(1 - \kappa_t \left(\frac{Y_i}{\sqrt{\hat{\lambda} + A_i}} \right) \right) \pm e^{t^2/2} \sqrt{\ln \delta^{-1} . J^{*c}}.$$

Note that since $|J^*| = n/2$ and $k_0 \ll n$, the above confidence interval is practically more conservative. This is not surprising as confidence intervals derived based on Chebyshev's inequalities and its derivatives such as McDiarmid's or Hoeffding's inequalities are in general more conservative compared to the confidence intervals constructed using the exact distributions. An alternative approach to computing confidence interval is

$$\hat{k}_0 \pm S.E.(\hat{k}_0), \tag{3.11}$$

where

$$S.E.(\hat{k}_0) = \sqrt{\sum_{i \in J_n^{*c}} \left[1 - \kappa_t \left(\frac{Y_i}{\sqrt{\hat{\lambda} + A_i}} \right) - h \left(\frac{tY_i}{\sqrt{\hat{\lambda} + A_i}} \right) \right]^2}.$$

Given $\hat{k}_0$, the subset S can be estimated by

$$\hat{S} = \{R_1, R_2, \dots, R_{\hat{k}_0}\}, \tag{3.12}$$

where $\{R_1, R_2, \dots, R_{\hat{k}_0}\} \subset \{1, 2, \dots, n\}$ are the samples corresponding to the largest $|Y_i|$. However, this estimator can be biased without implementing an iterative estimation procedure. This is because under our model (2.1), the quantity $\sum_{i \in S^c} \frac{Y_i^2}{\hat{\lambda} + A_i}$ in theory has a central chi-square distribution over S^c. Whereas, this quantity often follows a noncentral chi-square distribution in practice due to the randomness of data. It is well-known that a central chi-square distribution is stochastically smaller than any noncentral chi-square with the same degrees of freedom. Therefore, in order to estimate S^c (thus S) accurately, we need to increase the size of $\hat{S}^c$ via a few iterative steps so that $\sum_{i \in \hat{S}^c} \frac{Y_i^2}{\hat{\lambda} + A_i}$ will approximately have a central chi-square distribution. Indeed, Proposition 3.1 below provides intuitive ideas of increasing the size of $\hat{S}^c$ till the central chi-square distribution is approximately reached for $\sum_{i \in \hat{S}^c} \frac{Y_i^2}{\hat{\lambda} + A_i}$.

We propose the following iterative procedure for practical purpose:

1. Randomly choose half of the sample as our initial estimate of S^c, i.e., $|\hat{S}^c| = n/2$.
2. Randomly choose 5% of the observations in $\hat{S}$ and add them to $\hat{S}^c$. If $\inf_{\hat{S}^c} X_{\hat{S}^c} = \inf_{\hat{S}^c} \frac{1}{|\hat{S}^c|} \sum_{i \in \hat{S}^c} \frac{Y_i^2}{\hat{\lambda} + A_i} \leq 1.1$, we accept this new $\hat{S}^c$. Otherwise, we reject the new $\hat{S}^c$ and keep the previous $\hat{S}^c$.
3. Repeat Step 1 until $\hat{S}^c$ contains at least 90% of the data. This guarantees $k_0/n \leq 0.01$ which empirically satisfies our assumption of $k_0 \ll n$.

The justification of the proposed iterative estimation procedure is stated in the following proposition:

Proposition 3.1. *Given $\hat{\lambda}$, $\hat{S}^c$, and under marginal model (3.1) with sparsity assumption of $\| \alpha \|_0 = k_0$, one has*

$$P\left(\frac{1}{|\hat{S}^c|} \sum_{i \in \hat{S}^c} \frac{Y_i^2}{\hat{\lambda} + A_i} > 1.1 \mid \hat{\lambda} \right) < e^{-|\hat{S}^c|/2}.$$

The proof for this Proposition is deferred to the Appendix.

3.3. Empirical Estimation of α_i's

To estimate the hyper-parameters α_i's, we apply the squared-loss function

$$L_2(\hat{\theta}, \theta) = \sum_{i \in S^*} (\hat{\theta}_i - \theta_i)^2, \tag{3.13}$$

to samples in S^*. Under this loss function, one can compute the risk function of the shrinkage estimators in (2.2):

$$\begin{aligned} R_2(\hat{\theta}^S, \theta) = E\left(L_2(\hat{\theta}^S, \theta)\right) &= E\left(\sum_{i\in\hat{S}^*}(\hat{\theta}_i^S - \theta_i)^2\right) \\ &= E\left(\sum_{i\in\hat{S}^*}\left(\frac{\lambda}{\lambda + A_i}Y_i - \theta_i\right)^2\right) \\ &= \sum_{i\in\hat{S}^*} E\left(\frac{\lambda}{\lambda + A_i}Y_i + \frac{A_i}{\lambda + A_i}\alpha_i - \theta_i\right)^2 \\ &= \sum_{i\in\hat{S}^*}\frac{\lambda^2 A_i}{(\lambda + A_i)^2} + \sum_{i\in\hat{S}^*}\frac{A_i^2}{(\lambda + A_i)^2}(\alpha_i - \theta_i)^2. \end{aligned} \tag{3.14}$$

An unbiased estimator for the risk $R_2(\hat{\theta}_i^S, \theta)$ is thus given by

$$SURE_2(\alpha_1, \alpha_2, \dots, \alpha_{k_0}) = \sum_{i\in\hat{S}^*}\frac{\lambda^2 A_i}{(\lambda + A_i)^2} - \sum_{i\in\hat{S}^*}\frac{A_i^3}{(\lambda + A_i)^2} + \sum_{i\in\hat{S}^*}\frac{A_i^2}{(\lambda + A_i)^2}(Y_i - \alpha_i)^2.$$

Given empirical estimates $\hat{\lambda}$, $\hat{k}_0$, and $\hat{S}$, it is straightforward to see that the SURE estimates of α_i's are

$$\hat{\alpha}_i = Y_i, \tag{3.15}$$

for $i \in \hat{S}$, which are obtained by minimizing $SURE_2(\alpha_1, \alpha_2, \dots, \alpha_{\hat{k}_0})$ with respect to $\alpha_1, \alpha_2, \dots$, and $\alpha_{\hat{k}_0}$. Combining (3.8) and (3.15), the SURE estimates of θ_i's are

$$\hat{\theta}_i^{SURE} = \begin{cases} Y_i & \text{if } i \in \hat{S} \\ \dfrac{\hat{\lambda}}{\hat{\lambda} + A_i}Y_i & \text{if } i \in \hat{S}^c \end{cases}, \tag{3.16}$$

4. RISK PROPERTIES OF SURE ESTIMATOR

We establish properties of the SURE estimator in (3.16) under the following squared-loss function that is applied to the entire data:

$$L(\hat{\theta}, \theta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2. \tag{4.1}$$

For ease of notation, in the following we assume $\hat{\theta} = \hat{\theta}^{SURE}$. Hence, we have that

$$L(\hat{\theta}, \theta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2 = \frac{1}{n}L_1(\hat{\theta}, \theta) + \frac{1}{n}L_2(\hat{\theta}, \theta).$$

This decomposition allows us to evaluate the performance of the SURE estimators using estimators derived based on partial data, that is,

$$SURE(\lambda, \alpha_1, \alpha_2, \dots, \alpha_{k_0}) = \frac{1}{n}SURE_1(\lambda) + \frac{1}{n}SURE_2(\alpha_1, \alpha_2, \dots, \alpha_{k_0}).$$

The following theorem shows how well $SURE(\lambda, \alpha_1, \alpha_2, \dots, \alpha_{k_0})$ approximates $L(\hat{\theta}, \theta)$.

Theorem 4.1. *Under Conditions C1 and C2 in the Appendix, we have*

$$\sup_{\lambda>0, \alpha_1, \alpha_2, \dots, \alpha_{k_0}} \left|SURE(\lambda, \alpha_1, \alpha_2, \dots, \alpha_{k_0}) - L(\hat{\theta}, \theta)\right| \to 0 \quad \text{in } L^2, \text{ as } n \to \infty.$$

The proof of this theorem is deferred to the Appendix.

5. NUMERICAL STUDIES

We carry out simulation studies to evaluate the performance of the SURE estimators for sparse heteroscedastic hierarchical normal models. To measure the performance of the estimators, we compute the risk $R(\hat{\theta}, \theta)$ of the proposed shrinkage estimators. We have $n = 5000$ repetitions in our simulation for each model defined in the four scenarios below. All four scenarios have the two-level model

$$\begin{aligned} Y_i &\sim N(\theta_i, A_i) \\ \theta_i &\sim N(\alpha_i, 0.4), \end{aligned}$$

with sparsity size $k = 10$ but with different settings for A_i and θ_i.

- *Scenario* I: Baseline model

$$\begin{aligned} A_i &\sim U(0,1), \qquad i = 1, 2, \dots, 5000, \\ \alpha_i &= \begin{cases} 5 & \text{for } i = 1, \dots, k_0 = 10 \\ 0 & \text{for } i = k+1, \dots, 5000 \end{cases} \end{aligned}$$

- *Scenario* II: Assume a two-component mixture model for α_i

$$\begin{aligned} A_i &\sim U(0,1), \quad i = 1, 2, \dots, 5000, \\ \alpha_i &= \begin{cases} 0.5N(3, A_i) + 0.5N(5, A_i) & \text{for } i = 1, \dots, k_0 = 10 \\ 0 & \text{for } i = k+1, \dots, 5000 \end{cases} \end{aligned}$$

  Scenario II introduces a more complex mean structure than Scenario I.

- *Scenario* III: Assume a three-component mixture model for α_i

$$\begin{aligned} A_i &\sim U(0,1), \quad i = 1, 2, \dots, 5000, \\ \alpha_i &= \begin{cases} \frac{1}{3}N(2, A_i) + \frac{1}{3}N(5, A_i) + \frac{1}{3}N(10, A_i) & \text{for } i = 1, \dots, k_0 = 10 \\ 0 & \text{for } i = k+1, \dots, 5000 \end{cases} \end{aligned}$$

- *Scenario* IV: Assume three-component mixture models for both A_i and α_i

$$\begin{aligned} A_i &\sim 0.3U(0,1) + 0.7U(1,2), \quad i = 1, 2, \dots, 5000, \\ \alpha_i &= \begin{cases} \frac{1}{3}N(2, A_i) + \frac{1}{3}N(5, A_i) + \frac{1}{3}N(10, A_i) & \text{for } i = 1, \dots, k_0 = 10 \\ 0 & \text{for } i = k+1, \dots, 5000 \end{cases} \end{aligned}$$

  Scenario IV has a more complex variance structure than Scenarios I-III.

For each simulated data, we estimate the hyper-parameters λ, k_0 and α_i, using Equations (3.8), (3.10), and (3.15), respectively. The 95% confidence interval of k_0 is also computed based on (3.11). Given $\hat{k}_0$, the nonzero set S will be estimated by (3.12) and we further compute the misclassification rate (MCR) of S defined as the ratio of misclassified α_i's to $n = 5000$. We run simulation 200 times to obtain 200 sets of hyper-parameters estimates and MCR, then compute the empirical coverage probability (ECP) defined as the rate of true k_0 being covered by its confidence interval over all simulation runs and the risk of parameter estimates, $R(\hat{\theta}, \theta)$, where $R(\hat{\theta}, \theta)$ is defined in (4.1). As is seen the parameter estimates depend on t values. A general rule for determining t is through the cross-validation method, that is, one opts for t that minimizes the MSPE defined as

$$MSPE(t) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \hat{\theta}_j(t)\right)^2.$$

Using this criterion, we find $t = 1$ roughly gives the smallest MSPE with our simulation data, so we fix $t = 1$ to compare the performance of different methods.

To evaluate the performance of our estimators, we consider the oracle estimator of λ, i.e., $\tilde{\lambda} = arg\min_{\lambda \geq 0} \sum_{i \in J_n^*} \left(\frac{\lambda}{\lambda + A_i} Y_i - \theta_i\right)^2$ and its corresponding shrinkage estimator $\tilde{\theta}_i = \frac{\tilde{\lambda}}{\tilde{\lambda} + A_i} Y_i$, for $i \in J_n^*$. Here, the notation $\tilde{\theta}_i$ rather than $\hat{\theta}_i$ is used for emphasizing that $\tilde{\theta}_i$ depends on unknown θ_i and hence is not really an estimator. However, since $\tilde{\theta}_i$ has smaller loss or risk within the class of estimators in the form $\hat{\theta}_i = \frac{\hat{\lambda}}{\hat{\lambda} + A_i} Y_i$, we simply treat it as a reference for evaluating the performance of our SURE estimators.

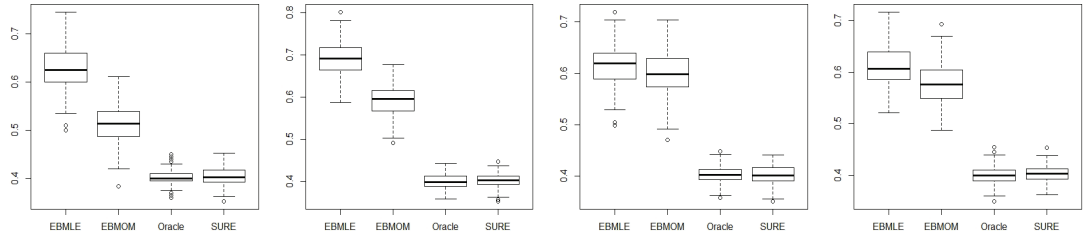
In addition to the oracle estimator, we also compute the EBMOM and EBML estimates and their MCR, ECP and risk. The summary of simulation results over 200 runs is reported in Table 1. In particular, we show the variability of $\hat{\lambda}$, $\hat{k}_0$, and MCR over 200 simulation runs through boxplots in Figures 1, 2 and 3. The results show that under the sparse heteroscedastic hierarchical normal model with nonconstant means we consider in this article, our proposed SURE estimator clearly outperforms EBMOM and EBMLE and performs similarly as the oracle estimator in terms of both parameter estimation and identification of nonzero locations.

To examine whether the true sparsity of the data affects the sparsity parameter estimation, we rerun the simulation by setting k_0 in the data generation process to 15, 20, 30 and 50, respectively. Table 2 summarizes the k_0 estimates together with the ECP and MCR at different k

Table 1 | Simulation results under various scenarios. The columns of $\hat{\lambda}$, $\hat{k}_0$, MCR, and ECP are mean estimates and their standard deviations inside the parentheses over 200 simulation runs.

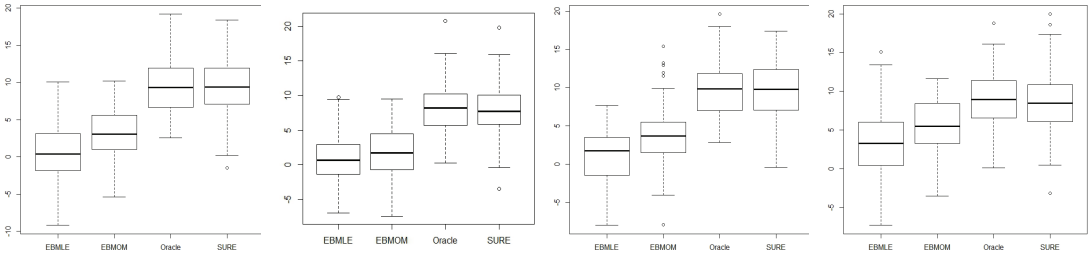
Scenarios	Method	$\hat{\lambda}$	$\hat{k}_0$	MCR	ECP	$R(\hat{\theta}, \theta)$
	Oracle	0.401 (0.023)	9.421 (3.09)	0.137 (0.083)	0.70 (0.173)	0.217
Scenario I	SURE	0.404 (0.043)	9.381 (3.24)	0.191 (0.092)	0.63 (0.173)	0.249
	EBMOM	0.513 (0.058)	3.299 (3.41)	0.671 (0.173)	0.21 (0.121)	0.734
	EBMLE	0.627 (0.067)	0.632 (4.33)	0.819 (0.192)	0.08 (0.114)	1.117
	Oracle	0.401 (0.022)	7.948 (3.19)	0.212 (0.108)	0.66 (0.165)	0.422
Scenario II	SURE	0.401 (0.023)	7.700 (3.32)	0.241 (0.123)	0.61 (0.154)	0.437
	EBMOM	0.592 (0.051)	1.890 (3.89)	0.757 (0.208)	0.14 (0.102)	0.801
	EBMLE	0.693 (0.055)	0.860 (3.61)	0.901 (0.203)	0.02 (0.093)	1.115
	Oracle	0.402 (0.027)	9.808 (3.27)	0.163 (0.099)	0.66 (0.163)	0.267
Scenario III	SURE	0.403 (0.027)	9.592 (3.35)	0.182 (0.104)	0.64 (0.160)	0.298
	EBMOM	0.601 (0.057)	3.978 (3.48)	0.610 (0.123)	0.22 (0.123)	0.721
	EBMLE	0.617 (0.056)	1.154 (3.62)	0.857 (0.173)	0.07 (0.099)	1.112
	Oracle	0.401 (0.023)	8.883 (3.11)	0.220(0.008)	0.65(0.176)	0.333
Scenario IV	SURE	0.403 (0.026)	8.576 (3.23)	0.236(0.008)	0.61(0.142)	0.421
	EBMOM	0.577 (0.054)	5.396 (3.88)	0.498(0.049)	0.42(0.133)	0.877
	EBMLE	0.611 (0.051)	3.479 (4.64)	0.619(0.058)	0.24(0.114)	1.234

EBMLE, empirical Bayes maximum likelihood estimator; EBMOM, empirical Bayes maximum likelihood estimator; SURE, Stein's unbiased risk estimate; MCR, misclassification rate; ECP, empirical coverage probability.



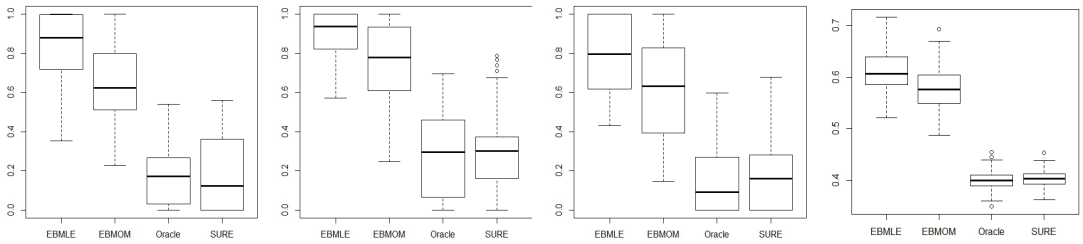
(a) Scenario I (b) Scenario II (c) Scenario III (d) Scenario IV

Figure 1 | Boxplots for $\hat{\lambda}$ with EBMLE, EBMOM, Oracle and SURE methods for four scenarios. The true value of λ is 4.



(a) Scenario I (b) Scenario II (c) Scenario III (d) Scenario IV

Figure 2 | Boxplots for $\hat{k}_0$ with EBMLE, EBMOM, Oracle and SURE methods for four scenarios. The true value of k_0 is 10.



(a) Scenario I (b) Scenario II (c) Scenario III (d) Scenario IV

Figure 3 | Boxplots for MCR with EBMLE, EBMOM, Oracle and SURE methods for four scenarios.

Table 2 | Estimation of $\hat{k}_0$ ECP, and MCR with SURE method under various k and scenarios.

		$\hat{k}_0$				
		10	**15**	**20**	**30**	**50**
	$\hat{k}_0$	9.381	13.612	18.704	28.375	46.441
Scenario-I	ECP	0.63	0.60	0.71	0.68	0.44
	MCR	0.187	0.146	0.104	0.077	0.078
	$\hat{k}$	7.700	11.890	15.658	24.276	40.442
Scenario-II	ECP	0.61	0.50	0.44	0.31	0.19
	MCR	0.282	0.243	0.228	0.194	0.191
	$\hat{k}$	9.592	14.705	18.892	19.284	19.709
Scenario-III	ECP	0.64	0.62	0.69	0.14	0.11
	MCR	0.177	0.113	0.357	0.676	0.605
	$\hat{k}$	8.576	13.632	17.482	17.574	17.681
Scenario-IV	ECP	0.61	0.67	0.56	0.09	0.05
	MCR	0.238	0.142	0.159	0.414	0.646

SURE, Stein's unbiased risk estimate; MCR, misclassification rate; ECP, empirical coverage probability.

under the four scenarios. The results show that k_0 estimates in Scenario I are consistently close to the true values across all k_0, but when the mean structure and the variance structure become more complex, as in Scenario II to IV but especially in Scenario III and IV, $\hat{k}_0$ tends to underestimate the true k_0 as k_0 increases. As a consequence, the ECP of confidence intervals for k_0 deteriorates and the MCR of S estimates is elevated, whenever k_0 is underestimated.

6. APPLICATION TO REAL DATA

We analyze the baseball data that was first introduced by Brown [19] and then later analyzed by Xie *et al.* [14]. This dataset contains batting records for all Major League Baseball players in the season of 2005. We first divide the dataset into two half seasons, then our goal is to compare the batting average of two halves for each player and detect the players who have had different performance in two halves of the baseball season. Following Brown [19] and Xie *et al.* [14], we removed the players whose number of at-bats is less than 10 to improve the accuracy of estimation. After this screening procedure, only 503 players out of 929 were selected.

Following the notations in Brown [19] and Xie *et al.* [14], we use N to denote the number of at-bats and H to denote the number of successful battings. Then we have

$$H_{ij} \sim B(N_{ij}, p_{ij}),$$

where $i = 1, 2$ stands for the season and $j = 1, 2, \ldots, n$ stands for the player. Again following Brown [19], we take a variance-stabilizing transformation of the data to obtain X:

$$X_{ij} = \arcsin \sqrt{\frac{H_{ij} + 0.25}{N_{ij} + 0.4}},$$

which approximately follows the distribution of

$$X_{ij} \sim N\left(\theta_{ij}, \frac{1}{4N_{ij}}\right),$$

where $\theta_{ij} = \arcsin\sqrt{p_{ij}}$. Finally, we apply our methodology to variables

$$Y_j = X_{1j} - X_{2j} \sim N\left(\theta_j,\ \frac{1}{4}\left(\frac{1}{N_{1j}} + \frac{1}{N_{2j}}\right)\right); \quad j = 1, 2, \ldots, 503,$$

where $\theta_j = \theta_{1j} - \theta_{2j}$.

Our analysis assumes independency between the performance of each player in two half seasons. This assumption may seem unreasonable at first glance. However, if we look at the performance of each player in two half-seasons, the Yule association measure, defined as

$$\frac{\frac{H_{1j}}{N_{1j}} - \frac{H_{2j}}{N_{2j}}}{\frac{H_{1j}}{N_{1j}} + \frac{H_{2j}}{N_{2j}}},$$

will be a small number around zero, because for each player $\frac{H_{1j}}{N_{1j}} \simeq \frac{H_{2j}}{N_{2j}}$. Therefore, the independency assumption for each individual is not unrealistic.

The cross-validation yields $t = 0.973$, and our SURE estimates for λ is $1.6e^{-4}$, and for k_0 is 1 with a 95% confidence interval [0, 10]. We conclude that at most 10 players perform differently in two half seasons. We also compute the SURE estimate of θ_i, and select 10 players that correspond to the largest $\hat{\theta}_i^{SURE}$ in absolute value. The name of the 10 players (i.e., the estimate of S), their total number of at-bats and total number of successful battings for two halves seasons, a two-sample Z statistic defined by

$$Z = \frac{X_1 - X_2}{\frac{1}{4}\left(\frac{1}{N_2} + \frac{1}{N_1}\right)},$$

and the SURE estimate of θ_i are presented in Table 3. These 10 players include 9 nonpitchers and 1 pitcher. According to $\hat{\theta}^{SURE}$, the most different performance is related to the nonpitcher players. For example, Ordonez and Balanco are the two nonpitch players that perform most differently over the two half seasons although their differences are in opposite directions.

7. DISCUSSION

We develop heteroscedastic hierarchical normal models with sparse and unequal mean as well as derive SURE for such models at demand of the data sets which exhibit heteroscedasticity mainly due to the lack of constant means rather than unequal variances. Our model fills in the gap that no existing literatures are devoted to the heteroscedastic hierarchical normal models with sparse and unequal mean structure. Both theory and simulation study show that our SURE holds nice properties and outperforms the classic EBML and EBMOM estimator for our proposed model. However, we also notice that the SURE estimates are sensitive to the mean and variance structures as the sparsity decreases. We leave the investigation on improving parameter estimates for future research.

Table 3 Information of the 10 players and their θ estimates.

Full Name	Pitchers (1)/Nonpitchers (0)	N_1	N_2	H_1	H_2	Z-Stat.	$\hat{\theta}^{SURE}$
Magglio Ordonez	0	10	208	0	92	2.716	0.4379
Brandon Webb	1	27	35	0	6	2.622	0.3375
Victor Martinez	0	252	442	61	106	2.864	0.1281
Todd Helton	0	269	414	71	92	2.769	0.1278
Rafael Furcal	0	304	507	71	104	2.625	0.1109
Aaron Hill	0	131	273	47	52	2.609	0.0258
Moises Alou	0	225	341	73	37	−3.257	−0.1632
Desi Relaford	0	175	210	45	2	−2.976	−0.2781
Tony Blanco	0	44	56	11	0	−2.921	−0.4202

ACKNOWLEDGMENTS

We wish to thank the editor and two anonymous referees whose comments greatly improved the article.

REFERENCES

1. B. Li, D.W. Nychka, C.M. Ammann, J. Am. Stat. Assoc. 105 (2010), 883–911.
2. L. Barboza, B. Li, M. Tingely, F. Viens, Ann. Appl. Stat. 8 (2014), 1966–2001.
3. L. Shand, B. Li, T. Park, D. Albraccin, J. R. Stat. Soc. Ser. C. 67 (2018), 1003–1022.
4. W. James, C.M. Stein, in Proceedings of the 4th Berkeley Symposium on Probability and Statistics, University of California Press, Berkeley, CA, USA, 1961, vol. I, pp. 367–379.
5. C.M. Stein, J. R. Stat. Soc. Ser. B. 24 (1962), 265–296.
6. B. Efron, C. Morris, J. Am. Stat. Assoc. 68 (1973), 117–130.
7. J. Berger, W.E. Strawderman, Ann. Stat. 24 (1996), 931–951.
8. L.D. Brown, E. Greenshtein, Ann. Stat. 37 (2009), 1685–1704.
9. A.J. Baranchik, Ann. Math. Stat. 41 (1970), 642–645.
10. W.E. Strawderman, Ann. Math. Stat. 42 (1971), 385–388.
11. L.D. Brown, Ann. Math. Stat. 42 (1971), 855–903.
12. L.D. Brown, J. Am. Stat. Assoc. 70 (1975), 417–427.
13. J. Berger, Ann. Stat. 4 (1976), 223–226.
14. X. Xie, S.C. Kou, L.D. Brown, J. Am. Stat. Assoc. 107 (2012), 1465–1479.
15. S.K. Ghoreishi, M.R. Meshkani, J. Multivar. Anal. 132 (2014), 129–137.
16. X. Xie, S.C. Kou, L.D. Brown, Ann. Stat. 44 (2016), 564–597.
17. S.K. Ghoreishi, J. Stat. Theory Appl. 16 (2017), 53–64.
18. C. Morris, J. Am. Stat. Assoc. 78 (1983), 47–55.
19. L.D. Brown, Ann. Appl. Stat. 2 (2008), 113–152.
20. A. Carpentier, N. Verzelen, Ann. Statist. 47 (2019), 93–126.
21. S. Bouchern, G. Lugosi, P. Massart, Concentration Inequalities: a Nonasymptotic Theory of Independence, Clarendon Press, Oxford, UK, 2013.
22. T.T. Cai, Z. Guo, Accuracy assessment for high-dimensional linear regression, arXiv preprint arXiv:1603.03474, 2016.

APPENDIX

To establish the asymptotic results, we assume two mild conditions following Ghoreishi and Meshkani [15] and Xie *et al.* [14],

C1) $\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} A_i < \infty$.

C2) $\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \theta_i^2 < \infty$.

Moreover, without loss of generality, we assume that the sub-populations were re-indexed such that we have $0 < A_1 \le A_2 \le \cdots \le A_n$.

Proof of Proposition 3.1 Define $u_i = \dfrac{Y_i^2}{\hat{\lambda} + A_i}$, and $X = \sum_{i=1}^{|\hat{S}^c|} u_{(i)}$. Let $\hat{S}^c$ is a set of size $|\hat{S}^c|$ that does not intersect with the support of α. Then

$$X = \sum_{i=1}^{|\hat{S}^c|} u_{(i)} \sim \chi^2_{|\hat{S}^c|}$$

That is, X follows a χ^2 distribution with $|\hat{S}^c|$ degrees of freedom. By Cai and Guo [22], we know that

$$P\left(X > 1.1|\hat{S}^c|\right) \le e^{-C|\hat{S}^c|},$$

and this completes the proof. Recall that $|\hat{S}^c| \approx n$.

Proof of Theorem 4.1 Consider the squared-loss function

$$L(\hat{\theta}, \theta) = \frac{1}{n}\sum_{i\in n}(\hat{\theta}_i - \theta_i)^2 = \frac{1}{n}L_1(\hat{\theta}, \theta) + \frac{1}{n}L_2(\hat{\theta}, \theta).$$

It is easy to see that the corresponding risk function of this loss function is given by

$$R(\hat{\theta}, \theta) = \frac{1}{n}\sum_{i\in J_n^*} \frac{A_i}{\lambda + A_i}(A_i\theta_i^2 + \lambda^2) + \frac{1}{n}\sum_{i\in\hat{S}^*} \frac{\lambda^2 A_i}{(\lambda + A_i)^2}$$
$$+\frac{1}{n}\sum_{i\in\hat{S}^*} \frac{A_i^2}{(\lambda + A_i)^2}(\alpha_i - \theta_i)^2.$$

The SURE unbiased estimator of $R(\hat{\theta}, \theta)$ is

$$SURE(\lambda, \alpha_1, \alpha_2, \dots, \alpha_{k_0}) = \frac{1}{n}SURE_1(\lambda) + \frac{1}{n}SURE_2(\alpha_1, \alpha_2, \dots, \alpha_{k_0}).$$

For a given $k_0 \ll n$, one has the following decomposition

$$\begin{aligned} &SURE(\lambda, \alpha_1, \alpha_2, \dots, \alpha_{k_0}) - L(\hat{\theta}, \theta) \\ &= \frac{1}{n}\left(SURE_1(\lambda) - L_1(\hat{\theta}, \theta)\right) + \frac{1}{n}\left(SURE_2(\alpha_1, \alpha_2, \dots, \alpha_{k_0}) - L_2(\hat{\theta}, \theta)\right). \end{aligned}$$

Hence, by taking the absolute value on both sides of the above equation, one has

$$\sup_{\lambda>0,\alpha_1,\alpha_2,\dots,\alpha_{k_0}} \left|SURE(\lambda, \alpha_1, \alpha_2, \dots, \alpha_{k_0}) - L(\hat{\theta}, \theta)\right|$$
$$\le \frac{1}{n}\sup_{\lambda>0}\left|SURE_1(\lambda) - L_1(\hat{\theta}, \theta)\right| + \frac{1}{n}\sup_{\alpha_1,\alpha_2,\dots,\alpha_{k_0}} \left|SURE_2(\alpha_1, \alpha_2, \dots, \alpha_{k_0}) - L_2(\hat{\theta}, \theta)\right|.$$

We consider the two terms of the right hand side separately. For the first term, we have

$$SURE_1(\lambda) - L_1(\hat{\theta}, \theta) = \sum_{i\in J_n^*}\left[Y_i^2 - A_i - \theta_i^2 - 2\frac{\lambda}{\lambda + A_i}(Y_i^2 - \theta_i Y_i - A_i)\right].$$

It is known that for a decreasing sequence of positive numbers $c_j = \frac{\lambda}{\lambda + A_i}$

$$\begin{aligned}
&\sup_{\lambda>0} \left| SURE_1(\lambda) - L_1(\hat{\theta}, \theta) \right| \\
&\leq \sum_{i \in S^{*c}} \left| Y_i^2 - A_i - {\theta_i}^2 \right| + \sup_{\lambda>0} \left| \sum_{i \in S^{*c}} \frac{2\lambda}{\lambda + A_i} (Y_i^2 - \theta_i Y_i - A_i) \right| \\
&\leq \sum_{i \in S^{*c}} \left| Y_i^2 - A_i - {\theta_i}^2 \right| + d \sup_{c_1 \geq c_2 \geq, \ldots \geq c_n} \left| \sum_{i \in S^{*c}} 2c_i (Y_i^2 - \theta_i Y_i - A_i) \right|.
\end{aligned}$$

Applying Theorem 3.1 in Ghoreishi and Meshkani [15] together with Conditions C1) and C2), we have

$$\sup_{\lambda>0} \frac{1}{n} \mid SURE_1(\lambda) - L_1(\hat{\theta}, \theta) \mid \to 0 \quad \text{in } L^2, \text{ as } |S^{*c}| \text{ or } n \to \infty. \tag{A.1}$$

For the second term, it is obvious that

$$\begin{aligned}
&SURE_2(\alpha_1, \alpha_2, \ldots, \alpha_{k_0}) - L_2(\hat{\theta}, \theta) \\
&= \sum_{i \in S^*} \left[(Y_i - \alpha_i)^2 - A_i - (\theta_i - \alpha_i)^2 - \frac{2\lambda}{\lambda + A_i} \left((Y_i - \alpha_i)^2 - (\theta_i - \alpha_i)(Y_i - \alpha_i) - A_i \right) \right].
\end{aligned}$$

Now, by definition $X_i = Y_i - \alpha_i$ and $\beta_i = \theta_i - \alpha_i$, we see again that

$$\begin{aligned}
&\sup_{\alpha_1, \alpha_2, \ldots, \alpha_{k_0}} \left| SURE_2(\alpha_1, \alpha_2, \ldots, \alpha_{k_0}) - L_2(\hat{\theta}, \theta) \right| \\
&\leq \sum_{i \in S^*} \left| X_i^2 - A_i - {\beta_i}^2 \right| + \left| \sum_{i \in S^*} \frac{2\lambda}{\lambda + A_i} (X_i^2 - \beta_i X_i - A_i) \right| \\
&\leq \sum_{i \in S^{*c}} \left| X_i^2 - A_i - {\beta_i}^2 \right| + \sup_{\lambda>0} \left| \sum_{i \in S^*} \frac{2\lambda}{\lambda + A_i} (X_i^2 - \beta_i X_i - A_i) \right|.
\end{aligned}$$

In this case, our hierarchal model on S^* is defined as

$$\begin{aligned}
X_i &\sim N(\beta_i, A_i), \\
\beta_i &\sim N(0, \lambda).
\end{aligned}$$

Obviously from Conditions C1) and C2), we have

$$\sup_{\alpha_1, \alpha_2, \ldots, \alpha_{k_0}} \frac{1}{n} \mid SURE_2(\alpha_1, \alpha_2, \ldots, \alpha_{k_0}) - L_2(\hat{\theta}, \theta) \mid \to 0 \quad \text{in } L^2, \text{ as } n \to \infty. \tag{A.2}$$

Equations (A.1) and (A.2) complete the proof.