Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models

S.C. Kou and Justin J. Yang

Abstract Shrinkage estimators have profound impacts in statistics and in scientific and engineering applications. In this article, we consider shrinkage estimation in the presence of linear predictors. We formulate two heteroscedastic hierarchical regression models and study optimal shrinkage estimators in each model. A class of shrinkage estimators, both parametric and semiparametric, based on unbiased risk estimate (URE) is proposed and is shown to be (asymptotically) optimal under mean squared error loss in each model. Simulation study is conducted to compare the performance of the proposed methods with existing shrinkage estimators. We also apply the method to real data and obtain encouraging and interesting results.

1 Introduction

Shrinkage estimators, hierarchical models and empirical Bayes methods, dating back to the groundbreaking works of [24] and [21], have profound impacts in statistics and in scientific and engineering applications. They provide effective tools to pool information from (scientifically) related populations for simultaneous inference—the data on each population alone often do not lead to the most effective estimation, but by pooling information from the related populations together (for example, by shrinking toward their consensus "center"), one could often obtain more accurate estimate for each individual population. Ever since the seminal works of [24] and [10], an impressive list of articles has been devoted to the study of shrinkage estimators in normal models, including [1, 2, 4–6, 8, 12, 14, 16, 22, 25], among others.

In this article, we consider shrinkage estimation in the presence of linear predictors. In particular, we study *optimal* shrinkage estimators for *heteroscedastic* data under *linear* models. Our study is motivated by three main considerations. First, in many practical problems, one often encounters heteroscedastic (unequal variance) data; for example, the sample sizes for different groups are not all equal. Second, in many statistical applications, in addition to the heteroscedastic response variable,

S.C. Kou (✉) • J.J. Yang
Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA
e-mail: kou@stat.harvard.edu; juchenjustinyang@fas.harvard.edu

S.E. Ahmed (ed.), *Big and Complex Data Analysis*, Contributions to Statistics,
DOI 10.1007/978-3-319-41573-4_13

one often has predictors. For example, the predictors could represent longitudinal patterns [7, 9, 27], exam scores [22], characteristics of hospital patients [18], etc. Third, in applying shrinkage estimators to real data, it is quite natural to ask for the *optimal* way of shrinkage.

The (risk) optimality is not addressed by the conventional estimators, such as the empirical Bayes ones. One might wonder if such an optimal shrinkage estimator exists in the first place. We shall see shortly that in fact (asymptotically) optimal shrinkage estimators do exist and that the optimal estimators are *not* empirical Bayes ones but are characterized by an unbiased risk estimate (URE).

The study of optimal shrinkage estimators under the heteroscedastic normal model was first considered in [29], where the (asymptotic) optimal shrinkage estimator was identified for both the parametric and semiparametric cases. Xie et al. [30] extends the (asymptotic) optimal shrinkage estimators to exponential families and heteroscedastic location-scale families. The current article can be viewed as an extension of the idea of optimal shrinkage estimators to heteroscedastic linear models.

We want to emphasize that this article works on a theoretical setting somewhat different from [30] but can still cover its main results. Our theoretical results show that the optimality of the proposed URE shrinkage estimators does not rely on normality nor on the tail behavior of the sampling distribution. What we require here are the symmetry and the existence of the fourth moment for the standardized variable.

This article is organized as follows. We first formulate the heteroscedastic linear models in Sect. 2. Interestingly, there are two parallel ways to do so, and both are natural extensions of the heteroscedastic normal model. After reviewing the conventional empirical Bayes methods, we introduce the construction of our optimal shrinkage estimators for heteroscedastic linear models in Sect. 3. The optimal shrinkage estimators are based on an unbiased risk estimate (URE). We show in Sect. 4 that the URE shrinkage estimators are asymptotically optimal in risk. In Sect. 5 we extend the shrinkage estimators to a semiparametric family. Simulation studies are conducted in Sect. 6. We apply the URE shrinkage estimators in Sect. 7 to the baseball data set of [2] and observe quite interesting and encouraging results. We conclude in Sect. 8 with some discussion and extension. The appendix details the proofs and derivations for the theoretical results.

2 Heteroscedastic Hierarchical Linear Models

Consider the heteroscedastic estimation problem

$$Y_i|\boldsymbol{\theta} \overset{\text{indep.}}{\sim} \mathcal{N}(\theta_i, A_i), \quad i = 1, \ldots, p, \tag{1}$$

where $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_p\right)^T$ is the unknown mean vector, which is to be estimated, and the variances $A_i > 0$ are unequal, which are assumed to be known. In many statistical applications, in addition to the heteroscedastic $\boldsymbol{Y} = \left(Y_1, \ldots, Y_p\right)^T$, one often has predictors $\boldsymbol{X}$. A natural question is to consider a heteroscedastic linear model that incorporates these covariates. Notation-wise, let $\{Y_i, \boldsymbol{X}_i\}_{i=1}^p$ denote the p independent statistical units, where Y_i is the response variable of the i-th unit, and $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{ki})^T$ is a k-dimensional column vector that corresponds to the k covariates of the i-th unit. The $k \times p$ matrix

$$\boldsymbol{X} = \left[\boldsymbol{X}_1 | \cdots | \boldsymbol{X}_p\right], \quad \boldsymbol{X}_1, .., \boldsymbol{X}_p \in \mathbb{R}^k,$$

where $\boldsymbol{X}_i$ is the i-th column of $\boldsymbol{X}$, then contains the covariates for all the units. Throughout this article we assume that $\boldsymbol{X}$ has full rank, i.e., $\text{rank}(\boldsymbol{X}) = k$.

To include the predictors, we note that, interestingly, there are *two* different ways to build up a heteroscedastic hierarchical linear model, which lead to different structure for shrinkage estimation.

Model I: Hierarchical linear model. On top of (1), the θ_i's are $\theta_i \overset{\text{indep.}}{\sim} \mathcal{N}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}, \lambda\right)$, where $\boldsymbol{\beta}$ and λ are both *unknown* hyper-parameters. Model I has been suggested as early as [26]. See [16] and [17] for more discussions. The special case of no covariates (i.e., $k = 1$ and $\boldsymbol{X} = [1|\cdots|1]$) is studied in depth in [29].

Model II: Bayesian linear regression model. Together with (1), one assumes $\boldsymbol{\theta} = \boldsymbol{X}^T \boldsymbol{\beta}$ with $\boldsymbol{\beta}$ following a conjugate prior distribution $\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{\beta}_0, \lambda \boldsymbol{W}\right)$, where $\boldsymbol{W}$ is a *known* $k \times k$ positive definite matrix and $\boldsymbol{\beta}_0$ and λ are *unknown* hyper-parameters. Model II has been considered in [3, 15, 20] among others; it includes ridge regression as a special case when $\boldsymbol{\beta}_0 = \boldsymbol{0}_k$ and $\boldsymbol{W} = \boldsymbol{I}_k$.

Figure 1 illustrates these two hierarchical linear models. Under Model I, the posterior mean of $\boldsymbol{\theta}$ is $\hat{\theta}_i^{\lambda,\boldsymbol{\beta}} = \lambda\left(\lambda + A_i\right)^{-1} Y_i + A_i\left(\lambda + A_i\right)^{-1} \boldsymbol{X}_i^T \boldsymbol{\beta}$ for $i = 1, \ldots, p$, so the shrinkage estimation is formed by directly shrinking the raw observation Y_i toward a linear combination of the k covariates $\boldsymbol{X}_i$. If we denote $\mu_i = \boldsymbol{X}_i^T \boldsymbol{\beta}$,

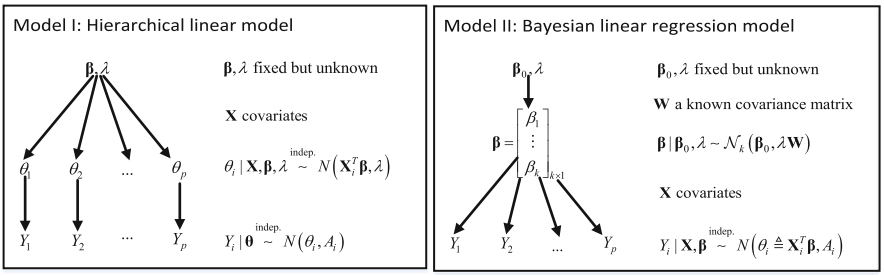

Fig. 1 Graphical illustration of the two heteroscedastic hierarchical linear models

and $\boldsymbol{\mu} = \left(\mu_1, \ldots, \mu_p\right)^T \in \mathscr{L}_{\text{row}}(\boldsymbol{X})$, the row space of $\boldsymbol{X}$, then we can rewrite the posterior mean of $\boldsymbol{\theta}$ under Model I as

$$\hat{\theta}^{\lambda,\boldsymbol{\mu}} = \frac{\lambda}{\lambda + A_i} Y_i + \frac{A_i}{\lambda + A_i} \mu_i, \quad \text{with } \boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X}). \tag{2}$$

Under Model II, the posterior mean of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\beta}_0} = \boldsymbol{X}^T \hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0}, \quad \text{with } \hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0} = \lambda \boldsymbol{W}(\lambda \boldsymbol{W} + \boldsymbol{V})^{-1} \hat{\boldsymbol{\beta}}^{\text{WLS}} + \boldsymbol{V}(\lambda \boldsymbol{W} + \boldsymbol{V})^{-1} \boldsymbol{\beta}_0, \tag{3}$$

where $\hat{\boldsymbol{\beta}}^{\text{WLS}} = \left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)^{-1} \boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Y}$ is the weighted least squares estimate of the regression coefficient, $\boldsymbol{A}$ is the diagonal matrix $\boldsymbol{A} = \text{diag}\left(A_1, \ldots, A_p\right)$, and $\boldsymbol{V} = (\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T)^{-1}$. Thus, the estimate for θ_i is linear in $\boldsymbol{X}_i$, and the "shrinkage" is achieved by shrinking the regression coefficient from the weighted least squares estimate $\hat{\boldsymbol{\beta}}^{\text{WLS}}$ toward the prior coefficient $\boldsymbol{\beta}_0$.

As both Models I and II are natural generalizations of the heteroscedastic normal model (1), we want to investigate if there is an optimal choice of the hyper-parameters in each case. Specifically, we want to investigate the best empirical choice of the hyper-parameters in each case under the mean squared error loss

$$l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{p} \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|^2 = \frac{1}{p} \sum_{i=1}^{p} \left(\theta_i - \hat{\theta}_i\right)^2 \tag{4}$$

with the associated risk of $\hat{\boldsymbol{\theta}}$ defined by

$$R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}} \left(l_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \right),$$

where the expectation is taken with respect to $\boldsymbol{Y}$ given $\boldsymbol{\theta}$.

Remark 1 Even though we start from the Bayesian setting to motivate the form of shrinkage estimators, our discussion will be all based on the frequentist setting. Hence all probabilities and expectations throughout this article are fixed at the unknown true $\boldsymbol{\theta}$, which is free in $\mathbb{R}^p$ for Model I and confined in $\mathscr{L}_{\text{row}}(\boldsymbol{X})$ for Model II.

Remark 2 The diagonal assumption of $\boldsymbol{A}$ is quite important for Model I but not so for Model II, as in Model II we can always apply some linear transformations to obtain a diagonal covariance matrix. Without loss of generality, we will keep the diagonal assumption for $\boldsymbol{A}$ in Model II.

For the ease of exposition, we will next overview the conventional empirical Bayes estimates in a general two-level hierarchical model, which includes both

Models I and II:

$$\boldsymbol{Y}|\boldsymbol{\theta} \sim \mathscr{N}_p(\boldsymbol{\theta}, \boldsymbol{A}) \text{ and } \boldsymbol{\theta} \sim \mathscr{N}_p(\boldsymbol{\mu}, \boldsymbol{B}), \tag{5}$$

where $\boldsymbol{B}$ is a non-negative definite symmetric matrix that is restricted in an allowable set $\mathscr{B}$, and $\boldsymbol{\mu}$ is in the row space $\mathscr{L}_{\text{row}}(\boldsymbol{X})$ of $\boldsymbol{X}$.

Remark 3 Under Model I, $\boldsymbol{\mu}$ and $\boldsymbol{B}$ take the form of $\boldsymbol{\mu} = \boldsymbol{X}^T\boldsymbol{\beta}$ and $\boldsymbol{B} \in \mathscr{B} = \{\lambda \boldsymbol{I}_p : \lambda > 0\}$, whereas under Model II, $\boldsymbol{\mu}$ and $\boldsymbol{B}$ take the form of $\boldsymbol{\mu} = \boldsymbol{X}^T\boldsymbol{\beta}_0$ and $\boldsymbol{B} \in \mathscr{B} = \{\lambda \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} : \lambda > 0\}$. It is interesting to observe that in Model I, $\boldsymbol{B}$ is of full rank, while in Model II, $\boldsymbol{B}$ is of rank k. As we shall see, this distinction will have interesting theoretical implications for the optimal shrinkage estimators.

Lemma 1 *Under the two-level hierarchical model (5), the posterior distribution is*

$$\boldsymbol{\theta}|\boldsymbol{Y} \sim \mathscr{N}_p\left(\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{\mu}, \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}\right),$$

and the marginal distribution of $\boldsymbol{Y}$ *is* $\boldsymbol{Y} \sim \mathscr{N}_p(\boldsymbol{\mu}, \boldsymbol{A}+\boldsymbol{B})$.

For given values of $\boldsymbol{B}$ and $\boldsymbol{\mu}$, the posterior mean of the parameter $\boldsymbol{\theta}$ leads to the Bayes estimate

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}} = \boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{\mu}. \tag{6}$$

To use the Bayes estimate in practice, one has to specify the hyper-parameters in $\boldsymbol{B}$ and $\boldsymbol{\mu}$. The conventional empirical Bayes method uses the marginal distribution of $\boldsymbol{Y}$ to estimate the hyper-parameters. For instance, the empirical Bayes maximum likelihood estimates (EBMLE) $\hat{\boldsymbol{B}}^{\text{EBMLE}}$ and $\hat{\boldsymbol{\mu}}^{\text{EBMLE}}$ are obtained by maximizing the marginal likelihood of $\boldsymbol{Y}$:

$$\left(\hat{\boldsymbol{B}}^{\text{EBMLE}}, \hat{\boldsymbol{\mu}}^{\text{EBMLE}}\right) = \underset{\substack{\boldsymbol{B}\in\mathscr{B} \\ \boldsymbol{\mu}\in\mathscr{L}_{\text{row}}(\boldsymbol{X})}}{\text{argmax}} \; -(\boldsymbol{Y}-\boldsymbol{\mu})^T(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) - \log\left(\det(\boldsymbol{A}+\boldsymbol{B})\right).$$

Alternatively, the empirical Bayes method-of-moment estimates (EBMOM) $\hat{\boldsymbol{B}}^{\text{EBMOM}}$ and $\hat{\boldsymbol{\mu}}^{\text{EBMOM}}$ are obtained by solving the following moment equations for $\boldsymbol{B} \in \mathscr{B}$ and $\boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X})$:

$$\boldsymbol{\mu} = \boldsymbol{X}^T\left(\boldsymbol{X}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{Y},$$

$$\boldsymbol{B} = (\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T - \boldsymbol{A}.$$

If no solutions of $\boldsymbol{B}$ can be found in $\mathscr{B}$, we then set $\hat{\boldsymbol{B}}^{\text{EBMOM}} = \boldsymbol{0}_{p\times p}$. Adjustment for the loss of k degrees of freedom from the estimation of $\boldsymbol{\mu}$ might be applicable for $\boldsymbol{B} = \lambda\boldsymbol{C}$ ($\boldsymbol{C} = \boldsymbol{I}_p$ for Model I and $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ for Model II): we can replace the second

moment equation by

$$\lambda = \left(\frac{p}{p-k} \frac{\|\boldsymbol{Y} - \boldsymbol{\mu}\|^2}{\mathrm{tr}\,(\boldsymbol{C})} - \frac{\mathrm{tr}\,(\boldsymbol{A})}{\mathrm{tr}\,(\boldsymbol{C})} \right)^+ .$$

The corresponding empirical Bayes shrinkage estimator $\hat{\boldsymbol{\theta}}^{\mathrm{EBMLE}}$ or $\hat{\boldsymbol{\theta}}^{\mathrm{EBMOM}}$ is then formed by plugging $(\hat{\boldsymbol{B}}^{\mathrm{EBMLE}}, \hat{\boldsymbol{\mu}}^{\mathrm{EBMLE}})$ or $(\hat{\boldsymbol{B}}^{\mathrm{EBMOM}}, \hat{\boldsymbol{\mu}}^{\mathrm{EBMOM}})$ into Eq. (6).

3 URE Estimates

The formulation of the empirical Bayes estimates raises a natural question: which one is preferred $\hat{\boldsymbol{\theta}}^{\mathrm{EBMLE}}$ or $\hat{\boldsymbol{\theta}}^{\mathrm{EBMOM}}$? More generally, is there an optimal way to choose the hyper-parameters? It turns out that neither $\hat{\boldsymbol{\theta}}^{\mathrm{EBMLE}}$ nor $\hat{\boldsymbol{\theta}}^{\mathrm{EBMOM}}$ is optimal. The (asymptotically) optimal estimate, instead of relying on the marginal distribution of $\boldsymbol{Y}$, is characterized by an unbiased risk estimate (URE). The idea of forming a shrinkage estimate through URE for heteroscedastic models is first suggested in [29]. We shall see that in our context of hierarchical linear models (both Models I and II) the URE estimators that we are about to introduce have (asymptotically) optimal risk properties.

The basic idea behind URE estimators is the following. Ideally we want to find the hyper-parameters that give the smallest risk. However, since the risk function depends on the unknown $\boldsymbol{\theta}$, we cannot directly minimize the risk function in practice. If we can find a good estimate of the risk function instead, then minimizing this proxy of the risk will lead to a competitive estimator.

To formally introduce the URE estimators, we start from the observation that, under the mean squared error loss (4), the risk of the Bayes estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}$ for fixed $\boldsymbol{B}$ and $\boldsymbol{\mu}$ is

$$R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}) = \frac{1}{p} \left\| \boldsymbol{A}\,(\boldsymbol{A}+\boldsymbol{B})^{-1}\,(\boldsymbol{\mu}-\boldsymbol{\theta}) \right\|^2 + \frac{1}{p}\mathrm{tr}\left(\boldsymbol{B}\,(\boldsymbol{A}+\boldsymbol{B})^{-1}\,\boldsymbol{A}\,(\boldsymbol{A}+\boldsymbol{B})^{-1}\,\boldsymbol{B} \right), \tag{7}$$

which can be easily shown using the bias-variance decomposition of the mean squared error. As the risk function involves the unknown $\boldsymbol{\theta}$, we cannot directly minimize it. However, an unbiased estimate of the risk is available:

$$\mathrm{URE}\,(\boldsymbol{B}, \boldsymbol{\mu}) = \frac{1}{p} \left\| \boldsymbol{A}\,(\boldsymbol{A}+\boldsymbol{B})^{-1}\,(\boldsymbol{Y}-\boldsymbol{\mu}) \right\|^2 + \frac{1}{p}\mathrm{tr}\left(\boldsymbol{A} - 2\boldsymbol{A}\,(\boldsymbol{A}+\boldsymbol{B})^{-1}\,\boldsymbol{A} \right), \tag{8}$$

which again can be easily shown using the bias-variance decomposition of the mean squared error. Intuitively, if URE $(\boldsymbol{B}, \boldsymbol{\mu})$ is a good approximation of the actual risk, then we would expect the estimator obtained by minimizing the URE to have good

properties. This leads to the URE estimator $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}$, defined by

$$\hat{\boldsymbol{\theta}}^{\mathrm{URE}} = \hat{\boldsymbol{B}}^{\mathrm{URE}}(\boldsymbol{A} + \hat{\boldsymbol{B}}^{\mathrm{URE}})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A} + \hat{\boldsymbol{B}}^{\mathrm{URE}})^{-1}\hat{\boldsymbol{\mu}}^{\mathrm{URE}}, \tag{9}$$

where

$$\left(\hat{\boldsymbol{B}}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}^{\mathrm{URE}}\right) = \underset{B\in\mathscr{B},\ \boldsymbol{\mu}\in\mathscr{L}_{\mathrm{row}}(\boldsymbol{X})}{\operatorname{argmin}} \mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{\mu}\right).$$

It is worth noting that the value of $\boldsymbol{\mu}$ that minimizes (8) for a given $\boldsymbol{B}$ is neither the ordinary least squares (OLS) nor the weighted least squares (WLS) regression estimate, echoing similar observation as in [29].

In the URE estimator (9), $\hat{\boldsymbol{B}}^{\mathrm{URE}}$ and $\hat{\boldsymbol{\mu}}^{\mathrm{URE}}$ are jointly determined by minimizing the URE. When the number of independent statistical units p is small or moderate, joint minimization of $\boldsymbol{B}$ and the vector $\boldsymbol{\mu}$, however, may be too ambitious. In this setting, it might be beneficial to set $\boldsymbol{\mu}$ by a predetermined rule and only optimize $\boldsymbol{B}$, as it might reduce the variability of the resulting estimate. In particular, we can consider shrinking toward a generalized least squares (GLS) regression estimate

$$\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} = \boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{Y},$$

where $\boldsymbol{M}$ is a *prespecified* symmetric positive definite matrix. This use of $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}$ gives the shrinkage estimate $\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}} = \boldsymbol{B}(\boldsymbol{A} + \boldsymbol{B})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A} + \boldsymbol{B})^{-1}\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}$, where one only needs to determine $\boldsymbol{B}$. We can construct another URE estimate for this purpose. Similar to the previous construction, we note that $\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}$ has risk

$$\begin{aligned} R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}) =& \frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\right)\boldsymbol{\theta}\right\|^2 \\ &+ \frac{1}{p}\mathrm{tr}\left(\left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\right)\right)\boldsymbol{A} \right. \\ &\left. \times \left(\boldsymbol{I}_p - \boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\right)\right)^T\right). \end{aligned} \tag{10}$$

An unbiased risk estimate of it is

$$\mathrm{URE}_{\boldsymbol{M}}\left(\boldsymbol{B}\right) = \frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}\right)\right\|^2 + \frac{1}{p}\mathrm{tr}\left(\boldsymbol{A} - 2\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\right)\boldsymbol{A}\right). \tag{11}$$

Both (10) and (11) can be easily proved by the bias-variance decomposition of mean squared error. Minimizing $\mathrm{URE}_{\boldsymbol{M}}\left(\boldsymbol{B}\right)$ over $\boldsymbol{B}$ gives the URE GLS shrinkage

estimator (which shrinks toward $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}$):

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}} = \hat{\boldsymbol{B}}_{\boldsymbol{M}}^{\mathrm{URE}} \left(\boldsymbol{A} + \hat{\boldsymbol{B}}_{\boldsymbol{M}}^{\mathrm{URE}}\right)^{-1} \boldsymbol{Y} + \boldsymbol{A} \left(\boldsymbol{A} + \hat{\boldsymbol{B}}_{\boldsymbol{M}}^{\mathrm{URE}}\right)^{-1} \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}, \tag{12}$$

where

$$\hat{\boldsymbol{B}}_{\boldsymbol{M}}^{\mathrm{URE}} = \underset{\boldsymbol{B} \in \mathscr{B}}{\operatorname{argmin}} \, \mathrm{URE}_{\boldsymbol{M}} (\boldsymbol{B}) .$$

Remark 4 When $\boldsymbol{M} = \boldsymbol{I}_p$, clearly $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} = \hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$, the ordinary least squares regression estimate. When $\boldsymbol{M} = \boldsymbol{A}^{-1}$, then $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} = \hat{\boldsymbol{\mu}}^{\mathrm{WLS}}$, the weighted least squares regression estimate.

Remark 5 Tan [28] briefly discussed the URE minimization approach for Model I without the covariates in [29] in relation to [11], where Model I is assumed but an unbiased estimate of the mean prediction error (rather than the mean squared error) is used to form a predictor (rather than an estimator).

Remark 6 In the homoscedastic case, (12) reduces to standard shrinkage toward a subspace $\mathscr{L}_{\mathrm{row}} (\boldsymbol{X})$, as discussed, for instance, in [23] and [19].

4 Theoretical Properties of URE Estimates

This section is devoted to the risk properties of the URE estimators. Our core theoretical result is to show that the risk estimate URE is not only unbiased for the risk but, more importantly, uniformly close to the actual loss. We therefore expect that minimizing URE would lead to an estimate with competitive risk properties.

4.1 Uniform Convergence of URE

To present our theoretical result, we first define $\mathscr{L}$ to be a subset of $\mathscr{L}_{\mathrm{row}} (\boldsymbol{X})$:

$$\mathscr{L} = \{\boldsymbol{\mu} \in \mathscr{L}_{\mathrm{row}} (\boldsymbol{X}) : \|\boldsymbol{\mu}\| \leq M p^{\kappa} \, \|\boldsymbol{Y}\|\},$$

where M is a large and fixed constant and $\kappa \in [0, 1/2)$ is a constant. Next, we introduce the following regularity conditions:

(A) $\sum_{i=1}^{p} A_i^2 = O(p)$; (B) $\sum_{i=1}^{p} A_i \theta_i^2 = O(p)$; (C) $\sum_{i=1}^{p} \theta_i^2 = O(p)$;
(D) $p^{-1} \boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^T \to \boldsymbol{\Omega}_D$; (E) $p^{-1} \boldsymbol{X} \boldsymbol{X}^T \to \boldsymbol{\Omega}_E > 0$;
(F) $p^{-1} \boldsymbol{X} \boldsymbol{A}^{-1} \boldsymbol{X}^T \to \boldsymbol{\Omega}_F > 0$; (G) $p^{-1} \boldsymbol{X} \boldsymbol{A}^{-2} \boldsymbol{X}^T \to \boldsymbol{\Omega}_G$.

The theorem below shows that URE $(\boldsymbol{B}, \boldsymbol{\mu})$ not only unbiasedly estimates the risk but also is (asymptotically) uniformly close to the actual loss.

Theorem 1 *Assume conditions (A)–(E) for Model I or assume conditions (A) and (D)–(G) for Model II. In either case, we have*

$$\sup_{\boldsymbol{B}\in\mathscr{B},\ \boldsymbol{\mu}\in\mathscr{L}} \left|\mathrm{URE}\left(\boldsymbol{B},\boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)\right| \to 0 \text{ in } L^1, \text{ as } p\to\infty.$$

We want to remark here that the set $\mathscr{L}$ gives the allowable range of $\boldsymbol{\mu}$: the norm of $\boldsymbol{\mu}$ is up to an $o\left(p^{1/2}\right)$ multiple of the norm of $\boldsymbol{Y}$. This choice of $\mathscr{L}$ does not lead to any difficulty in practice because, given a large enough constant M, it will cover the shrinkage location of any sensible shrinkage estimator. We note that it is possible to define the range of sensible shrinkage locations in other ways (e.g., one might want to define it by ∞-norm in $\mathbb{R}^p$), but we find our setting more theoretically appealing and easy to work with. In particular, our assumption of the exponent $\kappa < 1/2$ is flexible enough to cover most interesting cases, including $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$, the ordinary least squares regression estimate, and $\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}$, the weighted least squares regression estimate (as in Remark 4) as shown in the following lemma.

Lemma 2 *(i)* $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}} \in \mathscr{L}$. *(ii) Assume* (A) *and* (A′) $\sum_{i=1}^{p} A_i^{-2-\delta} = O\left(p\right)$ *for some* $\delta > 0$*; then* $\hat{\boldsymbol{\mu}}^{\mathrm{WLS}} \in \mathscr{L}$ *for* $\kappa = 4^{-1} + (4+2\delta)^{-1}$ *and a large enough* M.

Remark 7 We want to mention here that Theorem 1 in the case of Model I covers Theorem 5.1 of [29] (which is the special case of $k = 1$ and $\boldsymbol{X} = [1|1|\ldots|1]$) because the restriction of $|\mu| \le \max_{1\le i\le p} |Y_i|$ in [29] is contained in $\mathscr{L}$ as

$$\max_{1\le i\le p} |Y_i| = (\max_{1\le i\le p} Y_i^2)^{1/2} \le (\sum_{i=1}^{p} Y_i^2)^{1/2} = \|\boldsymbol{Y}\|.$$

Furthermore, we do not require the stronger assumption of $\sum_{i=1}^{p} |\theta_i|^{2+\delta} = O\left(p\right)$ for some $\delta > 0$ made in [29]. Note that in this case ($k = 1$ and $\boldsymbol{X} = [1|1|\ldots|1]$) we do not even require conditions (D) and (E), as condition (A) directly implies $\mathrm{tr}(\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T) = O\left(1\right)$, the result we need in the proof of Theorem 1 for Model I.

Remark 8 In the proof of Theorem 1, the sampling distribution of $\boldsymbol{Y}$ is involved only through the moment calculations, such as $\mathbb{E}(\mathrm{tr}(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{A} - \boldsymbol{\theta}\boldsymbol{\theta}^T)^2)$ and $\mathbb{E}(\|\boldsymbol{Y}\|^2)$. It is therefore straightforward to generalize Theorem 1 to the case of

$$Y_i = \theta_i + \sqrt{A_i}Z_i,$$

where Z_i follows *any* distribution with mean 0, variance 1, $\mathbb{E}\left(Z_i^3\right) = 0$, and $\mathbb{E}\left(Z_i^4\right) < \infty$. This is noteworthy as our result also covers that of [30] but the methodology we employ here does not require to control the tail behavior of Z_i as in [29, 30].

4.2 Risk Optimality

In this section, we consider the risk properties of the URE estimators. We will show that, under the hierarchical linear models, the URE estimators have (asymptotically) optimal risk, whereas it is not necessarily so for other shrinkage estimators such as the empirical Bayes ones.

A direct consequence of the uniform convergence of URE is that the URE estimator has a loss/risk that is asymptotically no larger than that of any other shrinkage estimators. Furthermore, the URE estimator is asymptotically as good as the oracle loss estimator. To be precise, let $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ be the oracle loss (OL) estimator defined by plugging

$$\begin{aligned}\left(\tilde{\boldsymbol{B}}^{\mathrm{OL}}, \tilde{\boldsymbol{\mu}}^{\mathrm{OL}}\right) &= \underset{B\in\mathscr{B},\ \boldsymbol{\mu}\in\mathscr{L}}{\operatorname{argmin}}\ l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)\\ &= \underset{B\in\mathscr{B},\ \boldsymbol{\mu}\in\mathscr{L}}{\operatorname{argmin}}\ \left\|\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{Y} + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{\mu} - \boldsymbol{\theta}\right\|^2\end{aligned}$$

into (6). Of course, $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ is not really an estimator, since it depends on the unknown $\boldsymbol{\theta}$ (hence we use the notation $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ rather than $\hat{\boldsymbol{\theta}}^{\mathrm{OL}}$). Although not obtainable in practice, $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ lays down the theoretical limit that one can ever hope to reach. The next theorem shows that the URE estimator $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}$ is asymptotically as good as the oracle loss estimator, and, consequently, it is asymptotically at least as good as any other shrinkage estimator.

Theorem 2 *Assume the conditions of Theorem 1 and that* $\hat{\boldsymbol{\mu}}^{\mathrm{URE}} \in \mathscr{L}$. *Then*

$$\begin{aligned}&\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) + \epsilon\right) = 0 \quad \forall \epsilon > 0,\\ &\limsup_{p\to\infty}\left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right)\right) = 0.\end{aligned}$$

Corollary 1 *Assume the conditions of Theorem 1 and that* $\hat{\boldsymbol{\mu}}^{\mathrm{URE}} \in \mathscr{L}$. *Then for any estimator* $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p,\hat{\boldsymbol{\mu}}_p} = \hat{\boldsymbol{B}}_p\left(\boldsymbol{A}+\hat{\boldsymbol{B}}_p\right)^{-1}\boldsymbol{Y} + \boldsymbol{A}\left(\boldsymbol{A}+\hat{\boldsymbol{B}}_p\right)^{-1}\hat{\boldsymbol{\mu}}_p$ *with* $\hat{\boldsymbol{B}}_p \in \mathscr{B}$ *and* $\hat{\boldsymbol{\mu}}_p \in \mathscr{L}$, *we always have*

$$\begin{aligned}&\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p,\hat{\boldsymbol{\mu}}_p}\right) + \epsilon\right) = 0 \quad \forall \epsilon > 0,\\ &\limsup_{p\to\infty}\left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p,\hat{\boldsymbol{\mu}}_p}\right)\right) \leq 0.\end{aligned}$$

Corollary 1 tells us that the URE estimator in either Model I or II is asymptotically optimal: it has (asymptotically) the smallest loss and risk among all shrinkage estimators of the form (6).

4.3 Shrinkage Toward the Generalized Least Squares Estimate

The risk optimality also holds when we consider the URE estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}$ that shrinks toward the GLS regression estimate $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} = \boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{Y}$ as introduced in Sect. 3.

Theorem 3 *Assume the conditions of Theorem 1,* $\hat{\boldsymbol{\mu}}^{\boldsymbol{M}} \in \mathscr{L}$*, and*

$$p^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T \to \boldsymbol{\Omega}_1 > 0, \quad p^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T \to \boldsymbol{\Omega}_2, \quad p^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}^2\boldsymbol{M}\boldsymbol{X}^T \to \boldsymbol{\Omega}_3, \tag{13}$$

where only the first and third conditions above are assumed for Model I and only the first and the second are assumed for Model II. Then we have

$$\sup_{\boldsymbol{B}\in\mathscr{B}} \left| \mathrm{URE}_{\boldsymbol{M}}(\boldsymbol{B}) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right) \right| \to 0 \text{ in } L^1 \text{ as } p \to \infty. \tag{14}$$

As a corollary, for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p,\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}} = \hat{\boldsymbol{B}}_p\left(\boldsymbol{A}+\hat{\boldsymbol{B}}_p\right)^{-1}\boldsymbol{Y} + \boldsymbol{A}\left(\boldsymbol{A}+\hat{\boldsymbol{B}}_p\right)^{-1}\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}$ *with* $\hat{\boldsymbol{B}}_p \in \mathscr{B}$*, we always have*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p,\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right) + \epsilon \right) = 0 \quad \forall \epsilon > 0,$$

$$\limsup_{p\to\infty} \left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p,\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right) \right) \leq 0.$$

Remark 9 For shrinking toward $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$, where $\boldsymbol{M} = \boldsymbol{I}_p$, we know from Lemma 2 that $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$ is automatically in $\mathscr{L}$, so we only need one more condition $p^{-1}\boldsymbol{X}\boldsymbol{A}^2\boldsymbol{X}^T \to \boldsymbol{\Omega}_3$ for Model I. For shrinking toward $\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}$, where $\boldsymbol{M} = \boldsymbol{A}^{-1}$, (13) is the same as the conditions (E) and (F) of Theorem 1, so additionally we only need to assume (A′) of Lemma 2 and (F) for Model I.

5 Semiparametric URE Estimators

We have established the (asymptotic) optimality of the URE estimators $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}$ in the previous section. One limitation of the result is that the class over which the URE estimators are optimal is specified by a parametric form: $\boldsymbol{B} = \lambda\boldsymbol{C}$ $(0 \leq \lambda \leq$

∞) in Eq. (6), where $\boldsymbol{C} = \boldsymbol{I}_p$ for Model I and $\boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ for Model II. Aiming to provide a more flexible and, at the same time, efficient estimation procedure, we consider in this section a class of semiparametric shrinkage estimators. Our consideration is inspired by Xie et al. [29].

5.1 Semiparametric URE Estimator Under Model I

To motivate the semiparametric shrinkage estimators, let us first revisit the Bayes estimator $\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\mu}}$ under Model I, as given in (2). It is seen that the Bayes estimate of each mean parameter θ_i is obtained by shrinking Y_i toward the linear estimate $\mu_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$, and that the amount of shrinkage is governed by A_i, the variance: the larger the variance, the stronger is the shrinkage. This feature makes intuitive sense.

With this observation in mind, we consider the following shrinkage estimators under Model I:

$$\hat{\theta}_i^{\boldsymbol{b},\boldsymbol{\mu}} = (1 - b_i)\, Y_i + b_i\mu_i, \quad \text{with } \boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X}),$$

where $\boldsymbol{b}$ satisfies the monotonic constraint

$$\text{MON}(\boldsymbol{A}) : b_i \in [0, 1],\ b_i \le b_j \text{ whenever } A_i \le A_j.$$

MON ($\boldsymbol{A}$) asks the estimator to shrink more for an observation with a larger variance. Since other than this intuitive requirement, we do not post any parametric restriction on b_i, this class of estimators is semiparametric in nature.

Following the optimality result for the parametric case, we want to investigate, for such a general estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}}$ with $\boldsymbol{b} \in \text{MON}(\boldsymbol{A})$ and $\boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X})$, whether there exists an optimal choice of $\boldsymbol{b}$ and $\boldsymbol{\mu}$. In fact, we will see shortly that such an optimal choice exists, and this asymptotically optimal choice is again characterized by an unbiased risk estimate (URE). For a general estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}}$ with *fixed* $\boldsymbol{b}$ and $\boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X})$, an unbiased estimate of its risk $R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}})$ is

$$\text{URE}^{SP}(\boldsymbol{b}, \boldsymbol{\mu}) = \frac{1}{p}\left\|\text{diag}(\boldsymbol{b})(\boldsymbol{Y} - \boldsymbol{\mu})\right\|^2 + \frac{1}{p}\text{tr}(\boldsymbol{A} - 2\text{diag}(\boldsymbol{b})\boldsymbol{A}),$$

which can be easily seen by taking $\boldsymbol{B} = \boldsymbol{A}(\text{diag}(\boldsymbol{b})^{-1} - \boldsymbol{I}_p)$ in (8). Note that we use the superscript "*SP*" (semiparametric) to denote it. Minimizing over $\boldsymbol{b}$ and $\boldsymbol{\mu}$ leads to the semiparametric URE estimator $\hat{\boldsymbol{\theta}}_{SP}^{\text{URE}}$, defined by

$$\hat{\boldsymbol{\theta}}_{SP}^{\text{URE}} = (\boldsymbol{I}_p - \text{diag}(\hat{\boldsymbol{b}}_{SP}^{\text{URE}}))\boldsymbol{Y} + \text{diag}(\hat{\boldsymbol{b}}_{SP}^{\text{URE}})\hat{\boldsymbol{\mu}}_{SP}^{\text{URE}}, \tag{15}$$

where

$$\left(\hat{\boldsymbol{b}}_{SP}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}_{SP}^{\mathrm{URE}}\right) = \underset{\boldsymbol{b}\in \mathrm{MON}(\boldsymbol{A}),\ \boldsymbol{\mu}\in\mathscr{L}_{\mathrm{row}}(\boldsymbol{X})}{\operatorname{argmin}} \mathrm{URE}^{SP}\left(\boldsymbol{b}, \boldsymbol{\mu}\right).$$

Theorem 4 *Assume conditions (A)–(E). Then under Model I we have*

$$\sup_{\boldsymbol{b}\in \mathrm{MON}(\boldsymbol{A}),\ \boldsymbol{\mu}\in\mathscr{L}} \left|\mathrm{URE}^{SP}\left(\boldsymbol{b}, \boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}}\right)\right| \to 0 \text{ in } L^1 \text{ as } p \to \infty.$$

As a corollary, for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\mu}}_p} = (\boldsymbol{I}_p - \mathrm{diag}(\hat{\boldsymbol{b}}_p))\boldsymbol{Y} + \mathrm{diag}(\hat{\boldsymbol{b}}_p)\hat{\boldsymbol{\mu}}_p$ *with* $\hat{\boldsymbol{b}}_p \in \mathrm{MON}\left(\boldsymbol{A}\right)$ *and* $\hat{\boldsymbol{\mu}}_p \in \mathscr{L}$*, we always have*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\mu}}_p}\right) + \epsilon\right) = 0 \quad \forall \epsilon > 0,$$

$$\limsup_{p\to\infty}\left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\mu}}_p}\right)\right) \leq 0.$$

The proof is the same as the proofs of Theorem 1 and Corollary 1 for the case of Model I except that we replace each term of $A_i/(\lambda + A_i)$ by b_i.

5.2 Semiparametric URE Estimator Under Model II

We saw in Sect. 2 that, under Model II, shrinkage is achieved by shrinking the regression coefficient from the weighted least squares estimate $\hat{\boldsymbol{\beta}}^{\mathrm{WLS}}$ toward the prior coefficient $\boldsymbol{\beta}_0$. This suggests us to formulate the semiparametric estimators through the regression coefficient. The Bayes estimate of the regression coefficient is

$$\hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0} = \lambda \boldsymbol{W}(\lambda \boldsymbol{W} + \boldsymbol{V})^{-1}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \boldsymbol{V}\left(\lambda \boldsymbol{W} + \boldsymbol{V}\right)^{-1}\boldsymbol{\beta}_0, \quad \text{with } \boldsymbol{V} = (\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T)^{-1}$$

as shown in (3). Applying the spectral decomposition on $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}$ gives $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda} = \mathrm{diag}\left(d_1, \ldots, d_k\right)$ with $d_1 \leq \cdots \leq d_k$. Using this decomposition, we can rewrite the regression coefficient as

$$\hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0} = \lambda \boldsymbol{W}^{1/2}\boldsymbol{U}\left(\lambda \boldsymbol{I}_k + \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \boldsymbol{W}^{1/2}\boldsymbol{U}\boldsymbol{\Lambda}\left(\lambda \boldsymbol{I}_k + \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0.$$

If we denote $\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{W}^{1/2}\boldsymbol{X}$ as the transformed covariate matrix, the estimate $\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\beta}_0} = \boldsymbol{X}^T\hat{\boldsymbol{\beta}}^{\lambda,\boldsymbol{\beta}_0}$ of $\boldsymbol{\theta}$ can be rewritten as

$$\hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\beta}_0} = \boldsymbol{Z}^T\left(\lambda\left(\lambda\boldsymbol{I}_k + \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \boldsymbol{\Lambda}\left(\lambda\boldsymbol{I}_k + \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0\right).$$

Now we see that $\lambda\left(\lambda\boldsymbol{I}_k + \boldsymbol{\Lambda}\right)^{-1} = \mathrm{diag}(\lambda/\left(\lambda + d_i\right))$ plays the role as the shrinkage factor. The larger the value of d_i, the smaller $\lambda/\left(\lambda + d_i\right)$, i.e., the stronger the shrinkage toward $\boldsymbol{\beta}_0$. Thus, d_i can be viewed as the effective "variance" component for the i-th regression coefficient (under the transformation). This observation motivates us to consider semiparametric shrinkage estimators of the following form

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\beta}_0} &= \boldsymbol{Z}^T\left(\left(\boldsymbol{I}_k - \mathrm{diag}\left(\boldsymbol{b}\right)\right)\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \mathrm{diag}\left(\boldsymbol{b}\right)\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0\right)\\ &= \boldsymbol{Z}^T\left(\left(\boldsymbol{I}_k - \mathrm{diag}\left(\boldsymbol{b}\right)\right)\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{Y} + \mathrm{diag}\left(\boldsymbol{b}\right)\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0\right), \qquad (16)\end{aligned}$$

where $\boldsymbol{b}$ satisfies the following monotonic constraint

$$\mathrm{MON}\left(\boldsymbol{D}\right): b_i \in [0,1],\ b_i \le b_j \text{ whenever } d_i \le d_j.$$

This constraint captures the intuition that, the larger the effective variance, the stronger is the shrinkage.

For *fixed* $\boldsymbol{b}$ and $\boldsymbol{\beta}_0$, an unbiased estimate of the risk $R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\beta}_0})$ is

$$\begin{aligned}\mathrm{URE}^{SP}\left(\boldsymbol{b},\boldsymbol{\beta}_0\right) = &\frac{1}{p}\left\|\boldsymbol{Z}^T\left(\boldsymbol{I}_k - \mathrm{diag}\left(\boldsymbol{b}\right)\right)\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{Y} + \boldsymbol{Z}^T\mathrm{diag}\left(\boldsymbol{b}\right)\boldsymbol{U}^T\boldsymbol{W}^{-1/2}\boldsymbol{\beta}_0 - \boldsymbol{Y}\right\|^2\\ &+ \frac{1}{p}\mathrm{tr}\left(2\boldsymbol{Z}^T\left(\boldsymbol{I}_k - \mathrm{diag}\left(\boldsymbol{b}\right)\right)\boldsymbol{\Lambda}\boldsymbol{Z} - \boldsymbol{A}\right),\end{aligned}$$

which can be shown using the bias-variance decomposition of the mean squared error. Minimizing it gives the URE estimate of $(\boldsymbol{b},\boldsymbol{\beta}_0)$:

$$\left(\hat{\boldsymbol{b}}^{\mathrm{URE}}_{SP}, \left(\hat{\boldsymbol{\beta}}_0\right)^{\mathrm{URE}}_{SP}\right) = \underset{\boldsymbol{b}\in\mathrm{MON}(\boldsymbol{D}),\ \boldsymbol{\beta}_0\in\mathbb{R}^k}{\mathrm{argmin}}\ \mathrm{URE}^{SP}\left(\boldsymbol{b},\boldsymbol{\beta}_0\right),$$

which upon plugging into (16) yields the semiparametric URE estimator $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}_{SP}$ under Model II.

Theorem 5 *Assume conditions (A), (D)–(G). Then under Model II we have*

$$\sup_{\boldsymbol{b}\in\mathrm{MON}(\boldsymbol{D}),\ \boldsymbol{X}^T\boldsymbol{\beta}_0\in\mathscr{L}}\left|\mathrm{URE}^{SP}\left(\boldsymbol{b},\boldsymbol{\beta}_0\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\beta}_0}\right)\right| \to 0 \text{ in } L^1 \text{ as } p\to\infty.$$

As a corollary, for any estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\beta}}_{0,p}}$ *obtained from (16) with* $\hat{\boldsymbol{b}}_p \in \mathrm{MON}(\boldsymbol{D})$ *and* $\boldsymbol{X}^T\hat{\boldsymbol{\beta}}_0 \in \mathscr{L}$*, we always have*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\beta}}_{0,p}}\right) + \epsilon\right) = 0 \quad \forall \epsilon > 0,$$

$$\limsup_{p\to\infty}\left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p,\hat{\boldsymbol{\beta}}_{0,p}}\right)\right) \leq 0.$$

The proof of the theorem is essentially identical to those of Theorem 1 and Corollary 1 for the case of Model II except that we replace each $d_i/(\lambda + d_i)$ by b_i.

6 Simulation Study

In this section, we conduct simulations to study the performance of the URE estimators. For the sake of space, we will focus on Model I. The four URE estimators are the parametric $\hat{\boldsymbol{\theta}}^{\mathrm{URE}}$ of Eq. (9), the parametric $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}$ of Eq. (12) that shrinks toward the OLS estimate $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$ (i.e., the matrix $\boldsymbol{M} = \boldsymbol{I}_p$), the semiparametric $\hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE}}$ of Eq. (15), and the semiparametric $\hat{\boldsymbol{\theta}}_{SP}^{\mathrm{URE,\,OLS}}$ that shrinks toward $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}$, which is formed similarly to $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}^{\mathrm{URE}}$ by replacing $A_i/(\lambda + A_i)$ with a sequence $\boldsymbol{b} \in \mathrm{MON}(\boldsymbol{A})$. The competitors here are the two empirical Bayes estimators $\hat{\boldsymbol{\theta}}^{\mathrm{EBMLE}}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{EBMOM}}$, and the positive part James-Stein estimator $\hat{\boldsymbol{\theta}}^{\mathrm{JS+}}$ as described in [2, 17]:

$$\hat{\theta}_i^{\mathrm{JS+}} = \hat{\mu}_i^{\mathrm{WLS}} + \left(1 - \frac{p-k-2}{\sum_{i=1}^{p}\left(Y_i - \hat{\mu}_i^{\mathrm{WLS}}\right)^2/A_i}\right)^{+}\left(Y_i - \hat{\mu}_i^{\mathrm{WLS}}\right).$$

As a reference, we also compare these shrinkage estimators with $\tilde{\boldsymbol{\theta}}^{\mathrm{OR}}$, the parametric oracle risk (OR) estimator, defined as plugging $\tilde{\lambda}^{\mathrm{OR}}\boldsymbol{I}_p$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{OR}}$ into Eq. (6), where

$$\left(\tilde{\lambda}^{\mathrm{OR}}, \tilde{\boldsymbol{\mu}}^{\mathrm{OR}}\right) = \underset{0\leq\lambda\leq\infty,\ \boldsymbol{\mu}\in\mathscr{L}_{\mathrm{row}}(\boldsymbol{X})}{\operatorname{argmin}} R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\mu}}\right)$$

and the expression of $R_p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\lambda,\boldsymbol{\mu}})$ is given in (7) with $\boldsymbol{B} = \lambda\boldsymbol{I}_p$. The oracle risk estimator $\tilde{\boldsymbol{\theta}}^{\mathrm{OR}}$ cannot be used without the knowledge of $\boldsymbol{\theta}$, but it does provide a sensible lower bound of the risk achievable by any shrinkage estimator with the given parametric form.

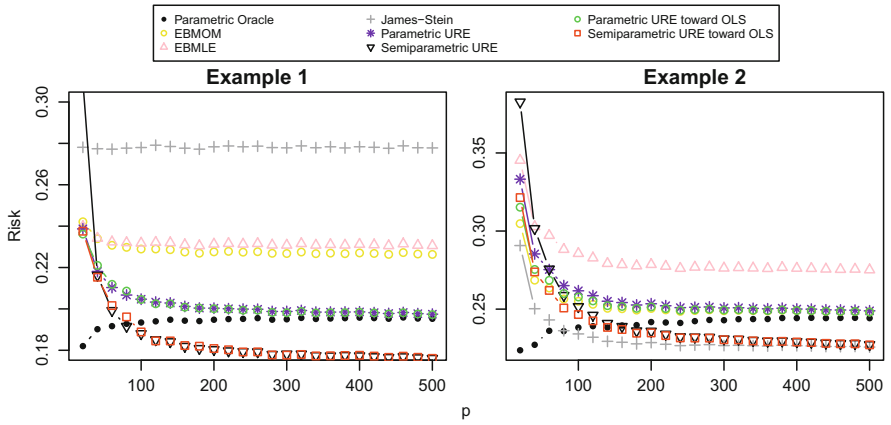

Fig. 2 Comparison of the risks of different shrinkage estimators for the two simulation examples

For each simulation, we draw (A_i, θ_i) $(i = 1, 2, \ldots, p)$ independently from a distribution $\pi(A_i, \theta_i | \boldsymbol{X}_i, \boldsymbol{\beta})$ and then draw Y_i given (A_i, θ_i). The shrinkage estimators are then applied to the generated data. This process is repeated 5000 times. The sample size p is chosen to vary from 20 to 500 with an increment of length 20. In the simulation, we fix a true but unknown $\boldsymbol{\beta} = (-1.5, 4, -3)^T$ and a known covariates $\boldsymbol{X}$, whose each element is randomly generated from Unif $(-10, 10)$. The risk performance of the different shrinkage estimators is given in Fig. 2.

Example 1 The setting in this example is chosen in such a way that it reflects grouping in the data:

$$A_i \sim 0.5 \cdot 1_{\{A_i=0.1\}} + 0.5 \cdot 1_{\{A_i=0.5\}};$$
$$\theta_i | A_i \sim N\left(2 \cdot 1_{\{A_i=0.1\}} + \boldsymbol{X}_i^T \boldsymbol{\beta}, 0.5^2\right);\ Y_i \sim N\left(\theta_i, A_i\right).$$

Here the normality for the sampling distribution of Y_i's is asserted. We can see that the four URE estimators perform much better than the two empirical Bayes ones and the James-Stein estimator. Also notice that both of the two (parametric and semiparametric) URE estimators that shrink towards $\hat{\boldsymbol{\mu}}^{\text{OLS}}$ is almost as good as the other two with general data-driven shrinkage location—largely due to the existence of covariate information. We note that this is quite different from the case of [29], where without the covariate information the estimator that shrinks toward the grand mean of the data performs significantly worse than the URE estimator with general data-driven shrinkage location.

Example 2 In this example, we allow Y_i to depart from the normal distribution to illustrate that the performance of those URE estimators does not rely on the

normality assumption:

$$A_i \sim \text{Unif}\,(0.1, 1)\,;\ \theta_i = A_i + \boldsymbol{X}_i^T \boldsymbol{\beta};$$
$$Y_i \sim \text{Unif}(\theta_i - \sqrt{3}A_i, \theta_i + \sqrt{3}A_i).$$

As expected, the four URE estimators perform better or at least as good as the empirical Bayes estimators. The EBMLE estimator performs the worst due to its sensitivity on the normality assumption. We notice that the EBMOM estimator in this example has comparable performance with the two parametric URE estimators, which makes sense as moment estimates are more robust to the sampling distribution. An interesting feature that we find in this example is that the positive part James-Stein estimator can beat the parametric oracle risk estimator and perform better than all the other shrinkage estimators for small or moderate p, even though the semiparametric URE estimators will eventually surpass the James-Stein estimator, as dictated by the asymptotic theory for large p. This feature of the James-Stein estimate is again quite different from the non-regression setting discussed in [29], where the James-Stein estimate performs the worst throughout all of their examples. In both of our examples only the semiparametric URE estimators are robust to the different levels of heteroscedasticity.

We can conclude from these two simulation examples that the semiparametric URE estimators give competitive performance and are robust to the misspecification of the sampling distribution and the different levels of the heteroscedasticity. They thus could be useful tools in analyzing large-scale data for applied researchers.

7 Empirical Analysis

In this section, we study the baseball data set of [2]. This data set consists of the batting records for all the Major League Baseball players in the 2005 season. As in [2] and [29], we build a given shrinkage estimator based on the data in the first half season and use it to predict the second half season, which can then be checked against the true record of the second half season. For each player, let the number of at-bats be N and the successful number of batting be H, then we have $H_{ij} \sim Binomial(N_{ij}, p_j)$, where $i = 1, 2$ is the season indicator and $j = 1, \cdots, p$ is the player indicator. We use the following variance-stabilizing transformation [2] before applying the shrinkage estimators

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

which gives $Y_{ij} \dot{\sim} N(\theta_j, (4N_{ij})^{-1})$, $\theta_j = \arcsin \sqrt{p_j}$. We use

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

as the error measurement for the prediction [2].

7.1 Shrinkage Estimation with Covariates

As indicated in [29], there exists a significant positive correlation between the player's batting ability and his total number of at-bats. Intuitively, a better player will be called for batting more frequently; thus, the total number of at-bats will serve as the main covariate in our analysis. The other covariate in the data set is the categorical variable of a player being a pitcher or not.

Table 1 summarizes the result, where the shrinkage estimators are applied three times—to all the players, the pitchers only, and the non-pitchers only. We use all the covariate information (number of at-bats in the first half season and being a pitcher or not) in the first analysis, whereas in the second and the third analyses we only use the number of at-bats as the covariate. The values reported are ratios of the error of a given estimator to that of the benchmark naive estimator, which simply uses the first half season Y_{1j} to predict the second half Y_{2j}. Note that in Table 1, if no covariate is involved (i.e., when $\boldsymbol{X} = [1|\cdots|1]$), the OLS reduces to the grand mean of the training data as in [29].

Table 1 Prediction errors of batting averages using different shrinkage estimators

	All		Pitchers		Non-pitchers	
p for estimation	567		81		486	
p for validation	499		64		435	
Covariates?	No	Yes	No	Yes	No	Yes
Naive	1	NA	1	NA	1	NA
Ordinary least squares (OLS)	0.852	0.242	0.127	0.115	0.378	0.333
Weighted least squares (WLS)	1.074	0.219	0.127	0.087	0.468	0.290
Parametric EBMOM	0.593	0.194	0.129	0.117	0.387	**0.256**
Parametric EBMLE	0.902	0.207	0.117	0.096	0.398	0.277
James-Stein	0.525	**0.184**	0.164	0.142	0.359	0.262
Parametric URE toward OLS	0.505	0.203	0.123	0.124	0.278	0.300
Parametric URE toward WLS	0.629	0.188	0.127	0.112	0.385	0.268
Parametric URE	0.422	0.215	0.123	0.130	0.282	0.310
Semiparametric URE toward OLS	0.409	0.197	0.081	0.097	0.261	0.299
Semiparametric URE toward WLS	0.499	**0.184**	0.098	**0.083**	0.336	**0.256**
Semiparametric URE	0.419	0.201	0.077	0.126	0.278	0.314

Bold numbers highlight the best performance with covariate(s) in each case

7.2 Discussion of the Numerical Result

There are several interesting observations from Table 1.

1. A quick glimpse shows that including the covariate information improves the performance of essentially all shrinkage estimators. This suggests that in practice incorporating good covariates would significantly improve the estimation and prediction.
2. In general, shrinking towards WLS provides much better performance than shrinking toward OLS or a general data-driven location. This indicates the importance of a good choice of the shrinkage location in a practical problem. An improperly chosen shrinkage location might even negatively impact the performance. The reason that shrinking towards a general data-driven location is not as good as shrinking toward WLS is probably due to that the sample size is not large enough for the asymptotics to take effect.
3. Table 1 also shows the advantage of semiparametric URE estimates. For each fixed shrinkage location type (toward OLS, WLS, or general), the semiparametric URE estimator performs almost always better than their parametric counterparts. The only one exception is in the non-pitchers only case with the general data-driven location, but even there the performance difference is ignorable.
4. The best performance in all three cases (all the players, the pitchers only, and the non-pitchers only) comes from the semiparametric URE estimator that shrinks toward WLS.
5. The James-Stein estimator *with* covariates performs quite well except in the pitchers only case, which is in sharp contrast with the performance of the James-Stein estimator *without* covariates. This again highlights the importance of covariate information. In the pitchers only case, the James-Stein performs the worst no matter one includes the covariates or not. This can be attributed to the fact that the covariate information (the total number of at-bats) is very weak for the pitchers only case; in the case of weak covariate information, how to properly estimate the shrinkage factors becomes the dominating issue, and the fact that the James-Stein estimator has only *one* uniform shrinkage factor makes it not competitive.

7.3 Shrinkage Factors

Figure 3 shows the shrinkage factors of all the shrinkage estimators with or without the covariates for the all-players case of Table 1. We see that the shrinkage factors are all reduced after including the covariates. This makes intuitive sense because the shrinkage location now contains the covariate information, and each shrinkage estimator uses this information by shrinking more toward it, resulting in smaller shrinkage factors.

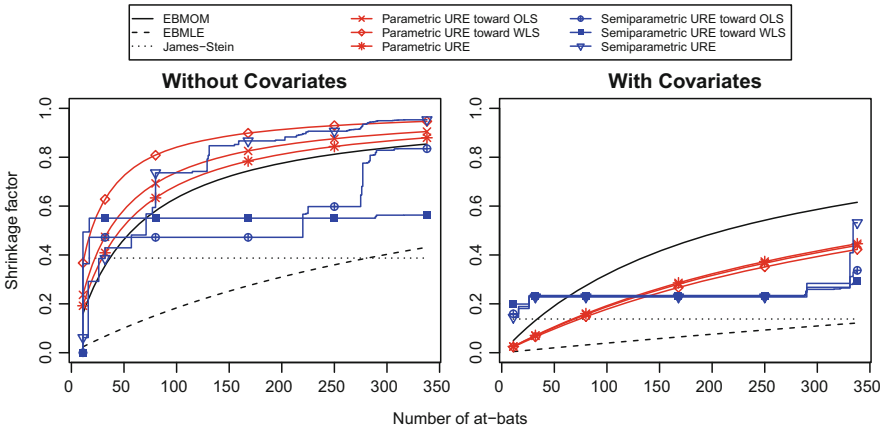

Fig. 3 Plot of the shrinkage factors $\hat{\lambda}/\left(\hat{\lambda}+A_i\right)$ or $1-\hat{b}_i$ of all the shrinkage estimators for the case of all players

8 Conclusion and Discussion

Inspired by the idea of unbiased risk estimate (URE) proposed in [29], we extend the URE framework to multivariate heteroscedastic linear models, which are more realistic in practical applications, especially for regression data that exhibits heteroscedasticity. Several parallel URE shrinkage estimators in the regression case are proposed, and these URE shrinkage estimators are all asymptotically optimal in risk compared to other shrinkage estimators, including the classical empirical Bayes ones. We also propose semiparametric estimators and conduct simulation to assess their performance under both normal and non-normal data. For data sets that exhibit a good linear relationship between the covariates and the response, a semiparametric URE estimator is expected to provide good estimation result, as we saw in the baseball data. It is also worth emphasizing that the risk optimality for the parametric and semiparametric URE estimators does not depend on the normality assumption of the sampling distribution of Y_i. Possible future work includes extending this URE minimization approach to simultaneous estimation in generalized linear models (GLMs) with canonical or more general link functions.

We conclude this article by extending the main results to the case of weighted mean squared error loss.

Weighted Mean Squared Error Loss One might want to consider the more general *weighted mean squared error* as the loss function:

$$l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}; \boldsymbol{\psi}\right) = \frac{1}{p}\sum_{i=1}^{p} \psi_i \left(\theta_i - \hat{\theta}_i\right)^2,$$

where $\psi_i > 0$ are known weights such that $\sum_{i=1}^{p} \psi_i = p$. The framework proposed in this article is straightforward to generalize to this case.

For Model II, we only need to study the equivalent problem by the following transformation

$$Y_i \to \sqrt{\psi_i} Y_i,\ \theta_i \to \sqrt{\psi_i}\theta_i,\ \boldsymbol{X}_i \to \sqrt{\psi_i}\boldsymbol{X}_i,\ A_i \to \psi_i A_i, \tag{17}$$

and restate the corresponding regularity conditions in Theorem 1 by the transformed data and parameters. We then reduce the weighted mean square error problem back to the same setting we study in this article under the classical loss function (4).

Model I is more sophisticated than Model II to generalize. In addition to the transformation in Eq. (17), we also need $\lambda \to \psi_i \lambda$ in every term related to the individual unit i. Thus,

$$\sqrt{\psi_i}\theta_i | \boldsymbol{X}, \boldsymbol{\beta}, \lambda \overset{\text{indep.}}{\sim} N\left(\sqrt{\psi_i}\boldsymbol{X}_i^T \boldsymbol{\beta}, \lambda\psi_i\right),$$

so these transformed parameters $\sqrt{\psi_i}\theta_i$ are also heteroscedastic in the sense that they have different weights, while the setting we study before assumes all the weights on the θ_i are one. However, if we carefully examine the proof of Theorem 1 for the case of Model I, we can see that actually we do not much require the equal weights on the θ_i's. What is important in the proof is that the shrinkage factor for unit i is always of the form $A_i/(A_i + \lambda)$, which is *invariant* under the transformation $A_i \to \psi_i A_i$ and $\lambda \to \psi_i \lambda$. Thus, after reformulating the regularity conditions in Theorem 1 by the transformed data and parameters, we can still follow the same proof to conclude the risk optimality of URE estimators (parametric or semiparametric) even under the consideration of weighted mean squared error loss.

For completeness, here we state the most general result under the semiparametric setting for Model I. Let

$$\hat{\boldsymbol{\theta}}^{\text{URE}}_{SP,\boldsymbol{\psi}} = \left(\boldsymbol{I}_p - \text{diag}\left(\hat{\boldsymbol{b}}^{\text{URE}}_{\boldsymbol{\psi}}\right)\right)\boldsymbol{Y} + \text{diag}\left(\hat{\boldsymbol{b}}^{\text{URE}}_{\boldsymbol{\psi}}\right)\hat{\boldsymbol{\mu}}^{\text{URE}}_{\boldsymbol{\psi}},$$

$$\text{URE}\left(\boldsymbol{b}, \boldsymbol{\mu}; \boldsymbol{\psi}\right) = \frac{1}{p}\sum_{i=1}^{p} \psi_i \left(b_i^2 \left(Y_i - \mu_i\right)^2 + \left(1 - 2b_i\right) A_i\right),$$

$$\left(\hat{\boldsymbol{b}}^{\text{URE}}_{\boldsymbol{\psi}}, \hat{\boldsymbol{\mu}}^{\text{URE}}_{\boldsymbol{\psi}}\right) = \underset{\boldsymbol{b}\in \text{MON}(\boldsymbol{A}),\ \boldsymbol{\mu}\in\mathscr{L}_{\text{row}}(\boldsymbol{X})}{\text{argmin}} \text{URE}\left(\boldsymbol{b}, \boldsymbol{\mu}; \boldsymbol{\psi}\right).$$

Theorem 6 *Assume the following five conditions* (ψ-A) $\sum_{i=1}^{p} \psi_i^2 A_i^2 = O(p)$, ($\psi$-B) $\sum_{i=1}^{p} \psi_i^2 A_i \theta_i^2 = O(p)$, ($\psi$-C) $\sum_{i=1}^{p} \psi_i \theta_i^2 = O(p)$, ($\psi$-D) $p^{-1}\sum_{i=1}^{p} \psi_i^2 A_i \boldsymbol{X}_i \boldsymbol{X}_i^T$ *converges, and* (ψ-E) $p^{-1}\sum_{i=1}^{p} \psi_i \boldsymbol{X}_i \boldsymbol{X}_i^T \to \boldsymbol{\Omega}_{\boldsymbol{\psi}} > 0$. *Then we have*

$$\sup_{\boldsymbol{b}\in\text{MON}(\boldsymbol{A}),\ \boldsymbol{\mu}\in\mathscr{L}_{\boldsymbol{\psi}}} \left|\text{URE}\left(\boldsymbol{b}, \boldsymbol{\mu}; \boldsymbol{\psi}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{b},\boldsymbol{\mu}}; \boldsymbol{\psi}\right)\right| \underset{p\to\infty}{\to} 0 \text{ in } L^1,$$

where $\boldsymbol{\mu} \in \mathscr{L}_{\boldsymbol{\psi}}$ *if and only if* $\boldsymbol{\mu} \in \mathscr{L}_{\text{row}}(\boldsymbol{X})$ *and*

$$\sum_{i=1}^{p} \psi_i \mu_i^2 \leq M p^{\kappa} \sum_{i=1}^{p} \psi_i Y_i^2$$

for a large and fixed constant M *and a fixed exponent* $\kappa \in [0, 1/2)$. *As a corollary, for any estimator* $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p, \hat{\boldsymbol{\mu}}_p} = (\boldsymbol{I}_p - \text{diag}(\hat{\boldsymbol{b}}_p))\boldsymbol{Y} + \text{diag}(\hat{\boldsymbol{b}}_p)\hat{\boldsymbol{\mu}}_p$ *with* $\hat{\boldsymbol{b}}_p \in \text{MON}(\boldsymbol{A})$ *and* $\hat{\boldsymbol{\mu}}_p \in \mathscr{L}_{\boldsymbol{\psi}}$, *we have*

$$\lim_{p\to\infty} \mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\text{URE}}_{SP,\boldsymbol{\psi}}\right) \geq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p, \hat{\boldsymbol{\mu}}_p}\right) + \epsilon \right) = 0 \quad \forall \epsilon > 0,$$

$$\limsup_{p\to\infty} \left(R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\text{URE}}_{SP,\boldsymbol{\psi}}\right) - R_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{b}}_p, \hat{\boldsymbol{\mu}}_p}\right) \right) \leq 0.$$

Acknowledgements S. C. Kou's research is supported in part by US National Science Foundation Grant DMS-1510446.

Appendix: Proofs and Derivations

Proof of Lemma 1 We can write $\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{Z}_1$ and $\boldsymbol{Y} = \boldsymbol{\theta} + \boldsymbol{Z}_2$, where $\boldsymbol{Z}_1 \sim \mathscr{N}_p(\boldsymbol{0}, \boldsymbol{B})$ and $\boldsymbol{Z}_2 \sim \mathscr{N}_p(\boldsymbol{0}, \boldsymbol{A})$ are independent. Jointly $\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{\theta} \end{pmatrix}$ is still multivariate normal with mean vector $\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}$ and covariance matrix $\begin{pmatrix} \boldsymbol{A}+\boldsymbol{B} & \boldsymbol{B} \\ \boldsymbol{B} & \boldsymbol{B} \end{pmatrix}$. The result follows immediately from the conditional distribution of a multivariate normal distribution.

Proof of Theorem 1 We start from decomposing the difference between the URE and the actual loss as

$$\begin{aligned}
&\text{URE}(\boldsymbol{B}, \boldsymbol{\mu}) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right) \\
=&\text{URE}\left(\boldsymbol{B}, \boldsymbol{0}_p\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{0}_p}\right) - \frac{2}{p}\text{tr}\left(\boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{\mu}(\boldsymbol{Y}-\boldsymbol{\theta})^T\right) \qquad (18)\\
=&\frac{1}{p}\text{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{A} - \boldsymbol{\theta}\boldsymbol{\theta}^T\right) - \frac{2}{p}\text{tr}\left(\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}\boldsymbol{\theta}^T - \boldsymbol{A}\right)\right) \\
&- \frac{2}{p}\text{tr}\left(\boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{\mu}(\boldsymbol{Y}-\boldsymbol{\theta})^T\right) \qquad (19)\\
=&\,\text{(I)} + \text{(II)} + \text{(III)}.
\end{aligned}$$

To verify the first equality (18), note that

$$
\begin{aligned}
&\mathrm{URE}\left(\boldsymbol{B},\boldsymbol{\mu}\right)-\mathrm{URE}\left(\boldsymbol{B},\mathbf{0}_p\right)\\
&=\frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)\right\|^2-\frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{Y}\right\|^2\\
&=-\frac{1}{p}\mathrm{tr}\left(\boldsymbol{\mu}^T\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)^T\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(2\boldsymbol{Y}-\boldsymbol{\mu}\right)\right),\\
&\quad l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)-l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\mathbf{0}_p}\right)\\
&=\frac{1}{p}\left\|\left(\boldsymbol{I}_p-\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y}+\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{\mu}-\boldsymbol{\theta}\right\|^2\\
&\quad-\frac{1}{p}\left\|\left(\boldsymbol{I}_p-\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y}-\boldsymbol{\theta}\right\|^2\\
&=\frac{1}{p}\mathrm{tr}\left(\boldsymbol{\mu}^T\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)^T\left(2\left(\left(\boldsymbol{I}_p-\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y}-\boldsymbol{\theta}\right)+\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{\mu}\right)\right).
\end{aligned}
$$

Equation (18) then follows by rearranging the terms. To verify the second equality (19), note

$$
\begin{aligned}
&\mathrm{URE}\left(\boldsymbol{B},\mathbf{0}_p\right)-l_p\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{\boldsymbol{B},\mathbf{0}_p}\right)\\
&=\frac{1}{p}\left\|\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{Y}\right\|^2-\frac{1}{p}\left\|\left(\boldsymbol{I}_p-\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y}-\boldsymbol{\theta}\right\|^2\\
&\quad+\frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}-2\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{A}\right)\\
&=\frac{1}{p}\mathrm{tr}\left(\left(\boldsymbol{Y}-2\left(\boldsymbol{I}_p-\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\right)\boldsymbol{Y}+\boldsymbol{\theta}\right)^T\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right)\\
&\quad+\frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}-2\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{A}\right)\\
&=\frac{1}{p}\mathrm{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T-\boldsymbol{A}-\boldsymbol{\theta}\boldsymbol{\theta}^T\right)-\frac{2}{p}\mathrm{tr}\left(\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T-\boldsymbol{A}\right)\right).
\end{aligned}
$$

With the decomposition, we want to prove separately the uniform L^1 convergence of the three terms (I), (II), and (III).

Proof for the case of Model I.

The uniform L^2 convergence of (I) and (II) has been shown in Theorem 3.1 of [29] under our assumptions (A) and (B), so we focus on (III), i.e., we want to show that $\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathscr{L}}|(\mathrm{III})|\to 0$ in L^1 as $p\to\infty$.

Without loss of generality, let us assume $A_1 \le A_2 \le \cdots \le A_p$. We have

$$\begin{aligned}\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathscr{L}} |(\mathrm{III})| &= \frac{2}{p}\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathscr{L}}\left|\sum_{i=1}^{p}\frac{A_i}{A_i+\lambda}\mu_i\left(Y_i-\theta_i\right)\right| \\ &\le \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sup_{0\le c_1\le\cdots\le c_p\le 1}\left|\sum_{i=1}^{p}c_i\mu_i\left(Y_i-\theta_i\right)\right| \\ &= \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\max_{1\le j\le p}\left|\sum_{i=j}^{p}\mu_i\left(Y_i-\theta_i\right)\right|,\end{aligned}$$

where the last equality follows from Lemma 2.1 of [13]. For a generic p-dimensional vector $\boldsymbol{v}$, we denote $[\boldsymbol{v}]_{j:p} = (0,\ldots 0, v_j, v_{j+1},\ldots,v_p)$. Let $\boldsymbol{P_X} = \boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}$ be the projection matrix onto $\mathscr{L}_{\mathrm{row}}(\boldsymbol{X})$. Then since $\mathscr{L}\subset\mathscr{L}_{\mathrm{row}}(\boldsymbol{X})$, we have

$$\begin{aligned}&\frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\max_{1\le j\le p}\left|\sum_{i=j}^{p}\mu_i\left(Y_i-\theta_i\right)\right| = \frac{2}{p}\max_{1\le j\le p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\left|\boldsymbol{\mu}^T[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right| \\ =&\frac{2}{p}\max_{1\le j\le p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\left|\boldsymbol{\mu}^T\boldsymbol{P_X}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right| \le \frac{2}{p}\max_{1\le j\le p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\|\boldsymbol{\mu}\|\times\left\|\boldsymbol{P_X}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right\| \\ =&\frac{2}{p}\max_{1\le j\le p}Mp^{\kappa}\,\|\boldsymbol{Y}\|\times\left\|\boldsymbol{P_X}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right\|.\end{aligned}$$

Cauchy-Schwarz inequality thus gives

$$\mathbb{E}\left(\sup_{0\le\lambda\le\infty,\boldsymbol{\mu}\in\mathscr{L}}|(\mathrm{III})|\right) \le 2Mp^{\kappa-1}\sqrt{\mathbb{E}\left(\|\boldsymbol{Y}\|^2\right)}\times\sqrt{\mathbb{E}\left(\max_{1\le j\le p}\left\|\boldsymbol{P_X}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right\|^2\right)}. \tag{20}$$

It is straightforward to see that, by conditions (A) and (C),

$$\sqrt{\mathbb{E}\left(\|\boldsymbol{Y}\|^2\right)} = \sqrt{\mathbb{E}(\textstyle\sum_{i=1}^{p}Y_i^2)} = \sqrt{\textstyle\sum_{i=1}^{p}\left(\theta_i^2+A_i\right)} = O\left(p^{1/2}\right).$$

For the second term on the right-hand side of (20), let $\boldsymbol{P_X} = \boldsymbol{\Gamma}\boldsymbol{D}\boldsymbol{\Gamma}^T$ denote the spectral decomposition. Clearly,

$$\boldsymbol{D} = \mathrm{diag}\left(\underbrace{1,\ldots,1}_{k\text{ copies}},\underbrace{0,\ldots,0}_{p-k\text{ copies}}\right).$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left(\max_{1\le j\le p}\left\|\boldsymbol{P}_{\boldsymbol{X}}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right\|^2\right)=\mathbb{E}\left(\max_{1\le j\le p}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}^T\boldsymbol{P}_{\boldsymbol{X}}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right)\\
&=\mathbb{E}\left(\max_{1\le j\le p}\operatorname{tr}\left(\boldsymbol{D}\boldsymbol{\Gamma}^T[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\left(\boldsymbol{\Gamma}^T[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right)^T\right)\right)\\
&=\mathbb{E}\left(\max_{1\le j\le p}\sum_{l=1}^{k}\left[\boldsymbol{\Gamma}^T[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right]_l^2\right)\\
&=\mathbb{E}\left(\max_{1\le j\le p}\sum_{l=1}^{k}\left(\sum_{m=j}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)\right)^2\right)\\
&\le\mathbb{E}\left(\sum_{l=1}^{k}\max_{1\le j\le p}\left(\sum_{m=j}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)\right)^2\right)\\
&=\sum_{l=1}^{k}\mathbb{E}\left(\max_{1\le j\le p}\left(\sum_{m=j}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)\right)^2\right).
\end{aligned}$$

For each l, $M_j^{(l)}=\sum_{m=p-j+1}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)$ forms a martingale, so by Doob's L^p maximum inequality,

$$\begin{aligned}
\mathbb{E}\left(\max_{1\le j\le p}\left(M_j^{(l)}\right)^2\right)&\le 4\mathbb{E}\left(M_p^{(l)}\right)^2=4\mathbb{E}\left(\sum_{m=1}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}(Y_m-\theta_m)\right)^2\\
&=4\sum_{m=1}^{p}\left[\boldsymbol{\Gamma}^T\right]_{lm}^2A_m=4\left[\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\right]_{ll}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left(\max_{1\le j\le p}\left\|\boldsymbol{P}_{\boldsymbol{X}}[\boldsymbol{Y}-\boldsymbol{\theta}]_{j:p}\right\|^2\right)\le\sum_{l=1}^{k}4\left[\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\right]_{ll}\\
&=4\sum_{l=1}^{p}[\boldsymbol{D}]_{ll}\left[\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\right]_{ll}=4\operatorname{tr}\left(\boldsymbol{D}\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\right)=4\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{A}\right)\\
&=4\operatorname{tr}\left(\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}\right)=4\operatorname{tr}\left(\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right)=O(1),
\end{aligned}$$

where the last equality uses conditions (D) and (E). We finally obtain

$$\mathbb{E}\left(\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathscr{L}}|(\mathrm{III})|\right)\le o\left(p^{-1/2}\right)\times O\left(p^{1/2}\right)\times O\left(1\right)=o\left(1\right).$$

Proof for the case of Model II.
Under Model II, we know that

$$\sum_{i=1}^{p}A_i\theta_i^2=\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta}=\boldsymbol{\beta}^T(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T)\boldsymbol{\beta}=O\left(p\right)$$

by condition (D). In other words, condition (D) implies condition (B). Therefore, we know that the term (I) $\to 0$ in L^2 as shown in Theorem 3.1 of [29], and we only need to show the uniform L^1 convergence of the other two terms, (II) and (III).

Recall that $\boldsymbol{B}\in\mathscr{B}=\left\{\lambda\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}:\lambda>0\right\}$ has only rank k under Model II. We can reexpress (II) and (III) in terms of low rank matrices. Let $\boldsymbol{V}=\left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)^{-1}$. Woodbury formula gives

$$\begin{aligned}(\boldsymbol{A}+\boldsymbol{B})^{-1}&=\left(\boldsymbol{A}+\lambda\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\lambda\boldsymbol{X}^T\left(\boldsymbol{W}^{-1}+\lambda\boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\\&=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\lambda\boldsymbol{X}^T\boldsymbol{W}\left(\lambda\boldsymbol{W}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{X}\boldsymbol{A}^{-1},\end{aligned}$$

which tells us

$$\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}=\boldsymbol{I}_p-\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}=\lambda\boldsymbol{X}^T\boldsymbol{W}\left(\lambda\boldsymbol{W}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{X}\boldsymbol{A}^{-1}.$$

Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ be the spectral decomposition of $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}$, i.e., $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda}=\mathrm{diag}\left(d_1,\ldots,d_k\right)$ with $d_1\le\cdots\le d_k$. Then $\left(\lambda\boldsymbol{W}+\boldsymbol{V}\right)^{-1}=\boldsymbol{W}^{-1/2}\left(\lambda\boldsymbol{I}_k+\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}\right)^{-1}\boldsymbol{W}^{-1/2}=\boldsymbol{W}^{-1/2}\boldsymbol{U}\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^T\boldsymbol{W}^{-1/2}$, from which we obtain

$$\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}=\lambda\boldsymbol{X}^T\boldsymbol{W}\left(\lambda\boldsymbol{W}+\boldsymbol{V}\right)^{-1}\boldsymbol{V}\boldsymbol{X}\boldsymbol{A}^{-1}=\lambda\boldsymbol{X}^T\boldsymbol{W}^{1/2}\boldsymbol{U}\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{A}^{-1}.$$

If we denote $\boldsymbol{Z}=\boldsymbol{U}^T\boldsymbol{W}^{1/2}\boldsymbol{X}$, i.e., $\boldsymbol{Z}$ is the transformed covariate matrix, then $\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}=\lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}$. It follows that

$$\begin{aligned}(\mathrm{II})&=-\frac{2}{p}\mathrm{tr}\left(\boldsymbol{B}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^T-\boldsymbol{Y}\boldsymbol{\theta}^T-\boldsymbol{A}\right)\right)\\&=-\frac{2}{p}\mathrm{tr}\left(\lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^T-\boldsymbol{Y}\boldsymbol{\theta}^T-\boldsymbol{A}\right)\right)\\&=-\frac{2}{p}\mathrm{tr}\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^T-\boldsymbol{Y}\boldsymbol{\theta}^T-\boldsymbol{A}\right)\boldsymbol{Z}^T\right),\end{aligned}$$

$$\begin{aligned}
\text{(III)} &= -\frac{2}{p}\text{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^{T}\right) \\
&= -\frac{2}{p}\text{tr}\left(\left(\boldsymbol{I}_p-\lambda\boldsymbol{Z}^{T}\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\right)\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^{T}\right) \\
&= -\frac{2}{p}\text{tr}\left(\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^{T}\right)+\frac{2}{p}\text{tr}\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^{T}\boldsymbol{Z}^{T}\right) \\
&= \text{(III)}_1+\text{(III)}_2\,.
\end{aligned}$$

We will next show that (II), $(\text{III})_1$, and $(\text{III})_2$ all uniformly converge to zero in L^1, which will then complete our proof.

Let $\boldsymbol{\Xi}=\boldsymbol{Z}\boldsymbol{A}^{-1}\left(\boldsymbol{Y}\boldsymbol{Y}^{T}-\boldsymbol{Y}\boldsymbol{\theta}^{T}-\boldsymbol{A}\right)\boldsymbol{Z}^{T}$. Then

$$\begin{aligned}
\sup_{0\le\lambda\le\infty}|\text{(II)}| &= \frac{2}{p}\sup_{0\le\lambda\le\infty}\left|\sum_{i=1}^{k}\frac{\lambda d_i}{\lambda+d_i}\left[\boldsymbol{\Xi}\right]_{ii}\right| \\
&\le \frac{2}{p}\sup_{0\le c_1\le\cdots\le c_k\le d_k}\left|\sum_{i=1}^{k}c_i\left[\boldsymbol{\Xi}\right]_{ii}\right| = \frac{2}{p}\max_{1\le j\le k}\left|\sum_{i=j}^{k}d_k\left[\boldsymbol{\Xi}\right]_{ii}\right|,
\end{aligned}$$

where the last equality follows as in Lemma 2.1 of [13]. As there are finite number of terms in the summation and the maximization, it suffices to show that

$$d_k\left[\boldsymbol{\Xi}\right]_{ii}/p\to 0 \text{ in } L^2 \quad \text{for all } 1\le i\le k.$$

To establish this, we note that $\left[\boldsymbol{\Xi}\right]_{ii}=\sum_{n=1}^{p}\sum_{m=1}^{p}\left(A_n^{-1}Y_n\left(Y_m-\theta_m\right)-\delta_{nm}\right)\left[\boldsymbol{Z}\right]_{in}\left[\boldsymbol{Z}\right]_{im}$,

$$\begin{aligned}
\mathbb{E}\left(\left[\boldsymbol{\Xi}\right]_{ii}^2\right) = \sum_{n,m,n',m'}&\mathbb{E}\left(\left(A_n^{-1}Y_n\left(Y_m-\theta_m\right)-\delta_{nm}\right)\left(A_{n'}^{-1}Y_{n'}\left(Y_{m'}-\theta_{m'}\right)-\delta_{n'm'}\right)\right) \\
&\times\left[\boldsymbol{Z}\right]_{in}\left[\boldsymbol{Z}\right]_{im}\left[\boldsymbol{Z}\right]_{in'}\left[\boldsymbol{Z}\right]_{im'}.
\end{aligned}$$

Depending on n,m,n',m' taking the same or distinct values, we can break the summation into 15 disjoint cases:

$$\sum_{\text{all distinct}}+\sum_{\text{three distinct, }n=m}+\sum_{\text{three distinct, }n=n'}+\sum_{\text{three distinct, }n=m'}$$
$$+\sum_{\text{three distinct, }m=n'}+\sum_{\text{three distinct, }m=m'}+\sum_{\text{three distinct, }n'=m'}+\sum_{\text{two distinct, }n=m,\,n'=m'}$$

$$+\sum_{\text{two distinct, } n=n',\, m=m'} + \sum_{\text{two distinct, } n=m',\, n'=m} + \sum_{\text{two distinct, } n=m=n'} + \sum_{\text{two distinct, } n=m=m'}$$
$$+\sum_{\text{two distinct, } n=n'=m'} + \sum_{\text{two distinct, } m=n'=m'} + \sum_{n=m=n'=m'}.$$

Many terms are zero. Straightforward evaluation of each summation gives

$$\begin{aligned}
\mathbb{E}\left([\boldsymbol{\Xi}]_{ii}^2\right) &= \sum_{n=1}^{p} \mathbb{E}\left(\left(A_n^{-1} Y_n \left(Y_n - \theta_n\right) - 1\right)^2\right) [\boldsymbol{Z}]_{in}^4 \\
&\quad + \sum_{n=1}^{p} \sum_{m\neq n} \mathbb{E}\left(\left(A_n^{-1} Y_n \left(Y_m - \theta_m\right)\right)^2\right) [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 \\
&\quad + \sum_{n=1}^{p} \sum_{m\neq n} \mathbb{E}\left(\left(A_n^{-1} Y_n \left(Y_m - \theta_m\right)\right)\left(A_m^{-1} Y_m \left(Y_n - \theta_n\right)\right)\right) [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 \\
&\quad + 2\sum_{n=1}^{p} \sum_{m\neq n} \mathbb{E}\left(\left(A_n^{-1} Y_n \left(Y_n - \theta_n\right) - 1\right)\left(A_m^{-1} Y_m \left(Y_n - \theta_n\right)\right)\right) [\boldsymbol{Z}]_{in}^3 [\boldsymbol{Z}]_{im} \\
&\quad + \sum_{n=1}^{p} \sum_{m\neq n',\, n'\neq n,\, m\neq n} \mathbb{E}\left(\left(A_m^{-1} Y_m \left(Y_n - \theta_n\right)\right)\left(A_{n'}^{-1} Y_{n'} \left(Y_n - \theta_n\right)\right)\right) [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im} [\boldsymbol{Z}]_{in'} \\
&= \sum_{n=1}^{p} \frac{2A_n + \theta_n^2}{A_n} [\boldsymbol{Z}]_{in}^4 + \sum_{n=1}^{p} \sum_{m\neq n} \frac{A_n A_m + A_n \theta_m^2}{A_m^2} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 + \sum_{n=1}^{p} \sum_{m\neq n} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 \\
&\quad + 2\sum_{n=1}^{p} \sum_{m\neq n} \frac{\theta_n \theta_m}{A_m} [\boldsymbol{Z}]_{in}^3 [\boldsymbol{Z}]_{im} + \sum_{n=1}^{p} \sum_{m\neq n',\, n'\neq n,\, m\neq n} \frac{A_n \theta_m \theta_{n'}}{A_m A_{n'}} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im} [\boldsymbol{Z}]_{in'} \\
&= \sum_{n,m=1}^{p} \frac{A_n}{A_m} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 + \sum_{n,m=1}^{p} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im}^2 + \sum_{n,m,n'=1}^{p} \frac{A_n \theta_m \theta_{n'}}{A_m A_{n'}} [\boldsymbol{Z}]_{in}^2 [\boldsymbol{Z}]_{im} [\boldsymbol{Z}]_{in'}.
\end{aligned}$$

Using matrix notation, we can reexpress the above equation as

$$\begin{aligned}
\mathbb{E}\left([\boldsymbol{\Xi}]_{ii}^2\right) &= \left[\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\right]_{ii} \left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\right]_{ii} + \left[\boldsymbol{Z}\boldsymbol{Z}^T\right]_{ii}^2 + \left[\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\right]_{ii} \left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\theta}\right]_i^2 \\
&\leq \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\right) \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\right) + \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{Z}^T\right)^2 + \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\right) \operatorname{tr}\left(\boldsymbol{\theta}^T \boldsymbol{A}^{-1} \boldsymbol{Z}^T \boldsymbol{Z} \boldsymbol{A}^{-1} \boldsymbol{\theta}\right) \\
&= \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right) \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right) + \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^T\right)^2 \\
&\quad + \operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right) \operatorname{tr}\left(\boldsymbol{\beta}^T \left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right) \boldsymbol{W} \left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right) \boldsymbol{\beta}\right),
\end{aligned}$$

which is $O(p)\,O(p) + O(p)^2 + O(p)\,O\left(p^2\right) = O\left(p^3\right)$ by conditions (D)-(F). Note also that condition (F) implies

$$d_k \leq \sum_{i=1}^{k} d_i = \operatorname{tr}\left(\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}\right) = \operatorname{tr}\left(\boldsymbol{W}^{-1}\boldsymbol{V}\right) = \operatorname{tr}\left(\boldsymbol{W}^{-1}(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T)^{-1}\right) = O\left(p^{-1}\right).$$

Therefore, we have

$$\mathbb{E}\left(d_k^2\left[\boldsymbol{\Xi}\right]_{ii}^2/p^2\right) = O\left(p^{-2}\right)O\left(p^3\right)/p^2 = O\left(p^{-1}\right) \to 0,$$

which proves

$$\sup_{0\leq\lambda\leq\infty} |(\mathrm{II})| \to 0 \text{ in } L^2, \quad \text{as } p \to \infty.$$

To prove the uniform convergence of $(\mathrm{III})_1$ to zero in L^1, we note that

$$\begin{aligned}\sup_{\boldsymbol{\mu}\in\mathscr{L}} |(\mathrm{III})_1| &= \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}} \left|\boldsymbol{\mu}^T(\boldsymbol{Y}-\boldsymbol{\theta})\right| = \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}} \left|\boldsymbol{\mu}^T\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y}-\boldsymbol{\theta})\right| \\ &\leq \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}} \|\boldsymbol{\mu}\| \times \|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y}-\boldsymbol{\theta})\| = \frac{2}{p}Mp^{\kappa}\,\|\boldsymbol{Y}\| \times \|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y}-\boldsymbol{\theta})\|,\end{aligned}$$

so by Cauchy-Schwarz inequality

$$\mathbb{E}\left(\sup_{\boldsymbol{\mu}\in\mathscr{L}} |(\mathrm{III})_1|\right) \leq 2Mp^{\kappa-1}\sqrt{\mathbb{E}\left(\|\boldsymbol{Y}\|^2\right)}\sqrt{\mathbb{E}\left(\|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y}-\boldsymbol{\theta})\|^2\right)}. \tag{21}$$

Under Model II, $\boldsymbol{\theta} = \boldsymbol{X}^T\boldsymbol{\beta}$, so it follows that $\sum_{i=1}^{p}\theta_i^2 = \|\boldsymbol{\theta}\|^2 = \operatorname{tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}\boldsymbol{X}^T\right) = O(p)$ by condition (E). Hence $\sqrt{\mathbb{E}\left(\|\boldsymbol{Y}\|^2\right)} = \sqrt{\sum_{i=1}^{p}\left(\theta_i^2 + A_i\right)} = O\left(p^{1/2}\right)$. For the second term on the right-hand side of (21), note that

$$\begin{aligned}\mathbb{E}\left(\|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y}-\boldsymbol{\theta})\|^2\right) &= \mathbb{E}\left(\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y}-\boldsymbol{\theta})(\boldsymbol{Y}-\boldsymbol{\theta})^T\right)\right) \\ &= \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{A}) = \operatorname{tr}\left(\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right) = O(1)\end{aligned}$$

by conditions (D) and (E). Thus, in aggregate, we have

$$\mathbb{E}\left(\sup_{\boldsymbol{\mu}\in\mathscr{L}} |(\mathrm{III})_1|\right) \leq 2Mp^{\kappa-1}O\left(p^{1/2}\right)O(1) = o(1).$$

We finally consider the $(\mathrm{III})_2$ term. We have

$$\begin{aligned}
\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathscr{L}} |(\mathrm{III})_2| &= \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sup_{0\le\lambda\le\infty}\left|\sum_{i=1}^{k}\frac{\lambda d_i}{\lambda+d_i}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\,(\boldsymbol{Y}-\boldsymbol{\theta})^T\boldsymbol{Z}^T\right]_{ii}\right| \\
&\le \frac{2}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\max_{1\le j\le k}\left|\sum_{i=j}^{k} d_k\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\,(\boldsymbol{Y}-\boldsymbol{\theta})^T\boldsymbol{Z}^T\right]_{ii}\right| \\
&\le \frac{2d_k}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sum_{i=1}^{k}\left|\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\,(\boldsymbol{Y}-\boldsymbol{\theta})^T\boldsymbol{Z}^T\right]_{ii}\right| \\
&= \frac{2d_k}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sum_{i=1}^{k}\left|\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i\left[\boldsymbol{Z}\,(\boldsymbol{Y}-\boldsymbol{\theta})\right]_i\right| \\
&\le \frac{2d_k}{p}\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sqrt{\sum_{i=1}^{k}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i^2}\times\sqrt{\sum_{i=1}^{k}\left[\boldsymbol{Z}\,(\boldsymbol{Y}-\boldsymbol{\theta})\right]_i^2}.
\end{aligned}$$

Thus, by Cauchy-Schwarz inequality

$$\mathbb{E}\left(\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathscr{L}}|(\mathrm{III})_2|\right)\le\frac{2d_k}{p}\sqrt{\mathbb{E}\left(\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sum_{i=1}^{k}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i^2\right)}\times\sqrt{\mathbb{E}\left(\sum_{i=1}^{k}\left[\boldsymbol{Z}\,(\boldsymbol{Y}-\boldsymbol{\theta})\right]_i^2\right)}.$$

Note that

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sum_{i=1}^{k}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i^2=\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sum_{i=1}^{k}\left(\sum_{m=1}^{p}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\right]_{im}[\boldsymbol{\mu}]_m\right)^2 \\
&\le\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sum_{i=1}^{k}\left(\sum_{m=1}^{p}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\right]_{im}^2\times\sum_{m=1}^{p}[\boldsymbol{\mu}]_m^2\right)=\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sum_{i=1}^{k}\left(\left[\boldsymbol{Z}\boldsymbol{A}^{-2}\boldsymbol{Z}^T\right]_{ii}\|\boldsymbol{\mu}\|^2\right) \\
&=\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{A}^{-2}\boldsymbol{Z}^T\right)\sup_{\boldsymbol{\mu}\in\mathscr{L}}\|\boldsymbol{\mu}\|^2=\mathrm{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-2}\boldsymbol{X}^T\right)\left(Mp^{\kappa}\|\boldsymbol{Y}\|\right)^2=o\left(p^2\right)\|\boldsymbol{Y}\|^2,
\end{aligned}$$

where the last equality uses condition (G). Thus,

$$\mathbb{E}\left(\sup_{\boldsymbol{\mu}\in\mathscr{L}}\sum_{i=1}^{k}\left[\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{\mu}\right]_i^2\right)=o\left(p^3\right).$$

Also note that

$$\mathbb{E}\left(\sum_{i=1}^{k}\left[\boldsymbol{Z}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\right]_i^2\right)=\mathbb{E}\left(\operatorname{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)\left(\boldsymbol{Y}-\boldsymbol{\theta}\right)^T\right)\right)$$
$$=\operatorname{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}\right)=\operatorname{tr}\left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\right)=O\left(p\right)$$

by condition (D). Recall that $d_k=O\left(p^{-1}\right)$ by condition (F). It follows that

$$\mathbb{E}\left(\sup_{0\le\lambda\le\infty,\,\boldsymbol{\mu}\in\mathscr{L}}\left|\left(\mathrm{III}\right)_2\right|\right)\le\frac{2}{p}O\left(p^{-1}\right)o\left(p^{3/2}\right)O\left(p^{1/2}\right)=o\left(1\right),$$

which completes our proof.

Proof of Lemma 2 The fact that $\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}\in\mathscr{L}$ is trivial as

$$\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}=\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{Y}=\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y},$$

while the projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$ has induced matrix 2-norm $\left\|\boldsymbol{P}_{\boldsymbol{X}}\right\|_2=1$. Thus, $\left\|\hat{\boldsymbol{\mu}}^{\mathrm{OLS}}\right\|\le\left\|\boldsymbol{P}_{\boldsymbol{X}}\right\|_2\left\|\boldsymbol{Y}\right\|=\left\|\boldsymbol{Y}\right\|$. For $\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}$, note that

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}&=\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Y}\\&=\boldsymbol{A}^{1/2}\left(\boldsymbol{X}\boldsymbol{A}^{-1/2}\right)^T\left(\boldsymbol{X}\boldsymbol{A}^{-1/2}\left(\boldsymbol{X}\boldsymbol{A}^{-1/2}\right)^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{A}^{-1/2}\right)\boldsymbol{A}^{-1/2}\boldsymbol{Y}\\&=\boldsymbol{A}^{1/2}\left(\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{A}^{-1/2}}\right)\boldsymbol{A}^{-1/2}\boldsymbol{Y},\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{A}^{-1/2}}$ is the ordinary projection matrix onto the row space of $\boldsymbol{X}\boldsymbol{A}^{-1/2}$ and has induced matrix 2-norm 1. It follows

$$\left\|\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}\right\|\le\left\|\boldsymbol{A}^{1/2}\right\|_2\left\|\boldsymbol{P}_{\boldsymbol{A}^{-1/2}\boldsymbol{X}}\right\|_2\left\|\boldsymbol{A}^{-1/2}\right\|_2\left\|\boldsymbol{Y}\right\|=\max_{1\le i\le p}A_i^{1/2}\times\max_{1\le i\le p}A_i^{-1/2}\times\left\|\boldsymbol{Y}\right\|.$$

Condition (A) gives

$$\max_{1\le i\le p}A_i^{1/2}=(\max_{1\le i\le p}A_i^2)^{1/4}\le(\sum_{i=1}^{p}A_i^2)^{1/4}=O\left(p^{1/4}\right).$$

Similarly, condition (A′) gives

$$\max_{1\le i\le p}A_i^{-1/2}=(\max_{1\le i\le p}A_i^{-2-\delta})^{1/(4+2\delta)}\le(\sum_{i=1}^{p}A_i^{-2-\delta})^{1/(4+2\delta)}=O\left(p^{1/(4+2\delta)}\right).$$

We then have proved that

$$\left\|\hat{\boldsymbol{\mu}}^{\mathrm{WLS}}\right\| \leq O\left(p^{1/4}\right) O\left(p^{1/(4+2\delta)}\right) \|\boldsymbol{Y}\| = O\left(p^{\kappa}\right) \|\boldsymbol{Y}\|.$$

Proof of Theorem 2 To prove the first assertion, note that

$$\mathrm{URE}\left(\hat{\boldsymbol{B}}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}^{\mathrm{URE}}\right) \leq \mathrm{URE}\left(\tilde{\boldsymbol{B}}^{\mathrm{OL}}, \tilde{\boldsymbol{\mu}}^{\mathrm{OL}}\right)$$

by the definition of $\hat{\boldsymbol{B}}^{\mathrm{URE}}$ and $\hat{\boldsymbol{\mu}}^{\mathrm{URE}}$, so Theorem 1 implies that

$$\begin{aligned}
&l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) \\
&\leq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - \mathrm{URE}\left(\hat{\boldsymbol{B}}^{\mathrm{URE}}, \hat{\boldsymbol{\mu}}^{\mathrm{URE}}\right) + \mathrm{URE}\left(\tilde{\boldsymbol{B}}^{\mathrm{OL}}, \tilde{\boldsymbol{\mu}}^{\mathrm{OL}}\right) - l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) \\
&\leq 2 \sup_{\boldsymbol{B}\in\mathscr{B},\, \boldsymbol{\mu}\in\mathscr{L}} \left|\mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)\right| \underset{p\to\infty}{\to} 0 \text{ in } L^1 \text{ and in probability,}
\end{aligned} \tag{22}$$

where the second inequality uses the condition that $\hat{\boldsymbol{\mu}}^{\mathrm{URE}} \in \mathscr{L}$. Thus, for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) \geq l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) + \epsilon\right) \\
&\leq \mathbb{P}\left(2 \sup_{\boldsymbol{B}\in\mathscr{B},\, \boldsymbol{\mu}\in\mathscr{L}} \left|\mathrm{URE}\left(\boldsymbol{B}, \boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)\right| \geq \epsilon\right) \to 0.
\end{aligned}$$

To prove the second assertion, note that

$$l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) \leq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right)$$

by the definition of $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$ and the condition $\hat{\boldsymbol{\mu}}^{\mathrm{URE}} \in \mathscr{L}$. Thus, taking expectations on Eq. (22) easily gives the second assertion.

Proof of Corollary 1 Simply note that

$$l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right) \leq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}_p}\right)$$

by the definition of $\tilde{\boldsymbol{\theta}}^{\mathrm{OL}}$. Thus,

$$l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{B}}_p, \hat{\boldsymbol{\mu}}_p}\right) \leq l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\mathrm{URE}}\right) - l_p\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^{\mathrm{OL}}\right).$$

Then Theorem 2 clearly implies the desired result.

Proof of Theorem 3 We observe that

$$\begin{aligned}\mathrm{URE}_{\boldsymbol{M}}(\boldsymbol{B}) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right) =& \mathrm{URE}\left(\boldsymbol{B}, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right) \\ &+ \frac{2}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right).\end{aligned}$$

Since

$$\begin{aligned}\sup_{\boldsymbol{B}\in\mathscr{B}}\left|\mathrm{URE}\left(\boldsymbol{B}, \hat{\boldsymbol{\mu}}^{\boldsymbol{M}}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\hat{\boldsymbol{\mu}}^{\boldsymbol{M}}}\right)\right| &\le \sup_{\boldsymbol{B}\in\mathscr{B},\ \boldsymbol{\mu}\in\mathscr{L}}\left|\mathrm{URE}\left(\boldsymbol{B},\boldsymbol{\mu}\right) - l_p\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{\boldsymbol{B},\boldsymbol{\mu}}\right)\right| \\ &\to 0 \text{ in } L^1\end{aligned}$$

by Theorem 1, we only need to show that

$$\sup_{\boldsymbol{B}\in\mathscr{B}}\left|\frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)\right| \to 0 \quad \text{as } p\to\infty.$$

Under Model I,

$$\begin{aligned}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right) &= \sum_{i=1}^{p}\frac{A_i}{A_i+\lambda}[\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}]_{ii} \\ &\le \left(\sum_{i=1}^{p}\left(\frac{A_i}{A_i+\lambda}\right)^2 \times \sum_{i=1}^{p}[\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}]_{ii}^2\right)^{1/2} \\ &\le \left(p\times\sum_{i=1}^{p}[\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}]_{ii}^2\right)^{1/2} \\ &= p^{1/2}\sqrt{\mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A})^T\right)}, \quad \text{for all } \lambda\ge 0,\end{aligned}$$

but $\mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}^T\right) = \mathrm{tr}\left(\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}^2\boldsymbol{M}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\right)$

$= \mathrm{tr}\left(\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}^2\boldsymbol{M}\boldsymbol{X}^T\right)\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{X}^T\right)\right) = O(1)$ by (13) and condition (E). Therefore,

$$\sup_{\boldsymbol{B}\in\mathscr{B}}\left|\frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)\right| = \frac{1}{p}O\left(p^{1/2}\right)O(1) = O(p^{-1/2}) \to 0.$$

Under Model II, $\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}=\boldsymbol{I}_p-\lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}$, where $\boldsymbol{W}^{-1/2}\boldsymbol{V}\boldsymbol{W}^{-1/2}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, $\boldsymbol{\Lambda}=\mathrm{diag}\left(d_1,\ldots,d_k\right)$ with $d_1\le\cdots\le d_k$, and $\boldsymbol{Z}=\boldsymbol{U}^T\boldsymbol{W}^{1/2}\boldsymbol{X}$ as defined in the proof of Theorem 1. Thus,

$$\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)=\mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)-\mathrm{tr}\left(\lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right).$$

We know that $\mathrm{tr}\left(\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)=\mathrm{tr}\left(\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{X}^T\right)\right)=O(1)$ by the assumption (13). $\mathrm{tr}\left(\lambda\boldsymbol{Z}^T\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)=\mathrm{tr}\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\boldsymbol{Z}^T\right)=\mathrm{tr}\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\right)$. The Cauchy-Schwarz inequality for matrix trace gives

$$\begin{aligned}&\left|\mathrm{tr}\left(\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\right)\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\right)\right)\right|\\&\le\mathrm{tr}^{1/2}\left(\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\right)^2\right)\\&\quad\times\mathrm{tr}^{1/2}\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\right).\end{aligned}$$

Since

$$\mathrm{tr}\left(\left(\lambda\left(\lambda\boldsymbol{I}_k+\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\right)^2\right)=\sum_{i=1}^{k}\left(\frac{\lambda d_i}{\lambda+d_i}\right)^2\le kd_k^2=O\left(p^{-2}\right)\quad\text{for all }\lambda\ge0$$

as shown in the proof of Theorem 1 and

$$\begin{aligned}&\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\right)\\&=\mathrm{tr}\left(\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)\\&=\mathrm{tr}\left(\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{A}\boldsymbol{X}^T\right)\boldsymbol{W}\left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}^T\right)\left(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^T\right)^{-1}\left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)\boldsymbol{W}\left(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\right)\right)\\&=O(p^2)\end{aligned}$$

from (13) and condition (F), we have

$$\sup_{\boldsymbol{B}\in\mathscr{B}}\left|\frac{1}{p}\mathrm{tr}\left(\boldsymbol{A}\left(\boldsymbol{A}+\boldsymbol{B}\right)^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{X}}\boldsymbol{A}\right)\right|=\frac{1}{p}\left(O(1)+\sqrt{O\left(p^{-2}\right)\times O\left(p^2\right)}\right)=O(p^{-1})\to0.$$

This completes our proof of (14). With this established, the rest of the proof is identical to that of Theorem 2 and Corollary 1.

References

1. Berger, J.O., Strawderman, W.E.: Choice of hierarchical priors: admissibility in estimation of normal means. Ann. Stat. **24**(3), 931–951 (1996)
2. Brown, L.D.: In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. Ann. Appl. Stat. **2**(1), 113–152 (2008)
3. Copas, J.B.: Regression, prediction and shrinkage. J. R. Stat. Soc. Ser. B Methodol. **45**(3), 311–354 (1983)
4. Efron, B., Morris, C.: Empirical Bayes on vector observations: an extension of Stein's method. Biometrika **59**(2), 335–347 (1972)
5. Efron, B., Morris, C.: Stein's estimation rule and its competitors—an empirical Bayes approach. J. Am. Stat. Assoc. **68**(341), 117–130 (1973)
6. Efron, B., Morris, C.: Data analysis using Stein's estimator and its generalizations. J. Am. Stat. Assoc. **70**(350), 311–319 (1975)
7. Fearn, T.: A Bayesian approach to growth curves. Biometrika **62**(1), 89–100 (1975)
8. Green, E.J., Strawderman, W.E.: The use of Bayes/empirical Bayes estimation in individual tree volume equation development. For. Sci. **31**(4), 975–990 (1985)
9. Hui, S.L., Berger, J.O.: Empirical Bayes estimation of rates in longitudinal studies. J. Am. Stat. Assoc. **78**(384), 753–760 (1983)
10. James, W., Stein, C.: Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 361–379. University of California Press, Berkeley (1961)
11. Jiang, J., Nguyen, T., Rao, J.S.: Best predictive small area estimation. J. Am. Stat. Assoc. **106**(494), 732–745 (2011)
12. Jones, K.: Specifying and estimating multi-level models for geographical research. Trans. Inst. Br. Geogr. **16**(2), 148–159 (1991)
13. Li, K.C.: Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. Ann. Stat. **14(3)**, 1101–1102 (1986)
14. Lindley, D.V.: Discussion of a paper by C. Stein. J. R. Stat. Soc. Ser. B Methodol. **24**, 285–287 (1962)
15. Lindley, D.V.V., Smith, A.F.M.: Bayes estimates for the linear model. J. R. Stat. Soc. Ser. B Methodol. **34**(1), 1–41 (1972)
16. Morris, C.N.: Parametric empirical Bayes inference: theory and applications. J. Am. Stat. Assoc. **78**(381), 47–55 (1983)
17. Morris, C.N., Lysy, M.: Shrinkage estimation in multilevel normal models. Stat. Sci. **27**(1), 115–134 (2012)
18. Normand, S.L.T., Glickman, M.E., Gatsonis, C.A.: Statistical methods for profiling providers of medical care: issues and applications. J. Am. Stat. Assoc. **92**(439), 803–814 (1997)
19. Omen, S.D.: Shrinking towards subspaces in multiple linear regression. Technometrics **24**(4), 307–311 (1982). 1982
20. Raftery, A.E., Madigan, D., Hoeting, J.A.: Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. **92**(437), 179–191 (1997)
21. Robbins, H.: An empirical Bayes approach to statistics. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Contributions to the Theory of Statistics, vol. 1, pp. 157–163. University of California Press, Berkeley (1956)
22. Rubin, D.B.: Using empirical Bayes techniques in the law school validity studies. J. Am. Stat. Assoc. **75**(372), 801–816 (1980)
23. Sclove, S.L., Morris, C., Radhakrishnan, R.: Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. Ann. Math. Stat. **43**(5), 1481–1490 (1972)
24. Stein, C.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Contributions to the Theory of Statistics, vol. 1, pp. 197–206. University of California Press, Berkeley (1956)

25. Stein, C.M.: Confidence sets for the mean of a multivariate normal distribution (with discussion). J. R. Stat. Soc. Ser. B Stat Methodol. **24**, 265–296 (1962)
26. Stein, C.: An approach to the recovery of inter-block information in balanced incomplete block designs. In: Neyman, F.J. (ed.) Research Papers in Statistics, pp. 351–366. Wiley, London (1966)
27. Strenio, J.F., Weisberg, H.I., Bryk, A.S.: Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. Biometrics **39**(1), 71–86 (1983)
28. Tan, Z.: Steinized empirical Bayes estimation for heteroscedastic data. Stat. Sin. **26**, 1219–1248 (2016)
29. Xie, X., Kou, S.C., Brown, L.D.: SURE estimates for a heteroscedastic hierarchical model. J. Am. Stat. Assoc. **107**(500), 1465–1479 (2012)
30. Xie, X., Kou, S.C., Brown, L.D.: Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. Ann. Stat. **44**, 564–597 (2016)