Slab and Shrinkage Linear Regression Estimation

Vali Asimit[*†] Marina Anca Cidota[‡] Ziwei Chen[*] Jennifer Asimit[§]

Abstract

Shrinkage estimation is a statistical methodology that is used to improve parameter estimation by reducing the mean square error, which is expected to improve the out-of-sample performance. This paper focuses on multiple linear regression estimators since the standard ordinary least square estimator is often computationally instable. Its penalized variants such as ridge and LASSO are the usual non-parametric solutions, and such shrinkage methods lead to sparse models and reduce overfitting. Another shrinkage class is the Stein-type shrinkage estimators that use Bayesian arguments to leverage prior information so that the resulting estimators dominate the maximum likelihood estimator. A third class of shrinkage estimators has been used with great success where various estimators are optimally combined to take advantage of the positive traits of each estimator. We provide seven non-parametric multiplicative and linear shrinkage estimators, and provide theoretical guarantees that these new estimators have a lower mean square error than the ordinary least square estimator. We illustrate that such theoretical guarantees are reflected in synthetic and real data experiments, and we choose genetics, machine learning, and finance applications to convince the reader about our contributions.

Keywords and phrases: Multivariate linear regression; Shrinkage estimation.

JEL classification: C13, C51, C52

[*]Bayes Business School, City St George's, University of London, UK. Email addresses: asimit@citystgeorges.ac.uk (Vali Asimit), Ziwei.Chen.3@citystgeorges.ac.uk (Ziwei Chen).

[†]Corresponding author.

[‡]Faculty of Mathematics and Computer Science, University of Bucharest, Romania. Email address: cidota@fmi.unibuc.ro (Marina Anca Cidota).

[§]University of Cambridge, UK. Email addresses: jennifer.asimit@mrc-bsu.cam.ac.uk (Jennifer Asimit).

1 Introduction

Shrinkage is a statistical method that was propelled by *Stein's Paradox* (Stein, 1956, 1960; James and Stein, 1961), which showed that the high-dimensional *Maximum Likelihood Estimator (MLE)* is not the estimator with the lowest estimation error whenever the data are drawn from Gaussian populations. Such a puzzling result surprised the statistical community, but this paradox is explained by the fact that shrinkage introduces a bias-variance tradeoff that improves estimation in a global sense. Stein's shrinkage estimator fundamentally changed the way statisticians could approach high-dimensional estimation, and the main idea stems from introducing bias, which could improve the overall estimation accuracy, and it has been influential in various areas such as applied mathematics, finance, machine learning and statistics.

The original Stein's result showed how to reduce the estimation error for a mean parameter vector under Gaussian parametric assumptions. Specifically, Stein (1956) demonstrated that by combining the information across all variables, one may reduce the *Mean Squared Error (MSE)* – which is the sum of the component-wise mean squared errors – even if the variables are independent. Stein's result was illustrated by applying a *multiplicative shrinkage* (also known as contraction) to a standard estimator (e.g., MLE), and therefore, this method is known to shrink around zero, since the Stein's estimator is a weighted average of the standard estimator and zero-valued estimator known as the *target* estimator. This Stein-type estimator can further be improved by choosing a more natural target estimator than shrinking around 0, and such a method is known as *linear shrinkage*; e.g., see (Lindley, 1962; Efron and Morris, 1972) that proposed shrinking around the mean of variables' means, which is a natural choice. A wide range of shrinkage results under parametric assumptions are extended from Gaussian distributions to spherically symmetric distributions, and a summary could be found in (Fourdrinier et al., 2018).

There is a wide range of Stein's Shrinkage applications in the literature, and we take stock of these applications. *First*, statistical decision theory was the first to use Stein's method at large scale to improve the high-dimensional estimation (Fourdrinier et al., 2018). *Second*, machine learning and statistical learning fields massively explored Stein's idea to achieve sparse and/or more stable estimation methods; e.g., see *Tikhonov penalization* (Tikhonov, 1963; Hoerl and Kennard, 1970), *Basic pursuit* (Chen and Donoho, 1994), *Least Absolute Shrinkage and Selection Operator (LASSO)* (Tibshirani, 1996), *Elastic-Net* (Zou and Hastie, 2005), *Generalized LASSO*, (She, 2009; Tibshirani and Taylor, 2011), and more details are provided in Section 1.1. *Third,*

a more recent adoption of Stein's idea has been in the finance literature where linear shrinkage has been used to stabilize the erratic behavior of the empirical covariance matrix (Ledoit and Wolf, 2004, 2022) via distribution-free methods, while multiplicative and linear shrinkage are used for the weights of industry standard portfolios to optimize the out-of-sample performance of such shrinkage portfolios (Kan and Zhou, 2007; Tu and Zhou, 2011; Kan and Lassance, 2024; Lassance et al., 2024) by assuming Gaussian or some elliptically distributed populations.

The aim of this paper is to provide multiplicative and linear shrinkage estimators for multivariate linear regression parameters that are distribution-free, i.e., without relying on any parametric assumption on the dependent variable, so that the estimation error (measured in terms of MSE) is reduced. Under parametric assumptions – e.g., Gaussian distributed dependent variable – the solution is well-known (Oman, 1991) as the vector parameter in a multivariate linear regression is Gaussian distributed and thus, the well-known Stein's contraction/shrinkage method for mean vectors drawn from multivariate Gaussian populations could be deployed in this particular setting. We provide five multiplicative shrinkage estimators, one linear shrinkage estimator and one non-standard shrinkage estimator that show good performance in simulated and real-data analyses.

Our main contributions are three-fold. *First,* we introduce seven shrinkage distribution-free estimators and show their asymptotic properties for multivariate linear regression parameters when the number of covariates is fixed. *Second,* we empirically show that two of our shrinkage estimators significantly outperform the OLS estimator when both the sample size and number of covariates are large. *Third,* we find that some shrinkage estimators are very effective in reducing the notoriously high estimation error in *Generalized Linear Modeling (GLM)*, though such a conclusion is validated in a follow-up paper of (Asimit et al., 2025a) via extensive simulated and real-data analyses.

1.1 Literature Review

Penalized *multivariate linear regression (MLR)* is a widely studied technique in multivariate analysis to produce accurate and/or parsimonious prediction models. That is, for a response vector $\mathbf{y} \in \Re^n$, covariates matrix $\mathbf{X} \in \Re^{n\times(p+1)}$ corresponding to the p covariates/features and penalty function $g : \Re^p \to \Re_+$, the problem is to understand the properties of the following

estimator:

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\Re^{p+1}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + g(\boldsymbol{\beta}), \tag{1.1}$$

where $||\cdot||_p$ is the usual p-norm on a real vector space. From the computational perspective, (1.1) implies solving an (optimization) instance that may or may not have a unique solution; some of these optimal solutions may be boundary solutions of the feasibility set, which could be problematic if boundary around infinity regions are attained. Such undesirable realizations are controlled through appropriate penalty functions g that have been effective to deploy accurate and/or parsimonious prediction models.

The trivial case with no penalization, $g(\cdot) = 0$ on $\Re^{p+1}$, is the *Ordinary Least Square (OLS)* estimator known as the *Best Linear Unbiased Estimator (BLUE)*, in the sense that there is no other linear and unbiased estimator with lower MSE; this property is a consequence of the *Gauss–Markov Theorem* and it has been vastly investigated in the linear modeling literature. One class of (convex) penalty functions is known as the Tikhonov penalization with $g(\boldsymbol{\beta}) = \lambda||\mathbf{D}\boldsymbol{\beta}||_2^2$ where $\lambda \geq 0$ and $\mathbf{D} \in \Re^{q\times(p+1)}$ with $q \geq 1$ (Tikhonov, 1963); $\mathbf{D}$ is known as the Tikhonov matrix and a special case is the identity matrix setting, i.e., $D = \mathbf{I}_{p+1}$, which is known as *Ridge Regression (RR)* that was formalized in (Hoerl and Kennard, 1970) but the authors discussed a natural extension known as *Generalized RR* where $g(\boldsymbol{\beta}) = ||\operatorname{diag}(\boldsymbol{\lambda})\boldsymbol{\beta}||_2^2$ where $\boldsymbol{\lambda} \geq 0$. Another class of (convex) penalty functions is known as *Least Absolute Shrinkage and Selection Operator (LASSO)* with $g(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1$ where $\lambda \geq 0$ and it was formalized in the seminal paper of (Tibshirani, 1996) which is mathematically equivalent to the *Basic pursuit* problem defined in (Chen and Donoho, 1994); an interesting generalization, known as *Generalized LASSO*, is defined in (Tibshirani and Taylor, 2011) with a similar formulation discussed in (She, 2009), where $g(\boldsymbol{\beta}) = \lambda||\mathbf{D}\boldsymbol{\beta}||_1$ with $\lambda \geq 0$ and $\mathbf{D} \in \Re^{q\times(p+1)}$ such that $q \geq 1$. Such penalization methods are convex and thus, solving (1.1) would require convex optimization algorithms that are scalable and have nice convergence properties. *Elastic-Net* is introduced by (Zou and Hastie, 2005) and combines the good properties of RR and LASSO.

There are other penalized regressions beyond L_1 and L_2 formulations. *Bridge regression* is defined in (Fu, 1998) $g(\boldsymbol{\beta}) = \lambda||\mathbf{D}\boldsymbol{\beta}||_\gamma^\gamma$ where $\lambda \geq 0$, $\gamma > 0$ and $D = \mathbf{I}_{p+1}$. A wide class of concave penalization is introduced in (Fan and Li, 2001) that is shown to equally apply to MLR and its well-known extension, *Generalized Linear Model (GLM)* discussed in (Nelder and

Wedderburn, 1972; McCullagh et al., 1989; Wood, 2017); solving such a general problem does not come without additional computational drawbacks and a Newton-Raphson-like algorithm is provided which (in principle) makes the parameters' tuning even more challenging than the convex 1-norm and 2-norm penalizations.

This paper aims to produce shrinkage MLR estimators that do not require cross-validation and do not consider variable selection aspects. For this reason, $\Sigma := \mathbf{X}^T\mathbf{X}$ is assumed to have full rank, and thus, the OLS and RR estimators $\widehat{\boldsymbol{\beta}}^{OLS} = \Sigma^{-1}\mathbf{X}^T\mathbf{y}$ and $\widehat{\boldsymbol{\beta}}^{RR}(\lambda) = (\Sigma + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$ are uniquely determined; since $\widehat{\boldsymbol{\beta}}^{RR}(0) = \widehat{\boldsymbol{\beta}}^{OLS}$ and $\widehat{\boldsymbol{\beta}}^{RR}(\infty) = \mathbf{0}$, it is sometimes inferred that RR is a shrinkage estimator around the origin. Similar arguments could be used to infer that all previously mentioned penalized MLR estimators are indeed shrinkage estimators, but all of them require cross-validation to estimate the penalty parameters. We aim not to rely on cross-validation that would likely improve the out-of-sample performance. The shrinkage methods considered in this paper are as follows

i) *multiplicative shrinkage* – $\widehat{\boldsymbol{\beta}}(\mathbf{D}) = \mathbf{D}\widehat{\boldsymbol{\beta}}^{OLS}$, where $\mathbf{D} \in \Re^{(p+1)\times(p+1)}$, and we say that shrinkage is made around $\mathbf{0}$ since $\widehat{\boldsymbol{\beta}}(\mathbf{0}) = \mathbf{0}$;

ii) *linear shrinkage* – $\widehat{\boldsymbol{\beta}}(\rho) = (1-\rho)\widehat{\boldsymbol{\beta}}^{OLS} + \rho\widehat{\boldsymbol{\beta}}^{target}$, where ρ is the shrinkage intensity, while $\widehat{\boldsymbol{\beta}}^{target}$ is a target estimator.

The optimal choices for $\mathbf{D}$ and ρ are made such that the theoretical MSE of $\widehat{\boldsymbol{\beta}}(\mathbf{D})$ and $\widehat{\boldsymbol{\beta}}(\rho)$ are minimized. The choice of the target estimator is expected to be a simplified model; e.g., assume a target estimator with uncorrelated covariates meaning that $\widehat{\boldsymbol{\beta}}^{ind} = \big(\operatorname{diag}(\Sigma)\big)^{-1}\mathbf{X}^T\mathbf{y}$.

The paper is organized as follows: our main results are amassed in Section 2; a summary of our numerical experiments are in Section 3, while the summary conclusions are gathered in Section 4. All proofs and supporting information are provided in the *SI Appendices*.

2 Main Results

We start with the multiplicative shrinkage estimators in Section 2.1, followed by the linear and slab shrinkage estimators in Section 2.2 and Section 2.3, respectively. The main features of these estimators are compared in Section 2.4, where we also provide another novel shrinkage estimator that is designed to outperform the RR estimator. Finally, empirical evidence is provided in

Section 2.5 to understand the performance of our novel estimators when the sample size and number of covariates are large.

The main set of assumptions used across all sections are given as Assumption 2.1.

Assumption 2.1. *Let* $\Sigma := \boldsymbol{X}^T\boldsymbol{X}$ *with* $\Sigma \succ 0$. *The linear model assumes that* $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ *for all* $1 \leq i \leq n$, *where* $\boldsymbol{x}_i$ *is the* i^{th} *column of* $\boldsymbol{X}^T$ *and* $\boldsymbol{\beta}$ *is the "true" model parameter vector. Further, the error is independent and identically distributed with zero mean and variance* σ^2.

We denote $a_l(\mathbf{u}) := \mathbf{u}^T\Sigma^{-l}\mathbf{u}$ for all $l \in \mathbb{Z}$ and $\mathbf{u} \in \Re^{p+1}$, where by definition, $\Sigma^{-0} := I_{p+1}$. Note that $a_l(\mathbf{u}) > 0$ for all $\mathbf{u} \in \Re^{p+1} \setminus \{\mathbf{0}\}$ and $l \in \mathbb{Z}$ if $\Sigma \succ 0$ as required by Assumption 2.1.

2.1 Multiplicative Shrinkage

This class – defined as $\widehat{\boldsymbol{\beta}}(\mathbf{D}) = \mathbf{D}\widehat{\boldsymbol{\beta}}^{OLS}$, where $\mathbf{D} \in \Re^{(p+1)\times(p+1)}$ – is discussed in (Hocking et al., 1976) where D is assumed to be a diagonal matrix and data are in canonical form. Specifically, the authors showed that the optimal shrinkage estimator – in terms of MSE – could be found over the following sets: i) $D = a\mathbf{I}_{p+1}$, where $a \in \Re$ (though $a > 0$ is desirable) which is first discussed in (Stein, 1960) in a Bayesian setting, and ii) $D = \text{diag}(\mathbf{b})$, where $\mathbf{b} \in \Re^{p+1}$. We could recover the results in (Hocking et al., 1976) by removing the data assumption of being in canonical form, but also a new result when the matrix D is no longer diagonal. These results are summarized in Theorem 1.

Theorem 1. *Let Assumption 2.1 hold, and multiplicative shrinkage is sought by solving*

$$\min_{D\in\mathcal{D}} MSE\Big(\boldsymbol{D}\widehat{\boldsymbol{\beta}}^{OLS}\Big). \tag{2.1}$$

i) If (2.1) is solved over the feasible set $\mathcal{D}_1 := \{a\boldsymbol{I}_{p+1} : a \in \Re\}$, *then its solution (known from now on as Stein (St) estimator) is unique and denoted as* $a^*\widehat{\boldsymbol{\beta}}^{OLS}$, *where*

$$a^* = \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{\boldsymbol{\beta}^T\boldsymbol{\beta} + M_0^*} \in [0,1), \quad M_1^* := MSE\Big(a^*\widehat{\boldsymbol{\beta}}^{OLS}\Big) = \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}M_0^*}{\boldsymbol{\beta}^T\boldsymbol{\beta} + M_0^*}, \tag{2.2}$$

where $M_0^* := MSE\big(\widehat{\boldsymbol{\beta}}^{OLS}\big) = \sigma^2\,\text{Tr}\left(\Sigma^{-1}\right)$.

ii) If (2.1) is solved over the feasible set $\mathcal{D}_2 := \{\text{diag}(\boldsymbol{b}) : \boldsymbol{b} \in \Re^{p+1}\}$, *then its solution (known from now on as Diagonal shrinkage (DSh) estimator) is unique and denoted as*

$\mathrm{diag}(\boldsymbol{b}^*)\widehat{\boldsymbol{\beta}}^{OLS}$, *where*

$$b_k^* = \frac{\beta_k^2}{\beta_k^2 + \sigma^2\sigma_k} \in [0,1),\ 0 \le k \le p, \quad M_2^* := MSE\Big(\mathrm{diag}(\boldsymbol{b}^*)\widehat{\boldsymbol{\beta}}^{OLS}\Big) = \sum_{k=0}^{p} b_k^*\sigma^2\sigma_k, \quad (2.3)$$

where $\sigma_k = \left(\Sigma^{-1}\right)_{kk} > 0$ *for all* $0 \le k \le p$.

iii) If (2.1) *is solved over the feasible set* $\mathcal{D}_3 := \left\{\boldsymbol{C} \in \Re^{(p+1)\times(p+1)}\right\}$, *then its solution (known from now on as Shrinkage (Sh) estimator) is unique and denoted as* $\boldsymbol{C}^*\widehat{\boldsymbol{\beta}}^{OLS}$, *where* $\boldsymbol{C}^*$ *is the unique solution of the Sylvester equation (in* $\boldsymbol{C}$*)* $\Sigma^{-1}\boldsymbol{C} + \boldsymbol{C}\boldsymbol{\beta}\boldsymbol{\beta}^T = \boldsymbol{\beta}\boldsymbol{\beta}^T$ *and*

$$M_3^* := MSE\Big(\boldsymbol{C}^*\widehat{\boldsymbol{\beta}}^{OLS}\Big) = \sigma^2\,\mathrm{Tr}\left(\left(\boldsymbol{C}^*\right)^T\Sigma^{-1}\boldsymbol{C}^*\right) + \boldsymbol{\beta}^T\left(\boldsymbol{C}^* - I_{p+1}\right)^T\left(\boldsymbol{C}^* - I_{p+1}\right)\boldsymbol{\beta}. \quad (2.4)$$

iv) It is also true that

$$M_3^* \le M_2^* \le M_1^* < M_0^*. \quad (2.5)$$

The middle inequality becomes an identity if and only if $\frac{\beta_k^2}{\sigma_k} = \frac{\beta_0^2}{\sigma_0}$ *for all* $1 \le k \le p$, *while the left-hand side inequality becomes an identity if and only if* $\boldsymbol{C}^*$ *is diagonal.*

Theorem 1 establishes the optimal shrinkage estimators for three search sets (solving (2.1) over $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$), but $a^*\widehat{\boldsymbol{\beta}}^{OLS}$, $\mathrm{diag}(\mathbf{b}^*)\widehat{\boldsymbol{\beta}}^{OLS}$ and $\mathbf{C}^*\widehat{\boldsymbol{\beta}}^{OLS}$ are oracle estimators as all depend upon the unknown population $\boldsymbol{\beta}$ and σ^2. We could replace $\boldsymbol{\beta}$ and σ^2 by their unbiased estimators

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{OLS} \text{ and } \widehat{\sigma^2} = \frac{1}{n-p-1}||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{OLS}||_2^2 \quad (2.6)$$

and define the corresponding plug-in bona fide estimators as follows:

$$\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS},\ \mathrm{diag}(\widehat{\mathbf{b}^*})\widehat{\boldsymbol{\beta}}^{OLS} \quad \text{and} \quad \widehat{\mathbf{C}^*}\widehat{\boldsymbol{\beta}}^{OLS}, \quad \text{where} \quad (2.7)$$

$$\widehat{a^*} := \frac{\left(\widehat{\boldsymbol{\beta}}^{OLS}\right)^T\widehat{\boldsymbol{\beta}}^{OLS}}{\left(\widehat{\boldsymbol{\beta}}^{OLS}\right)^T\widehat{\boldsymbol{\beta}}^{OLS} + \widehat{M_0^*}},\ \widehat{b_k^*} = \frac{\left(\widehat{\boldsymbol{\beta}}_k^{OLS}\right)^2}{\left(\widehat{\boldsymbol{\beta}}_k^{OLS}\right)^2 + \widehat{\sigma^2}\sigma_k} \ \text{ with } 0 \le k \le p,$$

$\widehat{M_0^*} := \widehat{\sigma^2}\,\mathrm{Tr}\left(\Sigma^{-1}\right)$ and $\widehat{\mathbf{C}^*}$ is the unique solution of the Sylvester equation (in $\mathbf{C}$)

$$\Sigma^{-1}\mathbf{C} + \mathbf{C}\widehat{\boldsymbol{\beta}}^{OLS}\left(\widehat{\boldsymbol{\beta}}^{OLS}\right)^T = \widehat{\boldsymbol{\beta}}^{OLS}\left(\widehat{\boldsymbol{\beta}}^{OLS}\right)^T. \quad (2.8)$$

Note that so far, n has been assumed to be fixed and the asymptotic properties of our estimators would require adding the index n to each quantity to signify the fact that the observed sample is based on the first n observations, but we refrain from further complicating our notations. For example, Theorem 2 requires $\frac{1}{n}\Sigma \to \mathbf{\Sigma}_0$ as $n \to \infty$, which means that the (non-random) covariates lead to a sequence of real-valued matrices $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ that converges to an unknown real-valued matrix $\mathbf{\Sigma}_0$, though we remove the index n. Theorem 2 shows the consistency of our St and DSh bona fide estimators, and unfortunately, we could not show the same property for the Sh bona fide estimator.

Theorem 2. *If Assumption 2.1 holds and* $\frac{1}{n}\Sigma \to \mathbf{\Sigma}_0$ *as* $n \to \infty$ *with* $\mathbf{\Sigma}_0 \succ 0$ *for a fixed* p*, then*

$$\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS} - a^*\widehat{\boldsymbol{\beta}}^{OLS} \xrightarrow{p} \mathbf{0}, \tag{2.9a}$$

$$\operatorname{diag}(\widehat{\boldsymbol{b}^*})\widehat{\boldsymbol{\beta}}^{OLS} - \operatorname{diag}(\boldsymbol{b}^*)\widehat{\boldsymbol{\beta}}^{OLS} \xrightarrow{p} \mathbf{0}, \tag{2.9b}$$

$$\widehat{a^*} - a^* \xrightarrow{L_2} 0 \quad \textit{and} \quad \widehat{\boldsymbol{b}^*} - \boldsymbol{b}^* \xrightarrow{L_2} \mathbf{0}. \tag{2.9c}$$

The two bona fide estimators ($\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS}$ *and* $\operatorname{diag}(\widehat{\boldsymbol{b}^*})\widehat{\boldsymbol{\beta}}^{OLS}$) *are consistent estimators of their equivalent oracle estimators* ($a^*\boldsymbol{\beta}^{OLS}$ *and* $\operatorname{diag}(\boldsymbol{b}^*)\boldsymbol{\beta}^{OLS}$)*, i.e., the St and DSh estimators have the following properties*

$$\left|\left|\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS} - a^*\widehat{\boldsymbol{\beta}}^{OLS}\right|\right|_2^2 \xrightarrow{L_2} 0 \quad \textit{and} \quad \left|\left|\operatorname{diag}(\widehat{\boldsymbol{b}^*})\widehat{\boldsymbol{\beta}}^{OLS} - \operatorname{diag}(\boldsymbol{b}^*)\widehat{\boldsymbol{\beta}}^{OLS}\right|\right|_2^2 \xrightarrow{L_2} 0; \tag{2.10}$$

furthermore, the two pairs of estimators have the same asymptotic expected loss, i.e., $\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS}$ *and* $\operatorname{diag}(\widehat{\boldsymbol{b}^*})\widehat{\boldsymbol{\beta}}^{OLS}$ *have the same asymptotic expected loss as* $a^*\boldsymbol{\beta}^{OLS}$ *and* $\operatorname{diag}(\boldsymbol{b}^*)\boldsymbol{\beta}^{OLS}$*, respectively since*

$$\mathbb{E}\left|\left|\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right|\right|_2^2 - \mathbb{E}\left|\left|a^*\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right|\right|_2^2 \to 0, \tag{2.11a}$$

$$\mathbb{E}\left|\left|\operatorname{diag}(\widehat{\boldsymbol{b}^*})\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right|\right|_2^2 - \mathbb{E}\left|\left|\operatorname{diag}(\boldsymbol{b}^*)\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right|\right|_2^2 \to 0. \tag{2.11b}$$

2.2 Linear Shrinkage

We now look at the linear shrinkage case which focuses on identifying the optimal linear shrinkage estimator that is a weighted average between an OLS estimator and a target estimator $\widehat{\boldsymbol{\beta}}^{target}$. We have identified one possible choice for the target estimator, namely, $\widehat{\boldsymbol{\beta}}^{ind}$, by assuming that data are standardized, i.e., the dependent variable and covariates have zero mean. Therefore, we

write the estimation problem in this case by excluding the intercept and looking for estimators that go through the origin, i.e., $y_i = \beta_1 x_{1i} + \dots \beta_p x_{pi} + \epsilon_i$ for all $1 \leq i \leq n$. This means that $\Sigma \in \Re^{p\times p}$ and we aim to choose the minimal MSE estimator from the following set of options

$$\widehat{\boldsymbol{\beta}}^{ind}(\rho) := (1-\rho)\widehat{\widehat{\boldsymbol{\beta}}}^{OLS} + \rho\widehat{\boldsymbol{\beta}}^{ind} = \big(\rho\widetilde{\Sigma}^{-1}\Sigma + (1-\rho)\mathbf{I}_p\big)\widehat{\widehat{\boldsymbol{\beta}}}^{OLS} := \Sigma(\rho)\widehat{\widehat{\boldsymbol{\beta}}}^{OLS}, \tag{2.12}$$

where $\widetilde{\Sigma} = \text{diag}(\Sigma)$ and $\widehat{\widehat{\boldsymbol{\beta}}}^{OLS}$ is the OLS estimator through the origin, while ρ is called the shrinkage intensity estimator. Note that such an optimal estimator has an MSE that is no smaller than M_3^*, but while the oracle Sh and its bona fide estimator are designed to be the "best" multiplicative shrinkage estimator, the latter has its computational drawbacks since the numerical solutions for solving a Sylvester equation for large p is very challenging and there is no theoretical result to ensure that it is a consistent estimator. Thus, one may prefer using a simpler optimal shrinkage estimator such as DSh or St, which (both or one of them) may have a smaller or larger MSE than the optimal Linear shrinkage estimator discussed in Theorem 3.

Theorem 3. *Let Assumption 2.1 hold, and linear shrinkage is sought by solving*

$$\min_{\rho\in\Re} MSE\big(\widehat{\boldsymbol{\beta}}^{ind}(\rho)\big), \tag{2.13}$$

where $\widehat{\boldsymbol{\beta}}^{ind} = \widetilde{\Sigma}^{-1}\boldsymbol{X}^T\boldsymbol{y}$. *Assume that the* p *covariates are standardized to have a zero mean. The unique solution of* (2.13) *(known from now on as Linear Shrinkage (LSh) estimator) is* $\widehat{\boldsymbol{\beta}}^{ind}(\rho^*) = \Sigma(\rho^*)\widehat{\widehat{\boldsymbol{\beta}}}^{OLS}$, *where* $\Sigma(\cdot)$ *on* $\Re$ *and* $\widehat{\widehat{\boldsymbol{\beta}}}^{OLS}$ *are defined in* (2.12), *and*

$$\rho^* = \frac{t_2 - t_1}{t_2 - t_1 + t_3} \in [0,1], \quad M^{*ind} := MSE\Big(\widehat{\boldsymbol{\beta}}^{ind}(\rho^*)\Big) = \frac{t_2(t_1+t_3) - t_1^2}{t_2 - t_1 + t_3}, \tag{2.14}$$

with

$$t_1 := \sigma^2\,\text{Tr}\left(\widetilde{\Sigma}^{-1}\right),\ t_2 := \sigma^2\,\text{Tr}\left(\Sigma^{-1}\right),\ t_3 := \boldsymbol{\beta}^T\left(\widetilde{\Sigma}^{-1}\Sigma - \boldsymbol{I}_p\right)^2\boldsymbol{\beta}. \tag{2.15}$$

Theorem 3 identifies the optimal oracle LSh estimator as it depends upon the unknown population $\boldsymbol{\beta}$ and σ^2. As before, we replace these unknown parameters by their unbiased estimators in (2.6) and define the corresponding plug-in bona fide estimator as follows:

$$\widehat{\boldsymbol{\beta}}^{ind}(\widehat{\rho^*}) \quad \text{where} \quad \widehat{\rho^*} = \frac{\widehat{t_2} - \widehat{t_1}}{\widehat{t_2} - \widehat{t_1} + \widehat{t_3}}, \tag{2.16}$$

$$\widehat{t_1} := \widehat{\sigma^2}\,\mathrm{Tr}\left(\widetilde{\Sigma}^{-1}\right),\ \widehat{t_2} := \widehat{\sigma^2}\,\mathrm{Tr}\left(\Sigma^{-1}\right),\ \widehat{t_3} := \left(\widehat{\widehat{\boldsymbol{\beta}}}^{OLS}\right)^T\left(\widetilde{\Sigma}^{-1}\Sigma - \mathbf{I}_p\right)^2\widehat{\widehat{\boldsymbol{\beta}}}^{OLS}.$$

The asymptotic properties of LSh are given as Theorem 4 which is a replica of Theorem 2.

Theorem 4. *Let Assumption 2.1 hold for a fixed p such that $\frac{1}{n}\Sigma \to \mathbf{\Sigma}_0$ as $n \to \infty$ with $\mathbf{\Sigma}_0 \succ 0$. If $\boldsymbol{\beta}^T\left(\widetilde{\Sigma}_0^{-1}\Sigma_0 - \boldsymbol{I}_p\right)^2\boldsymbol{\beta} \neq 0$ with $\widetilde{\Sigma}_0 := \mathrm{diag}(\mathbf{\Sigma}_0)$, then*

$$\widehat{\boldsymbol{\beta}}^{ind}(\widehat{\rho^*}) - \widehat{\boldsymbol{\beta}}^{ind}(\rho^*) \xrightarrow{p} \boldsymbol{0} \quad \textit{and} \quad \widehat{\rho^*} - \rho^* \xrightarrow{L_2} 0. \tag{2.17}$$

The bona fide LSh estimator, $\widehat{\boldsymbol{\beta}}^{ind}(\widehat{\rho^*})$, is a consistent estimator of its equivalent oracle estimator, $\widehat{\boldsymbol{\beta}}^{ind}(\rho^*)$, i.e.

$$\left|\left|\widehat{\boldsymbol{\beta}}^{ind}(\widehat{\rho^*}) - \widehat{\boldsymbol{\beta}}^{ind}(\rho^*)\right|\right|_2^2 \xrightarrow{L_2} 0; \tag{2.18}$$

furthermore, the bona fide and oracle LSh estimators have the same asymptotic expected loss, i.e.,

$$\mathbb{E}\left|\left|\widehat{\boldsymbol{\beta}}^{ind}(\widehat{\rho^*}) - \boldsymbol{\beta}\right|\right|_2^2 - \mathbb{E}\left|\left|\widehat{\boldsymbol{\beta}}^{ind}(\rho^*) - \boldsymbol{\beta}\right|\right|_2^2 \to 0. \tag{2.19}$$

Note that Theorem 4 assumes $\boldsymbol{\beta}^T\left(\widetilde{\Sigma}_0^{-1}\Sigma_0 - \mathbf{I}_p\right)^2\boldsymbol{\beta} \neq 0$, which is a mild assumption as $\boldsymbol{\beta}^T\left(\widetilde{\Sigma}_0^{-1}\Sigma_0 - \mathbf{I}_p\right)^2\boldsymbol{\beta} \geq 0$ is always true.

2.3 Slab Regression

We introduce a new penalized regression model, which is given as in (1.1) with $g(\boldsymbol{\beta}) := \mu\left(\mathbf{u}^T\boldsymbol{\beta}\right)^2$, where $\mu \geq 0$ and $\mathbf{u} \in \Re^{p+1}$. The new estimator, named *(simple) Slab Regression (SR)* estimator, is defined as follows:

$$\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u}) := \underset{\boldsymbol{\beta}\in\Re^{p+1}}{\mathrm{argmin}}\,\frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \mu\left(\mathbf{u}^T\boldsymbol{\beta}\right)^2 = \underset{\boldsymbol{\beta}\in\Re^{p+1}}{\mathrm{argmin}}\,\boldsymbol{\beta}^T\left(\Sigma + \mu\mathbf{u}\mathbf{u}^T\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}, \tag{2.20}$$

and has the following closed-form due to the well-known Sherman-Morrison identity

$$\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u}) = \left(\Sigma + \mu\mathbf{u}\mathbf{u}^T\right)^{-1}\mathbf{X}^T\mathbf{y} = \left(\Sigma^{-1} - \frac{\mu}{1+\mu\mathbf{u}^T\Sigma^{-1}\mathbf{u}}\Sigma^{-1}\mathbf{u}\mathbf{u}^T\Sigma^{-1}\right)\mathbf{X}^T\mathbf{y}. \tag{2.21}$$

Note that $\Sigma \succ 0$ due to Assumption 2.1, and $\mathbf{u}\mathbf{u}^T \succeq 0$ is true for any $\mathbf{u} \in \Re^{p+1}$, which in turn implies that $\Sigma + \mu\mathbf{u}\mathbf{u}^T \succ 0$, and thus, its inverse exists. In Euclidean geometry, "slab" is a

region between two parallel hyperplanes, which explains the chosen name for our proposed SR estimator.

It is interesting to note that the SR is a special case of the Generalized LASSO estimator introduced in (Tibshirani and Taylor, 2011), but SR is aimed to not rely on cross-validation and is not designed to achieve a parsimonious model as the Generalized LASSO is primarily aiming to. It is clear the Generalized LASSO has a very general formulation that has some interesting properties including uniqueness under some conditions (Ali and Tibshirani, 2019), which is very helpful when characterizing the asymptotic properties of such an estimator.

We are now ready to provide our first main result of this section, stated as Theorem 5, which provides a mathematical characterization of our proposed SR estimator.

Theorem 5. *Let $\mu \geq 0$ and $\boldsymbol{u} \in \Re^{p+1} \setminus \{\boldsymbol{0}\}$ such that Assumption 2.1 is in force.*

i) The instance in (2.20) has a unique solution as in (2.21) that is an interior point of its feasibility set $\Re^{p+1}$. Further,

$$MSE(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\boldsymbol{u})) = \sigma^2 \operatorname{Tr}\left(\left(I_{p+1} - \frac{\mu}{1+\mu\delta}A\right)^T \Sigma^{-1}\left(I_{p+1} - \frac{\mu}{1+\mu\delta}A\right)\right) + \left(\frac{\mu}{1+\mu\delta}\right)^2 \boldsymbol{\beta}^T A^T A \boldsymbol{\beta} \tag{2.22}$$

where $\delta := \boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u}$ and $A := \Sigma^{-1}\boldsymbol{u}\boldsymbol{u}^T$ with $\delta > 0$.

ii) For any $\tilde{\mu} \geq 0$

$$\min_{\boldsymbol{x}\in\Re^{p+1}} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad s.t. \quad -\tilde{\mu} \leq \boldsymbol{u}^T\boldsymbol{x} \leq \tilde{\mu}, \tag{2.23}$$

has a unique solution that is bounded away from neighborhoods of infinity, and strong duality holds in (2.23). There exists $\mu^* \geq 0$ such that the unique solution of (2.23) coincides with the optimal solution in (2.20) with $\mu = \mu^*$.

The second main result illustrates how to optimally find the penalty parameter μ, which is given as Theorem 6. A further MSE reduction is possible by looking within the class of SR estimators with $\mathbf{u}$ that have equal entries, i.e., $\mathbf{u} = v\mathbf{1}$ with $v > 0$.

Theorem 6. *Let $\mu \geq 0$ and Assumption 2.1 holds.*

i) Assume $\boldsymbol{u} \in \Re^{p+1} \setminus \{\boldsymbol{0}\}$. *There exists* $\mu^{**}(\boldsymbol{u}) \in (0, \infty]$ *such that* $MSE\Big(\widehat{\boldsymbol{\beta}}^{SR}(\cdot; \boldsymbol{u})\Big)$ *attains its global minimum on* $\overline{\Re}_+$ *at* $\mu^{**}(\boldsymbol{u})$, *where*

$$\mu^{**}(\boldsymbol{u}) = \begin{cases} \frac{\sigma^2 a_2(\boldsymbol{u})}{\Delta(\boldsymbol{u})}, & \text{if } \Delta(\boldsymbol{u}) > 0, \\ \infty, & \text{if } \Delta(\boldsymbol{u}) \leq 0, \end{cases} \tag{2.24}$$

and $\Delta(\boldsymbol{u}) := \sigma^2\big(a_0(\boldsymbol{u})a_3(\boldsymbol{u}) - a_1(\boldsymbol{u})a_2(\boldsymbol{u})\big) + a_3(\boldsymbol{u})\big(\boldsymbol{\beta}^T\boldsymbol{u}\big)^2$. *Then,*

$$MSE\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu^{**}(\boldsymbol{u}); \boldsymbol{u})\big) < MSE\Big(\hat{\boldsymbol{\beta}}^{OLS}\Big) \tag{2.25}$$

and there exists $\mu^{**}(\boldsymbol{u}) \leq \mu_U^{**}(\boldsymbol{u})$ *such that*

$$MSE\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu^{**}(\boldsymbol{u}); \boldsymbol{u})\big) < MSE\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu; \boldsymbol{u})\big) < MSE\big(\widehat{\boldsymbol{\beta}}^{OLS}\big) \text{ for all } 0 < \mu < \mu_U^{**}(\boldsymbol{u}), \tag{2.26}$$

where $\mu_U^{**}(\boldsymbol{u}) < \infty$ *if and only if* $\mu^{**}(\boldsymbol{u}) < \infty$. *Further,*

$$\widehat{\boldsymbol{\beta}}^{SR}(\mu^{**}(v\boldsymbol{1}); v\boldsymbol{1}) = \widehat{\boldsymbol{\beta}}^{SR}(\mu^{**}(\boldsymbol{1}); \boldsymbol{1}) = \left(I_{p+1} - \frac{\mu^{**}(\boldsymbol{1})}{1+\mu^{**}(\boldsymbol{1})a_1(\boldsymbol{1})}\Sigma^{-1}J_{p+1}\right)\widehat{\boldsymbol{\beta}}^{OLS} \tag{2.27}$$

for all $v \in (0, \infty)$, *where* J_{p+1} *an* $p+1$ *dimensional square matrix of ones.*

ii) Assume that $\boldsymbol{u} \in \Re_+^{p+1} \setminus \{\boldsymbol{0}\}$. *Then,* $\mu^{**}(\boldsymbol{u}) < \infty$ *if and only if*

$$\boldsymbol{u} \text{ is not an eigenvector of } \Sigma \quad \text{or} \quad \boldsymbol{\beta}^T\boldsymbol{u} \neq 0. \tag{2.28}$$

The main advantage of our estimator is the existence of an optimal tuning parameter μ – see (2.24) – that has a guaranteed lower MSE than the OLS estimator for any possible $\mathbf{u} \neq \mathbf{0}$. The next question is how to choose $\mathbf{u}$ and the most obvious choice would be $\mathbf{1}$ due to its simplicity and the MSE invariance property in (2.27). Recall that our SR estimators – either the one in (2.27) or it is equivalent with a general vector $\mathbf{u}$, i.e., $\widehat{\boldsymbol{\beta}}^{SR}\big(\mu^{**}(\mathbf{u}); \mathbf{u}\big)$ – are shown to have lower MSE than M_0^* (MSE of the OLS estimator), but we are not able to conclude whether the SR estimators have always a lower or higher MSE than M_1^* or M_2^*. By design, our SR estimator has no lower MSE than M_3^*. The simulation study in Section 3 shows the performance of these estimators.

SR estimator could be MSE optimal in many ways depending on the slab constraint $\boldsymbol{\beta}^T\mathbf{u}$, and

for implementation purposes, $\widehat{\boldsymbol{\beta}}^{SR}(\mu^{**}(\mathbf{u});\mathbf{u})$ is chosen for a simple choice $\mathbf{u}=\mathbf{1}$ as in (2.27). This simple slab regression estimator could be extended to multiple slab constraints, and we call this new estimator as *Generalized Slab Regression (GSR)* estimator and is defined as follows:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{GSR}(\boldsymbol{\mu}) &:= \underset{\boldsymbol{\beta}\in\Re^{p+1}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||_2^2+\sum_{l\in\mathcal{L}}\mu_l\left(\mathbf{u}_l^T\boldsymbol{\beta}\right)^2 \\ &= \underset{\boldsymbol{\beta}\in\Re^{p+1}}{\operatorname{argmin}} \boldsymbol{\beta}^T\left(\Sigma+\sum_{l\in\mathcal{L}}\mu_l\mathbf{u}_l\mathbf{u}_l^T\right)\boldsymbol{\beta}-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y},\end{aligned} \tag{2.29}$$

where $\mathbf{u}_l$ are some eigenvectors of Σ indexed through $l\in\mathcal{L}\subseteq\{0,\ldots,p\}$. Other slab constraint choices would be possible, but this choice simplifies the remaining derivations; e.g., by taking the standard basis in $\Re^{p+1}$, i.e. by choosing $\mathcal{L}=\{0,\ldots,p\}$, $u_{lk}=1$ if $0\leq l=k\leq p$ and $u_{lk}=0$ otherwise, then (2.29) becomes the so called *Generalized RR* discussed in (Hoerl and Kennard, 1970). The next main result, stated as Theorem 7, provides a mathematical characterization of our proposed GSR estimator.

Theorem 7. *Let $\mu\geq 0$ such that Assumption 2.1 is in force. Further, $\{\lambda_l, 0\leq l\leq p\}$ and $\{\boldsymbol{u}_l, 0\leq l\leq p\}$ are the paired eigenvalues and corresponding orthonormal eigenvectors of Σ meaning that $\boldsymbol{u}_l$ is the corresponding unit eigenvector of λ_l for any $0\leq l\leq p$. Let $\mathcal{L}\subseteq\{0,\ldots,p\}$ be an index set.*

i) The instance in (2.29) has a unique solution as in (2.30) that is an interior point of its feasibility set $\Re^{p+1}$.

$$\widehat{\boldsymbol{\beta}}^{GSR}(\boldsymbol{\mu})=\left(\boldsymbol{I}_{p+1}-\sum_{l\in\mathcal{L}}\frac{\mu_l\lambda_l^{-1}}{1+\mu_l\lambda_l^{-1}}\boldsymbol{u}_l\boldsymbol{u}_l^T\right)\widehat{\boldsymbol{\beta}}^{OLS}. \tag{2.30}$$

Further, the minimal MSE GSR is unique, it is attained at $\mu_l^*=\sigma^2/(\boldsymbol{u}_l^T\boldsymbol{\beta})^2$ for all $l\in\mathcal{L}$, and its MSE is given by

$$MSE\left(\widehat{\boldsymbol{\beta}}^{GSR}(\boldsymbol{\mu}^*)\right)=\sum_{l\notin\mathcal{L}}\sigma^2\lambda_l^{-1}+\sum_{l\in\mathcal{L}}\frac{\sigma^2\lambda_l^{-1}\left(\boldsymbol{u}_l^T\boldsymbol{\beta}\right)^2}{\sigma^2\lambda_l^{-1}+\left(\boldsymbol{u}_l^T\boldsymbol{\beta}\right)^2}. \tag{2.31}$$

ii) For any $\tilde{\mu}_l\geq 0$ with $l\in\mathcal{L}$

$$\min_{\boldsymbol{x}\in\Re^{p+1}}\frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad s.t. \quad -\tilde{\mu}_l\leq\boldsymbol{u}^T\boldsymbol{x}\leq\tilde{\mu}_l, \tag{2.32}$$

has a unique solution that is bounded away from neighborhoods of infinity, and strong duality holds in (2.32). Further, there exists $\tilde{\mu}_l^* \geq 0$ with $l \in \mathcal{L}$ such that the unique solution of (2.32) coincides with the optimal solution in (2.20) with $\mu = \mu^*$.

Theorem 7 provides a rich set of MSE optimal GSR estimators that depend upon the selection of eigenvalues that are adjusted in a certain way; for details, see Section 2.4. We get from (2.30) that GSR estimators share some properties with the MSE optimal SR estimators ($\widehat{\boldsymbol{\beta}}^{SR}(\mu^{**}(\mathbf{u}); \mathbf{u})$ in (2.26) with $\mathbf{u} \in \Re^{p+1} \setminus \{\mathbf{0}\}$) as defined in Theorem 6 with MSE no lower than M_3^*, but both are shrinkage estimators. Besides that, the two sets are quite different and we defer this discussion to Section 2.4. We could see from (2.31) that it is optimal to adjust all eigenvalues and choose the largest index set $\mathcal{L} = \{0, \ldots, p\}$.

Theorem 6 and 7 provide oracles estimators, and as before, we replace $\boldsymbol{\beta}$ and σ^2 by their unbiased estimators in (2.6) and define the corresponding plug-in bona fide estimator for (2.27) and (2.31) with $\mathcal{L} = \{0, \ldots, p\}$ (as it is optimal to adjust all eigenvalues):

$$\widehat{\boldsymbol{\beta}}^{SR}\left(\widehat{\mu^{**}(\mathbf{1})}; \mathbf{1}\right) = \left(I_{p+1} - \frac{\widehat{\mu^{**}(\mathbf{1})}}{1 + \widehat{\mu^{**}(\mathbf{1})} a_1(\mathbf{1})} \Sigma^{-1} J\right) \widehat{\boldsymbol{\beta}}^{OLS}, \tag{2.33a}$$

$$\widehat{\boldsymbol{\beta}}^{GSR}\left(\widehat{\boldsymbol{\mu}^*}\right) = \left(\mathbf{I}_{p+1} - \sum_{l \in \mathcal{L}} \frac{\widehat{\mu_l^*} \lambda_l^{-1}}{1 + \widehat{\mu_l^*} \lambda_l^{-1}} \mathbf{u}_l \mathbf{u}_l^T\right) \widehat{\boldsymbol{\beta}}^{OLS}, \quad \text{where} \tag{2.33b}$$

$$\widehat{\mu^{**}(\mathbf{1})} = \frac{\widehat{\sigma^2} a_2(\mathbf{1})}{\widehat{\sigma^2}\left(a_0(\mathbf{1}) a_3(\mathbf{1}) - a_1(\mathbf{1}) a_2(\mathbf{1})\right) + a_3(\mathbf{1}) \left(\mathbf{1}^T \widehat{\boldsymbol{\beta}}^{OLS}\right)^2} \quad \text{and} \tag{2.33c}$$

$$\widehat{\mu_l^*} = \widehat{\sigma^2} / \left(\mathbf{u}_l^T \widehat{\boldsymbol{\beta}}^{OLS}\right)^2 \quad \text{for all } 0 \leq l \leq p. \tag{2.33d}$$

Note that we assume a mild condition by imposing $\mathbf{1}$ to not be an eigenvector of Σ which guarantees $\mu^{**}(\mathbf{1}) < \infty$ in (2.33c); for details, see (2.28). In addition, $\widehat{\mu^{**}(\mathbf{1})}$ and $\widehat{\mu_l^*}$ are plug-in estimator by using (2.24) and Theorem 7 i).

The asymptotic properties of our SR and GSR estimators are given as Theorem 8 and is a replica of Theorem 2 and 4.

Theorem 8. *Let Assumption 2.1 hold for a fixed p such that $\frac{1}{n}\Sigma \to \boldsymbol{\Sigma}_0$ as $n \to \infty$ with $\boldsymbol{\Sigma}_0 \succ 0$. For any index set $\mathcal{L} \subseteq \{0, \ldots, p\}$, we have that*

$$\widehat{\boldsymbol{\beta}}^{SR}\left(\widehat{\mu^{**}(\boldsymbol{1})}; \boldsymbol{1}\right) - \widehat{\boldsymbol{\beta}}^{SR}\left(\mu^{**}(\boldsymbol{1}); \boldsymbol{1}\right) \xrightarrow{p} \boldsymbol{0} \quad \text{if } \boldsymbol{1} \text{ is not an eigenvector of } \Sigma, \tag{2.34a}$$

$$\widehat{\boldsymbol{\beta}}^{GSR}\left(\widehat{\boldsymbol{\mu}^*}\right) - \widehat{\boldsymbol{\beta}}^{GSR}\left(\boldsymbol{\mu}^*\right) \xrightarrow{p} \mathbf{0}, \quad \textit{if } \Sigma \textit{ have distinct eigenvalues.} \tag{2.34b}$$

Moreover, the following hold.

i) If $\mathbf{1}$ *is not an eigenvector of* Σ*, and there exists a universal (that does not depend on* n*) positive constant* M *such that* $\frac{1}{a_1(\mathbf{1})}\left|\left|\Sigma\right|\right|_F \leq M$*, where* $||A||_F := \sqrt{\mathrm{Tr}(A^T A)}$ *represents the Frobenius norm of matrix* A*, then the SR bona fide estimator,* $\widehat{\boldsymbol{\beta}}^{SR}\left(\widehat{\mu^{**}(\mathbf{1})}; \mathbf{1}\right)$*, is a consistent estimator of its oracle estimator,* $\widehat{\boldsymbol{\beta}}^{SR}\left(\mu^{**}(\mathbf{1}); \mathbf{1}\right)$*, i.e.,*

$$\left|\left|\widehat{\boldsymbol{\beta}}^{SR}\left(\widehat{\mu^{**}(\mathbf{1})}; \mathbf{1}\right) - \widehat{\boldsymbol{\beta}}^{SR}\left(\mu^{**}(\mathbf{1}); \mathbf{1}\right)\right|\right|_2^2 \xrightarrow{L_2} 0; \tag{2.35}$$

furthermore, the two estimators have the same asymptotic expected loss, i.e.,

$$\mathbb{E}\left|\left|\widehat{\boldsymbol{\beta}}^{SR}\left(\widehat{\mu^{**}(\mathbf{1})}; \mathbf{1}\right) - \boldsymbol{\beta}\right|\right|_2^2 - \mathbb{E}\left|\left|\widehat{\boldsymbol{\beta}}^{SR}\left(\mu^{**}(\mathbf{1}); \mathbf{1}\right) - \boldsymbol{\beta}\right|\right|_2^2 \to 0. \tag{2.36}$$

ii) If Σ *has distinct eigenvalues, then the GSR bona fide estimator,* $\widehat{\boldsymbol{\beta}}^{GSR}\left(\widehat{\boldsymbol{\mu}^*}\right)$*, is a consistent estimator of its oracle estimator,* $\widehat{\boldsymbol{\beta}}^{GSR}(\boldsymbol{\mu}^*)$*, i.e.,*

$$\left|\left|\widehat{\boldsymbol{\beta}}^{GSR}\left(\widehat{\boldsymbol{\mu}^*}\right) - \widehat{\boldsymbol{\beta}}^{GSR}\left(\boldsymbol{\mu}^*\right)\right|\right|_2^2 \xrightarrow{L_2} 0; \tag{2.37}$$

furthermore, the two estimators have the same asymptotic expected loss, i.e.,

$$\mathbb{E}\left|\left|\widehat{\boldsymbol{\beta}}^{GSR}\left(\widehat{\boldsymbol{\mu}^*}\right) - \boldsymbol{\beta}\right|\right|_2^2 - \mathbb{E}\left|\left|\widehat{\boldsymbol{\beta}}^{GSR}\left(\boldsymbol{\mu}^*\right) - \boldsymbol{\beta}\right|\right|_2^2 \to 0. \tag{2.38}$$

2.4 Comparative Description of Shrinkage Estimators

This section provides a succinct comparative description of our shrinkage estimators introduced in Sections 2.1 – 2.3. We can achieve that by looking at how the eigenvalues of the covariates covariance matrix are changed by various shrinkage methods introduced in this paper. It interesting to note that all estimators end up becoming multiplicative shrinkage estimators, $\widehat{\boldsymbol{\beta}}^* = \mathbf{D}\widehat{\boldsymbol{\beta}}^{OLS}$ with $\mathbf{D} \in \Re^{(p+1)\times(p+1)}$. If D has an inverse which is guaranteed for all estimators except Sh, i.e., St, DSh, LSh, SR, and GSR, then

$$\widehat{\boldsymbol{\beta}}^* = \mathbf{D}\widehat{\boldsymbol{\beta}}^{OLS} = \mathbf{D}\Sigma^{-1}\mathbf{X}^T\mathbf{y} = (\Sigma^*)^{-1}\mathbf{X}^T\mathbf{y}, \quad \text{with } \Sigma^* := \Sigma\mathbf{D}^{-1}. \tag{2.39}$$

This means that each of the five estimators replaces the empirical (covariates) covariance matrix estimator with an estimator Σ^* that is a multiplicative shrinkage estimator; this differs from the linear covariance shrinkage approach proposed in the seminal paper (Ledoit and Wolf, 2004) where one looks for covariance shrinkage estimators $\Sigma^* = \rho T + (1-\rho)\Sigma$ with T being a shrinkage target matrix. Note that we primarily aim to shrink the MLR's parameter vector and not the covariance matrix, which is indirectly done by our shrinkage estimators. We now analyze the eigenvalues and eigenvectors of Σ^* and compare them with the set of paired eigenvalue-eigenvector's of Σ, $\{(\lambda_k, \mathbf{u}_k) : 0 \leq k \leq p\}$. It is not difficult to see that the eigenvectors of Σ^* for DSh, LSh, and SR are different than those of Σ, while St and GSR preserve the eigenvectors. That is, the eigenvalues of Σ^* for St and GSR are

$$\lambda_k^{*St} = \lambda_k + \frac{M_0^*}{\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right)^2} \quad \text{and} \quad \lambda_k^{*GSR} = \lambda_k I_{\{k\notin\mathcal{L}\}} + \left(\lambda_k + \frac{\sigma^2}{\left(\mathbf{u}_k^T\boldsymbol{\beta}\right)^2}\right) I_{\{k\in\mathcal{L}\}}, \text{ respectively, } (2.40)$$

where $I_{\{A\}}$ is the indicator of set A that takes the value of 1 is A is true and 0, otherwise. We explain in *SI Appendix* III how important the eigenvalues (of the covariates' covariance matrix) are in MLR estimation, where we provide an ample discussion about their impact over the OLS and some shrinkage estimators. It is found in Section 2.5 that St and GSR consistently outperform OLS when both the sample size and number of covariates become large, and we believe that (2.40) plays an important role in justifying that empirical finding.

Any linear regression models requires a "good" estimator for the precision matrix and it is well-known that the inverse of the sample covariance matrix is an unbiased estimator (up to a multiplicative correction factor) of the inverse of the population covariance matrix if the multivariate Gaussian assumption is imposed, but no other equivalent result is known. The conjecture in (Ledoit and Wolf, 2004) suggests that a "good" estimator for Σ^{-1} would reduce (and increase) the large (and low) eigenvalues given Result 9 i). There are two practical solutions to rectify the precision matrix and one way is to adjust large and small eigenvalues of Σ, but the low eigenvalues (especially those close to 0) are the most influential eigenvalues in the estimation of Σ^{-1}, while the large eigenvalues are of lower importance in this case; the other way is to adjust all eigenvalues by keeping their sum ($\text{Tr}(\Sigma)$) unchanged and reduce MSE of the shrinkage covariance estimator, which (Ledoit and Wolf, 2004) had indirectly proposed.

Note that the RR estimator, $\widehat{\boldsymbol{\beta}}^{RR}(\lambda) = (\Sigma^*)^{-1}\mathbf{X}^T\mathbf{y}$ with $\Sigma^* = \Sigma + \lambda\mathbf{I}_{p+1}$, preserves the eigenvectors and $\lambda_k^{*RR} = \lambda_k + \lambda$ for all $0 \leq k \leq p$. Therefore, RR inflates all eigenvalues by the same

value λ, which is unknown and there is no optimal way to estimate it. On the contrary, St and GSR regressions inflate the eigenvalues of Σ, while preserving its eigenvectors, which is similar to RR, though the advantage of St and GSR is that such estimators are optimally estimated without relying on cross-validation, while RR does need cross-validation, which is expected to be sub-optimal.

We now provide a new shrinkage estimator that we name as *Shrinkage Ridge Regression (SRR)*, where Σ is replaced by its linear shrinkage estimator that shrinks around a diagonal matrix with equal entries, i.e., $v := \frac{1}{p+1}\operatorname{Tr}(\Sigma)$, which is similar to (Ledoit and Wolf, 2004). This means that

$$\widehat{\boldsymbol{\beta}}^{SRR}(\rho) = \big(\Sigma^*(\rho)\big)^{-1}\mathbf{X}^T\mathbf{y} \quad \text{with} \quad \Sigma^*(\rho) = (1-\rho)\Sigma + \rho v\mathbf{I}_{p+1}, \tag{2.41}$$

and the optimal ρ^* is chosen so that $MSE\big(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big)$ is minimized rather than minimizing $MSE\big(\Sigma^*(\rho)\big)$ as in (Ledoit and Wolf, 2004). The main SRR result is given as Proposition 1.

Proposition 1. *Let Assumption 2.1 hold. The shrinkage estimator in (2.41) is sought by solving*

$$\min_{0\leq\rho\leq 1} \widehat{MSE\big(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big)}. \tag{2.42}$$

The optimal solution in (2.42) is the Shrinkage Ridge Regression (SRR) estimator

$$\widehat{\boldsymbol{\beta}}^{SRR}(\rho^*) := \sum_{k=0}^{p} \frac{v_k^{1/2}}{(1-\rho^*)\lambda_k + \rho^* v}\boldsymbol{u}_k \quad \textit{with} \quad \rho^* := \operatorname*{argmin}_{0\leq\rho\leq 1} H(\rho), \tag{2.43}$$

where $v_k := \big(\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{u}_k\big)^2$ *for all* $0 \leq k \leq p$ *and*

$$H(\rho) := \frac{1}{n-p-1}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\sum_{l=0}^{p}\frac{v_l}{(1-\rho)\lambda_l+\rho v} + \sum_{l=0}^{p}\frac{\lambda_l v_l}{\big((1-\rho)\lambda_l+\rho v\big)^2}\right)\sum_{k=0}^{p}\frac{\lambda_k}{\big((1-\rho)\lambda_k+\rho v\big)^2}$$
$$+\rho^2\sum_{k=0}^{p}\frac{(\lambda_k - v)^2 v_k}{\big((1-\rho)\lambda_k+\rho v\big)^3}.$$

Note that SRR relies on $\Sigma^*(\rho^*)$ which preserves the eigenvectors, while the eigenvalues are $\lambda_k^{*SRR} = (1-\rho^*)\lambda_k + \rho^* v$ for all $0 \leq k \leq p$, which is another rotation-equivariant covariance matrix estimator that is often considered in linear and non-linear shrinkage estimation when the purpose is to find MSE optimal shrinkage covariance estimators (Donoho et al., 2018; Ledoit and Wolf, 2021, 2022). Since ρ^* is not available in closed-form, we cannot provide asymptotic results

to the SRR estimator, but we can numerically compare SRR to OLS and RR. We conduct a simulation study where the *data generating process (DGP)* is described in *SI Appendix* II.1 – for details, see the *Latent Space Features* setting – which consists of an overparametrized regime; that is, we generate covariates to lie close to a low-dimensional subspace (of dimension f with $f < p$) and Gaussian response variable with standard deviation $\sigma = 5$. Different scenarios are created by varying the ratios p/n and f/p so that we understand how effective SRR (when compared to OLS and RR) is in handling an unstable sample covariance matrix estimator induced by low-dimensional subspace factor structure.

Table 1: **Counts of models achieving the minimum L_2 error**

Normal Distribution: $\sigma = 5$									
$n = 1{,}000$									
p/n	**5%**			**10%**			**25%**		
f/p	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**
OLS	0	0	0	0	0	0	0	0	0
RR	54	45	80	43	60	48	0	0	0
SRR	196	205	170	207	190	202	250	250	250
p/n	**50%**			**75%**			**95%**		
f/p	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**
OLS	0	0	0	0	0	0	0	0	0
RR	0	0	0	0	0	0	0	0	0
SRR	250	250	250	250	250	250	250	250	250

Note: We tabulate counts of how many times each estimator (OLS, RR, and SRR) achieves the lowest L_2-distance (from the "true" regression parameters) across $N = 250$ replications of samples of size $n = 1,000$ for various choices of p/n and f/p; the best-performing method highlighted in red.

Our numerical results are summarized in Table 1 where we report how many times each estimator achieves the smallest L_2-distance (from the "true" regression parameters) across $N = 250$ replications under various settings. The SRR estimator consistently achieves the lowest L_2 error by large margins, and the evidence is overwhelming when $p/n \geq 25\%$. More granular empirical evidence to capture the outperformance of SRR over RR is available in *SI Appendix* II.4.

2.5 How Large is Large?

The asymptotic behavior of our shrinkage estimators has been discussed so far under the setting of fixed p and large n. A key element in our proofs is the uniform integrability of $\widehat{\boldsymbol{\beta}}^{OLS}$ that allows us to show the equivalence between the oracle shrinkage estimators and their bona fide estimators. Note that the Kolmogorov setting where $p/n \to k \in (0, 1)$ as $n \to \infty$ requires a very different setting and technical tools which is beyond the scope of this paper. We are actively thinking about how to perform MLR shrinkage in this setting, but a natural question is how

our estimators would behave under the Kolmogorov setting which is the purpose of this section. Some recent research outputs have shown that the OLS estimator has a non-zero asymptotic MSE under the Kolmogorov setting, which is shown via probabilistic heuristics in (El Karoui et al., 2013), though more rigorous arguments are available in (El Karoui, 2013; Donoho and Montanari, 2016). These papers assumed covariates as random, which is different than the classical fixed p setting that we have considered in this paper.

The supplementary material in *SI Appendix* III amasses a series of interesting findings. *First,* we illustrate in *SI Appendix* III.1 some patterns about the empirical eigenvalues that would give some empirical evidence about the covariance matrix empirical estimator; such empirical evidence is relevant as the eigenvalues play an important role in some of our novel shrinkage estimators as explained in Section 2.4. These findings are summarized as Result 9.

Result 9. *i) The largest and lowest empirical eigenvalues are overestimated and underestimated, respectively.*

ii) The overall estimation error in empirical eigenvalues is reduced when the strength of dependence becomes more extreme (either positive or negative); e.g., see Figure 3.

iii) The eigenvalues' bias does not uniformly decrease from the largest to the lowest empirical eigenvalue, especially when the "true" eigenvalues are clustered; e.g., see Figure 4.

Second, we found in *SI Appendix* III.2 that the model fitted with independent covariates yields a lower estimation error than the one with dependent covariates as long as the eigenvalues are preserved. This property is true for the fixed p and large n case, but also under the Kolmogorov setting, which implies that running MLR in a very high dimension would be more efficient by considering a sparse model without increasing the estimation error. *Third,* we also empirically found in *SI Appendix* III.2 that St and GSR shrinkage estimators outperform OLS in the Kolmogorov setting (though GSR outperforms St), which gives us confidence to validate the motivation of this section hoping that some of our novel shrinkage estimators may be more effective than OLS in the Kolmogorov setting for which we have not established theoretical results. These findings are summarized as Result 10.

Result 10. *i) Assuming that the population eigenvalues are preserved, the empirical eigenvalues for independent and dependent Gaussian covariates are estimated with the same error. This invariance property is not true for OLS regression parameters where the es-*

timation error is reduced for independent Gaussian covariates as compared to dependent Gaussian covariates.

ii) Assuming that the population eigenvalues are preserved, the estimation error of St, DSh and GSR estimators are lower for independent Gaussian covariates as compared to their corresponding dependent Gaussian covariates.

iii) Assuming the Kolmogorov setting with large n and p, St and GSR consistently outperform OLS. St and GSR perform similarly in estimation error for small p/n though St shows a slight advantage in such settings, while GSR clearly outperforms St for big choices of p/n.

3 Numerical Experiments

The theoretical properties of multiple shrinkage estimators have been investigated in the previous section and we now evaluate their performance through synthetic data (see Section 3.1) and three real datasets. We choose three applications from very different fields. The *first* application is given in Section 3.2 and examines how helpful our shrinkage estimators are to improve statistical fine-mapping; these methods aim to identify causal variants underlying genetic associations with a trait (response variable). The *second* application is given in Section 3.3 where we show that our shrinkage estimators could reduce the prediction error in GLM modeling; we chose a cyber-sickness dataset to make our point to predict motion sickness which is a research question raised in the virtual reality field. The implications of our findings go well beyond the small application in Section 3.3, and in parallel to this paper, we have finished another paper (Asimit et al., 2025a) that provides ample evidence that the estimation error could be massively reduced by using our shrinkage estimators. The *third* application is discussed in Section 3.4 where shrinkage estimators are shown to be very effective in enhancing investors' decisions under uncertainty, which is in accordance with the fast-growing finance literature focusing on shrinkage methodologies.

Note that the RR and SRR estimators are included only in Section 3.4, which is the only case where covariates may exhibit an ill-conditioned covariance matrix. These two estimators have performed much worse than OLS in most simulation scenarios considered in Section 3.1, which explains why we have discarded RR and SRR.

3.1 Simulation Results Analysis

Two sets of simulation studies are considered in this section, one is for continuous data and another one for counting data. We compare the performance of i) OLS estimator, ii) St estimator as in (2.2), iii) DSh estimator as in (2.3), iv) Sh estimator as in (2.4), v) SR estimator as in (2.27), and vi) GSR estimator as in (2.30). Note that the LSh estimator is not included, as it behaves similarly to OLS when applied to centered data. The simulation settings are described in *SI Appendix* II.1, with results being summarized in *SI Appendix* II.2 as Tables 3 – 6.

The *first* simulation study examines continuous dependent Gaussian covariates and all results are presented in Tables 3 – 5. This study compares performance across different dependent variable's distributions, distinguishing between lighter-tailed cases (see Tables 3 and 4) and heavier-tailed cases (see Table 5). The overall conclusions are that SR and GSR consistently outperform OLS and the other three shrinkage estimators in most scenarios. A further improvement of the St, DSh and GSR estimators is investigated in Section II.3 where cross-validation is introduced to provide different weights between bias and variance when selecting the optimal shrinkage estimator; we find that GSR may benefit from such adjustment, but a neutral effect is observed on St and DSh estimators.

The *second* simulation study considers counting covariates with Gaussian dependence, with results being summarized in Table 6. This setting is common in genetics applications, where covariates are genotype scores at genetic variants. The genotype score counts the number of effect alleles at a variant and follows a binomial distribution with $N_q = 2$ number of trials and $q_0 = EAF$ success probability, where EAF is the effect allele frequency. The picture is a tad different than what we have found for continuous covariates, and we note that OLS behaves very well only for cases with small variability and small sample size, while DSh and GSR outperform all estimators in the remaining settings. On the contrary, St is by far the best estimator when a larger variability in the response variable is observed.

3.2 Application to Statistical Fine-mapping in Genetics

Based on the results from our *second* simulation study in Section 3.1, where DSh, GSR, and St performed better in handling discrete correlated Gaussian covariates, we use these three estimators for this realistic simulation study and and compare them with the standard OLS

estimator. The Sh and SR estimators are not included because they showed weaker performance in the simulations. This choice ensures that the selected methods are suitable for statistical fine-mapping in genetics, where genetic variants are often highly correlated, and accurate effect estimation is important.

There is potential for our new regression effect estimates to improve current genetic analysis approaches, such as fine-mapping. In *genome-wide association studies (GWAS)*, genetic variants are each tested for association with a quantitative (e.g., cholesterol level) or binary trait (presence/absence of coronary artery disease) using a linear model. As many genetic variants are highly correlated, GWAS report the genetic variants with the lowest genome-wide significant ($P < 5 \times 10^{-8}$) p-value among correlated variants. However, the variant with the smallest p-value (lead variant) is not necessarily causal and may be detected because of correlation (i.e. *linkage-disequilibrium (LD)*) with the causal variant(s). The identification of causal variants that underlie genetic associations is key to facilitating translation into new therapeutic targets or elucidating new biological insights. Statistical fine-mapping is therefore needed to refine sets of potential causal variants within a region constructed around a lead variant (Hutchinson et al., 2020). Fine-mapping prioritization of likely causal variants (i.e., those with a high *Marginal Posterior Probability (MPP)* of causality) may be improved through joint analyses of multiple traits, as biologically related traits often share causal variants.

Bayesian methods are common in fine-mapping – e.g. JAM (Newcombe et al., 2016), FINEMAP (Benner et al., 2016) – and a *Bayes' factor (BF)* is used to summarize the evidence of association for each combination of *variants (SNPs)* compared to the null model of no causal variants. The *Joint Analysis of Marginal summary statistics (JAM)* fine-mapping approach uses a sparse Bayesian regression framework and infers joint LD-adjusted multi-SNP models, highlighting the best multi-SNP models (high posterior probability) considering a thinned subset of SNPs that are not in high correlation (Newcombe et al., 2016). JAM (and many other fine-mapping methods) uses the GWAS effect estimates from the thinned subset of SNPs to fit the multi-SNP models. JAM was expanded to account for all thinned out SNPs by considering all the possible models formed by all the combinations of SNPs in the JAM model, replacing SNPs in the model with highly correlated SNPs that were previously thinned out. The expanded version of JAM has been integrated into *flashfm (flexible and shared information fine-mapping)* multi-trait fine-mapping, where multi-trait model priors are upweighted when causal variant(s) are shared among traits (Hernández et al., 2021).

Both JAM and flashfm make use of GWAS summary statistics, in particular the genetic effect estimates at each genetic variant. In the case where a study consists of unrelated individuals, these effect estimates are calculated using ordinary least squares. Here, we modify JAM (expanded version) and flashfm such that the effect estimates for single-SNP and multi-SNP models are calculated using GSR, St, and DSh. In simulations within the region harboring the gene *IL2RA*, we show that estimates based on DSh have the potential to outperform those from OLS and that multi-trait fine-mapping gives further power improvements over each of the single-trait approaches.

Table 2: **Power and FDR (false discovery rate) comparison for single and multi-trait fine-mapping based on four different estimators.**

	Power		FDR	
Method	**Trait 1**	**Trait 2**	**Trait 1**	**Trait 2**
single-OLS	0.76	0.7	0.01	0.01
multi-OLS	0.835	0.815	0.01	0.005
single-St	0.78	0.65	0.005	0.035
multi-St	0.855	0.77	0.005	0.03
single-DSh	**0.815**	0.68	0.063	0.063
multi-DSh	**0.865**	0.785	0.063	0.058
single-GSR	0.78	0.645	0.005	0.025
multi-GSR	0.855	0.77	0.005	0.02

In our *IL2RA* simulations of 100 replications, we set plausible causal variants that have been extensively explored in previous studies *SI Appendix* IV and set uniform random effect sizes (between 0.15 and 0.4). Power is evaluated by using an MPP threshold of 0.9; all results are displayed in Table 2. Among the four single-trait versions of JAM (expanded), the highest power of 0.815 is attained by DSh estimation, which is an increase of 0.055 over that from OLS (power = 0.760). A further increase of 0.05 is achieved by DSh estimation within the flashfm multi-trait approach (power = 0.865), which is an increase of 0.03 over OLS estimation within flashfm (power = 0.835). For trait 2, the power attained by OLS and DSh are similar for single-trait fine-mapping (single-OLS power = 0.70; single-DSh power = 0.68) and multi-trait fine-mapping gives a further increase of more than 0.10 for each (multi-OLS power = 0.815; single-DSh power = 0.785).

The observed improvement with DSh is consistent with our previous simulation results, where DSh gave the best-performing regression model in the setting of two causal variants and low trait variability (see Table 6 when $\sigma = 1$).

3.3 Application to GLM Modeling – Cyber-sickness Data

GLM is a generalization of MLR by including a non-Gaussian response variable assumption and the standard implementation is through the *Iteratively Reweighted Least Squares (IRLS)* method that iteratively solves OLS instances – in fact, *Weighted Least Squares (WLS)* instances with known weights – which explains its computational efficiency (Nelder and Wedderburn, 1972; McCullagh et al., 1989; Wood, 2017); for details, also see *SI Appendix* VII. We show in this section that one may replace deploying IRLS with OLS by one of our five shrinkage methods – we choose SR, GSR, St, DSh, and Sh in this real data analysis – so that the estimation error is improved. Our small data analysis compares an OLS solver to solvers based on the five shrinkage methods when solving Logistic and Poisson GLMs. We conclude that improvement is i) very limited for Logistic regression (see Tables 7 and 8) and ii) 3% to 7% improvement when using St shrinkage for Poisson regression (see Table 9). The latter conclusion may look as an obsolete result and we further extend this analysis in a follow-up paper (Asimit et al., 2025a) where we found that St, DSh, SR and GSR consistently outperform OLS in IRLS implementations for Poisson and Gamma GLMs via extensive simulated and real-data analyses.

The real data analyses in this section rely on a cyber-sickness dataset* used in the machine learning literature. Cyber-sickness is similar to motion sickness, but it happens while using electronic screens rather than through actual movement. It refers to a set of symptoms that fall into three categories: nausea, oculomotor issues (such as eye strain and fatigue), and general disorientation. People may experience cyber-sickness when using *virtual reality (VR)* systems but also through using everyday electronic devices. Automatic real-time detection of cyber-sickness may help get a better understanding of the phenomenon and develop effective countermeasures, which in turn could reduce visual discomfort and improve the user's experience.

This physiological dataset includes recordings from 23 participants who were immersed in a VR roller coaster simulation. The data are labeled with cyber-sickness severity scores on a scale from 0 (no cyber-sickness) to 10 (high cyber-sickness), which is the target/response variable, but a more detailed data description is available in *SI Appendix* V. Two types of models are formulated to evaluate the effectiveness of our proposed estimators. *First,* a binary classification problem is performed, which is a Logistic GLM with a *logit* link function, where the *Fast Motion Scale (FMS)* scores are reduced to binary outcomes for specific pairs of classes. *Second,* a Poisson

*Available at https://github.com/shovonis/CyberSicknessClassification

GLM with a *log* link function is performed to mimic a multi-class classification by dividing the cyber-sickness severity into four ordinal levels, grouping the FMS scores into distinct categories. Details of these models and their formulations are provided in *SI Appendices* V.1 and V.2.

Logistic GLM with a *logit* link function and Poisson GLM with a *log* link function are deployed for two feature sets from the physiological dataset to represent different levels of multicollinearity within the feature space i) a smaller set with 13 features and ii) a larger set with 130 features, which exhibits a more pronounced multicollinearity effect. Our analyses include the features in both raw and standardized forms in order to understand whether the estimation error is influenced by this choice of data. The comparative performance among different models (measured by the estimated MSE) shows some positive and neutral benefits of using our proposed shrinkage estimators in solving GLMs through IRLS. *First,* solving Logistic GLMs with St, Sh, SR or GSR would lead to similar performance as compared to the baseline OLS solver, while the DSh estimator performed poorly in comparison to all the other methods; for details, see Tables 7 and 8. *Second,* solving Poisson GLMs with St clearly improves the estimation error as compared to the benchmark OLS solver when features are standardized (see Table 9).

3.4 Real Data Analysis – Portfolio Investment

As mentioned in Section 1, there is growing finance literature that adopts shrinkage methods to enhance investors' decisions under uncertainty, and portfolio theory has benefited the most from adopting shrinkage methodologies. Since L_2 linear regressions and investment decisions where investors orders their decisions (measure risk) via variance preferences are mathematically equivalent, we now compare OLS and our shrinkage estimators (St, DSh, Sh, SR, GSR, and SRR) to construct risk-minimizing portfolios. Factor models have been massively explored in the finance literature, where it is argued that asset returns can be represented by a smaller set of observable or engineered covariates. Thus, we include in our analysis the RR and SRR estimators.

Investment decisions with variance preferences mean constructing *Global Minimum Variance (GMV)* portfolios as defined in (VI.1) and discussed in *SI Appendix* VI.1. Unlike the mean-variance portfolio introduced by Markowitz (1952), which makes investment decisions by balancing risk (measured by variance) and reward (measured by realized expected return), the GMV portfolio focuses only on minimizing the risk. The mean-variance approach requires estimates

of both the mean and covariance matrix of asset returns, but mean return estimates often have large errors, making portfolios unstable and leading to poor out-of-sample performance (Merton, 1980). In contrast, the GMV portfolio relies only on covariance matrix estimates, reducing sensitivity to errors in mean estimates and achieving better out-of-sample performance; GMV remains affected by covariance estimation errors, for which robust methods have been proposed to improve its out-of-sample performance (DeMiguel and Nogales, 2009).

In this analysis, we construct GMV portfolios to S&P500 (Standard & Poor's 500) data, which is an index of 500 large U.S. firms widely used to measure the US market performance. Our dataset – that denoted as *DA441* – contains daily asset returns about 441 firms that had been S&P500 constituents for at least one time during the observation period (January 1, 2000 to December 31, 2023); note that these 441 firms are selected among the 1,070 S&P500 constituents that had been during the observation period, and the 441 selected firms are those that had been listed on the US stock exchanges without interruption. Additional details on the dataset are in *SI Appendix* VI.1 and (Asimit et al., 2025b), while details about the portfolio construction are also given in *SI Appendix* VI.1. Numerical experiments are made across periods with various market conditions in *SI Appendix* VI.2. The out-of-sample performance is investigated by applying a rolling-window scheme with five-year and ten-year training periods, each followed by a three-month testing window.

The main conclusions of our analyses are three-fold. *First,* eigenvalue-driven methods (RR, GSR and SRR) are useful to stabilize the risk, but are not effective in terms of reward (low expected returns) and risk-adjusted performance (low Sharpe ratios, which are calculated as expected return per a unit of risk). *Second,* St, DSh and Sh show very good performance in terms of reward and risk-adjusted performances with St being the "best" option. *Third,* OLS is showing very poor performance irrespective of the market conditions.

4 Conclusions

A wide range of distribution-free shrinkage estimators have been discussed within the topic of multivariate linear regression. Our theory is focused on the setting with a fixed number of covariates, but we empirically show that some of our shrinkage estimators outperform OLS by large margins when both the sample size and number of covariates get large. The advantage of

using our novel estimators has been illustrated through three very different applications, where we also find that our shrinkage estimators are very effective in significantly reducing the high estimation errors in GLM modeling.

References

Afendras, G. and Markatou, M. (2016). Uniform integrability of the ols estimators, and the convergence of their moments. *Test*, 25:775–784.

Ali, A. and Tibshirani, R. J. (2019). The generalized lasso problem and uniqueness. *Electronic Journal of Statistics*, 13(2):2307–2347.

Asimit, J. L., Rainbow, D. B., Fortune, M. D., Grinberg, N. F., Wicker, L. S., and Wallace, C. (2019). Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. *Nature communications*, 10(1):3216.

Asimit, V., Chen, Z., Drimitrova, D., Xie, Y., and Zhang, Y. (2025a). Shrinkage glm modelling. *preprint*.

Asimit, V., Peng, L., Tunaru, R., and Zhou, F. (2025b). Risk budgeting under general risk measures. *preprint*.

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501.

Bogachev, V. I. (2007). *Measure theory*, volume 1. Springer.

Chen, S. and Donoho, D. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE.

Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

Coppersmith, D., Hoffman, A. J., and Rothblum, U. G. (1997). Inequalities of rayleigh quotients and bounds on the spectral radius of nonnegative symmetric matrices. *Linear algebra and its applications*, 263:201–220.

DeMiguel, V. and Nogales, F. J. (2009). Portfolio selection with robust estimation. *Operations research*, 57(3):560–577.

Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969.

Donoho, D. L., Gavish, M., and Johnstone, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of statistics*, 46(4):1742.

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

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

Fourdrinier, D., Strawderman, W. E., and Wells, M. T. (2018). *Shrinkage estimation*. Springer.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949.

Hernández, N., Soenksen, J., Newcombe, P., Sandhu, M., Barroso, I., Wallace, C., and Asimit, J. (2021). The flashfm approach for fine-mapping multiple quantitative traits. *Nature communications*, 12(1):6147.

Hocking, R. R., Speed, F., and Lynn, M. (1976). A class of biased estimators in linear regression. *Technometrics*, pages 425–437.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hutchinson, A., Asimit, J., and Wallace, C. (2020). Fine-mapping genetic associations. *Human Molecular Genetics*, 29(R1):R81–R88.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 361–379. University of California Press.

Kan, R. and Lassance, N. (2024). Optimal portfolio choice with fat tails and parameter uncertainty. *Journal of Financial and Quantitative Analysis, forthcoming*.

Kan, R. and Zhou, G. (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3):621–656.

Kundu, R., Islam, R., Calyam, P., and Hoque, K. (2022). Truvr: Trustworthy cybersickness detection using explainable machine learning. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 777–786, Los Alamitos, CA, USA. IEEE Computer Society.

Lassance, N., Vanderveken, R., and Vrins, F. (2024). On the combination of naive and mean-variance portfolio strategies. *Journal of Business & Economic Statistics*, 42(3):875–889.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2021). Shrinkage estimation of large covariance matrices: Keep it simple, statistician? *Journal of Multivariate Analysis*, 186:104796.

Ledoit, O. and Wolf, M. (2022). The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 20(1):187–218.

Lindley, D. V. (1962). Discussion of the article by stein. *Journal of the Royal Statistical Society, Series B*, 24:265–296.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

McCullagh, P., Nelder, J., and Wedderburn, R. (1989). *Generalized Linear Models*. Second ed., Chapman and Hall/CRC.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361.

Muirhead, R. J. (1987). Developments in eigenvalue estimation. In *Advances in Multivariate Statistical Analysis: Pillai Memorial Volume*, pages 277–288. Springer.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384.

Newcombe, P. J., Conti, D. V., and Richardson, S. (2016). Jam: A scalable bayesian framework for joint analysis of marginal snp effects. *Genetic Epidemiology*, 40(3):188–201.

Oman, S. D. (1991). Random calibration with many measurements: An application of stein estimation. *Technometrics*, 33(2):187–195.

Pardo, L. (2005). *Statistical Inference Based on Divergence Measures (1st ed.)*. Chapman and Hall/CRC.

She, Y. (2009). Sparse regression with exact clustering. *Electronic Journal of Statistics*, 39(3):1360–1392.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 197–207. University of California Press.

Stein, C. (1960). Multiple regression contributions to probability and statistics. *Essays in Honor of Harold Hotelling*, 103.

Su, Z., Marchini, J., and Donnelly, P. (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady akademii nauk*, volume 151, pages 501–504. Russian Academy of Sciences.

Tu, J. and Zhou, G. (2011). Markowitz meets talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, 99(1):204–215.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

Supplementary material of "Slab and Shrinkage Linear Regression Estimation"

I Proofs

I.1 Proof of Theorem 1

We first prove Part i) for which we need to minimize in $a \in \Re$

$$MSE\Big(a\widehat{\boldsymbol{\beta}}^{OLS}\Big) = a^2\sigma^2 \operatorname{Tr}\left(\Sigma^{-1}\right) + (a-1)^2\boldsymbol{\beta}^T\boldsymbol{\beta} = a^2 M_0^* + (a-1)^2\boldsymbol{\beta}^T\boldsymbol{\beta},$$

which is strictly convex, and thus, it has a unique solution $a^* \in [0,1)$. This concludes (2.2) and thus, Part i) is fully justified.

Part ii) is now argued, where we minimize in $\mathbf{b} \in \Re^{p+1}$

$$\begin{aligned} MSE\Big(\operatorname{diag}(\mathbf{b})\widehat{\boldsymbol{\beta}}^{OLS}\Big) &= \sigma^2 \operatorname{Tr}\left(\operatorname{diag}(\mathbf{b})\Sigma^{-1}\operatorname{diag}(\mathbf{b})\right) + \boldsymbol{\beta}^T\left(\operatorname{diag}(\mathbf{b})-I_{p+1}\right)^T\left(\operatorname{diag}(\mathbf{b})-I_{p+1}\right)\boldsymbol{\beta} \\ &= \sigma^2 \operatorname{Tr}\Big(\Sigma^{-1}\operatorname{diag}\left(\mathbf{b}^2\right)\Big) + \boldsymbol{\beta}^T \operatorname{diag}\left(\mathbf{b}-\mathbf{1}\right)^2\boldsymbol{\beta} \qquad \text{(I.1)} \\ &= \sum_{k=0}^{p}\Big(\sigma^2\sigma_k b_k^2 + (b_k-1)^2\beta_k^2\Big), \end{aligned}$$

where squaring a vector is made component-wise. The above is a sum of separable (with respect to each b_k) strictly convex quadratic functions since $\sigma_k > 0$ for all $0 \le k \le p$. The latter is true since there exists an orthogonal matrix Q, i.e., $QQ^T = I_{p+1}$, and a diagonal matrix $D = \operatorname{diag}(\mathbf{d})$ with $\mathbf{d} > \mathbf{0}$ (since $\Sigma^{-1} \succ 0$) such that $\Sigma^{-1} = QDQ^T$, and in turn, we have that

$$\left(\Sigma^{-1}\right)_{kk} = \sum_{k'=0}^{p}\left(Q^T\right)_{kk'}\left(D\right)_{k'k'}\left(Q\right)_{k'k} = \sum_{k'=0}^{p} d_{k'}q_{k'k}^2 \ge 0.$$

The inequality becomes an identity if and only if $q_{k'k}^2 = 0$ for all $0 \le k' \le p$, but this can not be true since $1 = \sum_{k'=0}^{p} q_{k'k}^2$ for all $0 \le k \le p$ as $QQ^T = I_{p+1}$, and thus, the above holds with strict inequality and in turn, $\sigma_k > 0$ for all $0 \le k \le p$. The strict convexity of (I.1) and some simple algebraic manipulations conclude Part ii).

We now prove Part iii) for which we need to minimize in $C \in \Re^{(p+1)\times(p+1)}$

$$MSE\Big(C\widehat{\boldsymbol{\beta}}^{OLS}\Big) = \sigma^2 \operatorname{Tr}\left(C^T\Sigma^{-1}C\right) + \boldsymbol{\beta}^T(C-I_{p+1})^T(C-I_{p+1})\boldsymbol{\beta}. \tag{I.2}$$

Note that the above is convex in C and its global minimum is unique if Σ^{-1} and $-\boldsymbol{\beta}^T\boldsymbol{\beta}$ have no common eigenvalues, which coincides with the well-known result regarding the Sylvester equation. Specifically, one may find that by first using (I.2) to get that

$$\begin{aligned}\frac{\partial MSE\Big(C\widehat{\boldsymbol{\beta}}^{OLS}\Big)}{\partial C} &= \sigma^2\frac{\partial \operatorname{Tr}\left(C^T\Sigma^{-1}C\right)}{\partial C} - \frac{\partial \boldsymbol{\beta}^TC\boldsymbol{\beta}}{\partial C} - \frac{\partial \boldsymbol{\beta}^TC^T\boldsymbol{\beta}}{\partial C} + \frac{\partial \boldsymbol{\beta}^TC^TC\boldsymbol{\beta}}{\partial C} \\ &= 2\Sigma^{-1}C - 2\boldsymbol{\beta}\boldsymbol{\beta}^T + 2C\boldsymbol{\beta}\boldsymbol{\beta}^T\end{aligned} \tag{I.3}$$

and in turn, any optimal solution in (2.4) is the solution of the Sylvester equation (in C) $\Sigma^{-1}C + C\big(\boldsymbol{\beta}\boldsymbol{\beta}^T\big) = \boldsymbol{\beta}\boldsymbol{\beta}^T$. The latter equation has a unique solution if and only if Σ^{-1} and $-\boldsymbol{\beta}^T\boldsymbol{\beta}$ have no common eigenvalues, which is true since all eigenvalues of Σ^{-1} are positive (as $\Sigma^{-1} \succ 0$) and all eigenvalues of $-\boldsymbol{\beta}^T\boldsymbol{\beta}$ are non-positive (as $\boldsymbol{\beta}^T\boldsymbol{\beta} \succeq 0$). This concludes the proof of Part iii).

Part iv) is now justified. The non-strict variant of (2.5), namely, $M_3^* \leq M_2^* \leq M_1^* \leq M_0^*$, is clear since the feasibility set obtaining M_s^* is a subset of the feasibility set obtaining M_{s+1}^* for all $s \in \{0,1,2\}$. Clearly, $M_1^* < M_0^*$ since $M_0^* = \sigma^2\sum_{k=0}^p \sigma_k > 0$ as $\sigma_k > 0$ is proved in Part ii). It is not difficult to show that $M_3^* < M_2^*$ if and only if C^* is diagonal.

It only remains to find the necessary and sufficient conditions under which $M_2^* < M_1^*$ is true. By taking $u_k = \beta_k^2$ and $v_k = \sigma^2\sigma_k$ for all $0 \leq k \leq p$ in Lemma 1, one may get that

$$M_2^* = \sum_{k=0}^p \frac{\beta_k^2\sigma^2\sigma_k}{\beta_k^2+\sigma^2\sigma_k} \leq \frac{\left(\sum_{k=0}^p \beta_k^2\right)\left(\sum_{k=0}^p \sigma^2\sigma_k\right)}{\sum_{k=0}^p \beta_k^2 + \sum_{k=0}^p \sigma^2\sigma_k} = M_1^*,$$

since $M_0^* = \sigma^2 \operatorname{Tr}\left(\Sigma^{-1}\right) = \sigma^2\sum_{k=0}^p \sigma_k$; note that $\sigma_k > 0$ for all $0 \leq k \leq p$ is proved in Part ii). The above inequality becomes an identity if and only if $\frac{\beta_k^2}{\sigma_k} = \frac{\beta_0^2}{\sigma_0}$ for all $1 \leq k \leq p$, which is a direct consequence of Lemma 1. The proof is now complete.

I.2 Proof of Theorem 2

It is first noted that due to Assumption 2.1, we have that

$$\widehat{\boldsymbol{\beta}}^{OLS} \xrightarrow{p} \boldsymbol{\beta} \quad \text{and} \quad \widehat{\sigma^2} \xrightarrow{p} \sigma^2. \tag{I.4}$$

In addition, the fact that $\frac{1}{n}\Sigma \to \Sigma_0$ and $\left(\frac{1}{n}\Sigma\right)^{-1} = n\Sigma^{-1}$, we get that

$$\Sigma^{-1} = o(n)J_{p+1} \quad \text{and} \quad \mathrm{Tr}(\Sigma^{-1}) = o(n), \tag{I.5}$$

where J_{p+1} an $p+1$ dimensional square matrix of ones. Note that we have used in the latter that if the convergence of a sequence of matrices, $A_n \to A$, implies the convergence of their inverse (assuming that inverses exist), i.e. $A_n^{-1} \to A^{-1}$; this is ensured by the fact that $A_n^{-1} = adj(A_n)/\det(A_n)$ and $A^{-1} = adj(A)/\det(A)$, where $adj(A_n)$(and $adj(A)$) is the adjugate of A_n(and A), and the obvious convergences $adj(A_n) \to adj(A)$ and $\det(A_n) \to \det(A)$. Therefore, the continuous mapping property of the convergence in probability, (I.4) and (I.5) imply that $\widehat{a^*} \xrightarrow{p} 1$ and $a^* \xrightarrow{p} 1$ (though the latter convergence is a deterministic convergence), and in turn, (2.9a) yields due to the continuous mapping property of the convergence in probability and (I.4). The proof of (2.9b) is quite similar, and one can find that $\widehat{b_k^*} \xrightarrow{p} 1$ and $b_k^* \xrightarrow{p} 1$ for all $0 \le k \le p$ by recalling that $0 < \sigma_k < \mathrm{Tr}(\Sigma^{-1}) = o(n)$, which implies that $\sigma_k = o(n)$ for all $0 \le k \le p$.

The proof of the first claim in (2.9c) follows from (2.9a) and the uniform integrability of $\widehat{a^*} - a^*$ which is implied by the fact that $\widehat{a^*} - a^*$ is uniformly bounded as $|\widehat{a^*}| \le 1$ and $|a^*| \le 1$ are true. Finally, the second claim in (2.9c) can be shown in a similar manner.

The first claim in (2.10) holds due to the uniformly integrability of $\widehat{\boldsymbol{\beta}}^{OLS}$ and the fact that $\widehat{a^*} - a^*$ is uniformly bounded. The uniform integrability of $\widehat{\boldsymbol{\beta}}^{OLS}$ is discussed in (Afendras and Markatou, 2016) that relies on the same conditions as our Assumption 2.1 and asymptotic covariance condition

$$\frac{1}{n}\Sigma \to \boldsymbol{\Sigma}_0 \text{ as } n \to \infty \text{ with } \boldsymbol{\Sigma}_0 \succ 0 \text{ for a fixed } p, \tag{I.6}$$

but one may use a much simple proof and use Theorem 4.5.9 in (Bogachev, 2007) with $G(t) = t^2$ and find that there exists $M > 0$ such that $\mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}_k^{OLS}\right)^2\right] < M$ for n sufficiency large; for any

$\epsilon > 0$, there exists $n_0 \geq 1$ such that the latter claim is concluded as follows:

$$\begin{aligned}
\mathbb{E}\big[\big(\widehat{\boldsymbol{\beta}}_k^{OLS}\big)^2\big] &= \beta_k^2 + \mathbb{V}\big(\widehat{\boldsymbol{\beta}}_k^{OLS}\big) \\
&\leq \beta_k^2 + \operatorname{Tr}\Big(\mathbb{V}\big(\widehat{\boldsymbol{\beta}}_k^{OLS}\big)\Big) \\
&= \beta_k^2 + \sigma^2 \operatorname{Tr}\big(\Sigma^{-1}\big) \\
&\leq \beta_k^2 + \sigma^2\left(\frac{1}{n}\operatorname{Tr}\big(\Sigma_0^{-1}\big) + \epsilon\right),
\end{aligned} \tag{I.7}$$

which is true for any $n > n_0$, where the latter inequality is a consequence of (I.6). The second claim in (2.10) follows through similar arguments.

It remained to show (2.11a) and (2.11b), and as before, we have shown only the first one. Note that

$$\begin{aligned}
&\mathbb{E}\left|\left|\left|\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right|\right|_2^2 - \left|\left|a^*\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right|\right|_2^2\right| \\
&\qquad \leq \sqrt{\mathbb{E}\left|\left|\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS} - a^*\widehat{\boldsymbol{\beta}}^{OLS}\right|\right|_2^2}\sqrt{\mathbb{E}\left|\left|\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS} + a^*\widehat{\boldsymbol{\beta}}^{OLS} - 2\boldsymbol{\beta}\right|\right|_2^2},
\end{aligned} \tag{I.8}$$

which is a consequence of the Cauchy-Schwarz inequality. Now, the first term in the right-hand side of (I.8) convergences to 0 due to (2.10), and second term in the right-hand side of (I.8) is bounded as $\mathbb{E}\left|\left|\widehat{a^*}\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right|\right|_2^2$ and $\mathbb{E}\left|\left|a^*\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\right|\right|_2^2$ are bounded, and in turn, the left-hand side of (I.8) convergences to 0. The proof is now complete.

I.3 Proof of Theorem 3

Similar to the proof in Appendix I.1 and by keeping (2.12) in mind, one may find that

$$\begin{aligned}
MSE\big(\widehat{\boldsymbol{\beta}}^{ind}(\rho)\big) &= MSE\big(\Sigma(\rho)\widehat{\boldsymbol{\beta}}^{OLS}\big) \\
&= \sigma^2\big(\rho^2 \operatorname{Tr}(\Sigma\widetilde{\Sigma}^{-2}) + 2\rho(1-\rho)\operatorname{Tr}(\widetilde{\Sigma}^{-1}) + (1-\rho)^2\operatorname{Tr}(\Sigma^{-1})\big) \\
&\qquad + \rho^2\boldsymbol{\beta}^T(\widetilde{\Sigma}^{-1}\Sigma - \mathbf{I}_p)^2\boldsymbol{\beta} \\
&= t_1(2\rho - \rho^2) + t_2(1-\rho)^2 + t_3\rho^2,
\end{aligned} \tag{I.9}$$

since $\operatorname{Tr}\big(\widetilde{\Sigma}^{-1}\big) = \operatorname{Tr}\big(\Sigma\widetilde{\Sigma}^{-2}\big) = \sum_{k=1}^p(\Sigma_{kk})^{-1}$ given that $\widetilde{\Sigma} = \operatorname{diag}(\Sigma)$, which in turn justifies (2.14) via some algebraic manipulations that we skip in this proof. Note that by Lemma 2, one find that $t_2 \geq t_1$, which in turn we have that $\rho^* \in [0,1]$ as $t_3 \geq 0$. The proof is now complete.

I.4 Proof of Theorem 4

Similar to the proof of Theorem 2, we only need to show that $\widehat{\rho^*} \xrightarrow{p} 0$ and $\rho^* \to 0$ due to the linearity of our estimator in order to justify (2.17). The latter is ensured by keeping the equivalent of (I.4) for $\widehat{\widehat{\boldsymbol{\beta}}}^{OLS}$ and (I.5) in mind, which implies that $\widetilde{\Sigma}^{-1} = o(n)J_p$ and $\mathrm{Tr}(\widetilde{\Sigma}^{-1}) = o(n)$, but also the fact that $\boldsymbol{\beta}^T\big(\widetilde{\Sigma}_0^{-1}\Sigma_0 - \mathbf{I}_p\big)^2\boldsymbol{\beta} \neq 0$. This justifies our claim in (2.17).

The remaining claims, namely (2.18) and (2.19), could be shown in the same manner as their counterparts in Theorem 2, and thus, we skip the details. We should note that the uniform integrability of $\widehat{\widehat{\boldsymbol{\beta}}}^{OLS}$ can be concluded in the same manner as (I.7). The proof is now complete.

I.5 Proof of Theorem 5

We first prove Part i). The interior point claim follows from the fact that

$$\lim_{||\boldsymbol{\beta}||_\infty \to \infty} \frac{\frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \mu\big(\mathbf{u}^T\boldsymbol{\beta}\big)^2}{||\boldsymbol{\beta}||_\infty^2} > 0,$$

since $\Sigma + \mu\mathbf{u}\mathbf{u}^T \succ 0$ and in turn the diagonal elements of $\Sigma + \mu\mathbf{u}\mathbf{u}^T$ are positive, where $||\boldsymbol{\beta}||_\infty$ is the max norm. Clearly, $\delta > 0$ since $\Sigma^{-1} \succ 0$ and $\mathbf{u} \neq \mathbf{0}$. Now, (2.21) yields that

$$\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u}) = \left(I_{p+1} - \frac{\mu}{1+\mu\delta}A\right)\Sigma^{-1}\mathbf{X}^T\mathbf{y} = \left(I_{p+1} - \frac{\mu}{1+\mu\delta}A\right)\widehat{\boldsymbol{\beta}}^{OLS} \tag{I.10}$$

for any feasible μ and $\mathbf{u}$, and in turn, we have that

$$MSE\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big) \;=\; \mathrm{Tr}\left(\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big)\right) + \Big(bias\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big)\Big)^2. \tag{I.11}$$

Equation (I.10) together with some algebraic manipulations and well-known standard properties of the OLS estimator imply that

$$\begin{aligned}\mathrm{Tr}\left(\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big)\right) &= \mathrm{Tr}\left(\left(I_{p+1} - \frac{\mu}{1+\mu\delta}A\right)^T \mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}^{OLS}\big)\left(I_{p+1} - \frac{\mu}{1+\mu\delta}A\right)\right) \\ &= \sigma^2\,\mathrm{Tr}\left(\left(I_{p+1} - \frac{\mu}{1+\mu\delta}A\right)^T \Sigma^{-1}\left(I_{p+1} - \frac{\mu}{1+\mu\delta}A\right)\right)\end{aligned} \tag{I.12}$$

and

$$\begin{aligned}\Big(bias\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big)\Big)^2 &= \Big|\Big|\mathbb{E}\big[(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big]-\boldsymbol{\beta}\Big|\Big|_2^2 \\ &= \Big|\Big|\left(I_{p+1}-\frac{\mu}{1+\mu\delta}A\right)\boldsymbol{\beta}-\boldsymbol{\beta}\Big|\Big|_2^2 \\ &= \left(\frac{\mu}{1+\mu\delta}\right)^2\boldsymbol{\beta}^TA^TA\boldsymbol{\beta}.\end{aligned} \tag{I.13}$$

Putting together (I.11)–(I.13), we get (2.22), which concludes the proof of part i).

Part ii) is a clear implication of the Karush-Kuhn-Tucker conditions, the fact that the objective function in (2.23) is strictly convex and strong duality holds which is a consequence of the fact that the Slater's condition holds in (2.23) as those constraints are linear. The proof is now complete.

I.6 Proof of Theorem 6

We first prove part i). Theorem 5 i) and the fact that A and Σ^{-1} are symmetric matrices, one may use some basic matrix trace properties to get that

$$\begin{aligned}&\frac{\partial\,\mathrm{Tr}\left(\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big)\right)}{\partial\mu} \\ &= \sigma^2\frac{\partial}{\partial\mu}\mathrm{Tr}\left(\left(I_{p+1}-\frac{\mu}{1+\mu a_1(\mathbf{u})}\Sigma^{-1}\mathbf{u}\mathbf{u}^T\right)^T\Sigma^{-1}\left(I_{p+1}-\frac{\mu}{1+\mu a_1(\mathbf{u})}\Sigma^{-1}\mathbf{u}\mathbf{u}^T\right)\right) \\ &= \sigma^2\frac{\partial}{\partial\mu}\mathrm{Tr}\left(\left(-\frac{2\mu}{1+\mu a_1(\mathbf{u})}\Sigma^{-2}\mathbf{u}\mathbf{u}^T+\left(\frac{\mu}{1+\mu a_1(\mathbf{u})}\right)^2\mathbf{u}\mathbf{u}^T\Sigma^{-3}\mathbf{u}\mathbf{u}^T\right)\right) \\ &= \frac{2\sigma^2}{(1+\mu a_1(\mathbf{u}))^2}\left(-\mathrm{Tr}\left(\Sigma^{-2}\mathbf{u}\mathbf{u}^T\right)+\frac{\mu}{1+\mu a_1(\mathbf{u})}\mathrm{Tr}\left(\mathbf{u}\mathbf{u}^T\Sigma^{-3}\mathbf{u}\mathbf{u}^T\right)\right)\end{aligned} \tag{I.14}$$

and

$$\begin{aligned}\frac{\partial\Big(bias\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big)\Big)^2}{\partial\mu} &= \frac{\partial}{\partial\mu}\left(\frac{\mu}{1+\mu a_1(\mathbf{u})}\right)^2\boldsymbol{\beta}^T\left(\Sigma^{-1}\mathbf{u}\mathbf{u}^T\right)^T\left(\Sigma^{-1}\mathbf{u}\mathbf{u}^T\right)\boldsymbol{\beta} \\ &= \frac{2\mu}{(1+\mu a_1(\mathbf{u}))^3}\boldsymbol{\beta}^T\left(\Sigma^{-1}\mathbf{u}\mathbf{u}^T\right)^T\left(\Sigma^{-1}\mathbf{u}\mathbf{u}^T\right)\boldsymbol{\beta} \\ &= \frac{2\mu}{(1+\mu a_1(\mathbf{u}))^3}a_3(\mathbf{u})\big(\boldsymbol{\beta}^T\mathbf{u}\big)^2.\end{aligned} \tag{I.15}$$

Now,

$$\text{Tr}\left(\mathbf{u}\mathbf{u}^T\Sigma^{-3}\mathbf{u}\mathbf{u}^T\right) = a_0(\mathbf{u})a_3(\mathbf{u}) > 0 \quad \text{and} \quad \text{Tr}\left(\Sigma^{-2}\mathbf{u}\mathbf{u}^T\right) = a_2(\mathbf{u}) > 0. \tag{I.16}$$

Thus, (I.11) and (I.14)–(I.16) imply that

$$\frac{\partial MSE\big(\widehat{\boldsymbol{\beta}}^{SR}(\mu;\mathbf{u})\big)}{\partial\mu} = \frac{2}{(1+\mu a_1(\mathbf{u}))^2}\Big(-\sigma^2 a_2(\mathbf{u}) + \frac{\mu}{1+\mu a_1(\mathbf{u})}\big(\Delta(\mathbf{u})+\sigma^2 a_1(\mathbf{u})a_2(\mathbf{u})\big)\Big). \tag{I.17}$$

Since $\Delta(\mathbf{u})+\sigma^2 a_1(\mathbf{u})a_2(\mathbf{u}) > 0$, (I.16) and (I.17) imply (2.24)–(2.26).

Note that $\mu^{**}(v\mathbf{1}) = v^{-2}\mu^{**}(\mathbf{1})$ for all $v \in (0,\infty)$, and together with (2.22), one may conclude (2.27). This concludes the proof for part i).

We now argue part ii) for which $\mathbf{u} \in \Re_+^{p+1} \setminus \{\mathbf{0}\}$ is assumed. It is first shown that

$$\frac{a_1(\mathbf{u})}{a_0(\mathbf{u})} \le \frac{a_3(\mathbf{u})}{a_2(\mathbf{u})} \quad \text{for any } \mathbf{u} \in \Re_+^{p+1} \setminus \{\mathbf{0}\}. \tag{I.18}$$

Proposition 1 in (Coppersmith et al., 1997) tells us that

$$\frac{\mathbf{u}^T Z\mathbf{u}}{\mathbf{u}^T\mathbf{u}} \le \frac{\mathbf{u}^T Z^3\mathbf{u}}{\mathbf{u}^T Z^2\mathbf{u}} \quad \text{for any } \mathbf{u} \in \Re_+^{p+1} \setminus \{\mathbf{0}\} \text{ and } Z \succeq 0 \text{ such that } Z\mathbf{u} \neq \mathbf{0}. \tag{I.19}$$

By taking $Z = \Sigma^{-1}$ in (I.19), and noting that $Z\mathbf{u} \neq \mathbf{0}$ since otherwise $\mathbf{u}^T Z\mathbf{u} = 0$, which would be impossible as $\Sigma^{-1} \succ 0$ and $\mathbf{u} \neq \mathbf{0}$, and in turn, (I.18) holds. Consequently, $a_3(\mathbf{u})\big(\boldsymbol{\beta}^T\mathbf{u}\big)^2 \le \Delta(\mathbf{u})$ with equality if and only if (I.18) becomes a strict inequality or $\boldsymbol{\beta}^T\mathbf{u} = \mathbf{0}$. By applying Proposition 1 in (Coppersmith et al., 1997), one may find that (I.18) becomes an equality if and only if $\mathbf{u}$ is an eigenvector of Σ^{-1}, which is equivalent to $\mathbf{u}$ being an eigenvector of $\Sigma(0)$. This concludes that $\Delta(\mathbf{u}) > 0$ if and if and only (2.28) holds, and thus, $\mu^{**}(\mathbf{u}) < \infty$ if and if and only (2.28) holds. The proof is now complete.

I.7 Proof of Theorem 7

We first prove Part i). The interior point claim can be proved as in the proof of Theorem 5 in Appendix I.5 since $\Sigma + \sum_{l\in\mathcal{L}} \mu_l \mathbf{u}_l\mathbf{u}_l^T \succ 0$ is clearly true since the Spectral Decomposition

Theorem implies that $\Sigma = \sum_{k=0}^{p} \lambda_l \mathbf{u}_l \mathbf{u}_l^T$ and $\Sigma^{-1} = \sum_{k=0}^{p} \lambda_l^{-1} \mathbf{u}_l \mathbf{u}_l^T$. Now, (2.30) follows from

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{GSR}(\boldsymbol{\mu}) &= \left(\Sigma + \sum_{l\in\mathcal{L}} \mu_l \mathbf{u}_l \mathbf{u}_l^T\right)^{-1} \mathbf{X}^T\mathbf{y} \\
&= \left(\Sigma + \sum_{l\in\mathcal{L}} \mu_l \mathbf{u}_l \mathbf{u}_l^T\right)^{-1} \Sigma\widehat{\boldsymbol{\beta}}^{OLS} \\
&= \left(\mathbf{I}_{p+1} - \sum_{l\in\mathcal{L}} \frac{\mu_l\lambda_l^{-1}}{1+\mu_l\lambda_l^{-1}} \mathbf{u}_l \mathbf{u}_l^T\right) \widehat{\boldsymbol{\beta}}^{OLS},
\end{aligned} \tag{I.20}$$

where the latter is an implication of the Sherman-Morrison identity that could be proved by induction. We prove this result by considering the cases in which $\mathcal{L} = \{1\}$ and $\mathcal{L} = \{1, 2\}$, since the general case follows the same idea. First, the Sherman-Morrison identity yields the case $\mathcal{L} = \{1\}$ as follows

$$\left(\Sigma + \mu_1 \mathbf{u}_1 \mathbf{u}_1^T\right)^{-1} = \Sigma^{-1} - \frac{\mu_1 \Sigma^{-1} \mathbf{u}_1 \mathbf{u}_1^T \Sigma^{-1}}{1 + \mu_1 \mathbf{u}_1^T \Sigma^{-1} \mathbf{u}_1} = \Sigma^{-1} - \frac{\mu_1 \lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}} \mathbf{u}_1 \mathbf{u}_1^T \Sigma^{-1}, \tag{I.21}$$

where the latter is a consequence of the fact that $\mathbf{u}_l$'s are orthonormal vectors; specifically,

$$\mathbf{u}_1^T \Sigma^{-1} \mathbf{u}_1 = \sum_{k=0}^{p} \lambda_k^{-1} \mathbf{u}_1^T \mathbf{u}_k \mathbf{u}_k^T \mathbf{u}_1 = \lambda_1^{-1} \mathbf{u}_1^T \mathbf{u}_1 \mathbf{u}_1^T \mathbf{u}_1 = \lambda_1^{-1} \tag{I.22}$$

and

$$\Sigma^{-1} \mathbf{u}_1 \mathbf{u}_1^T = \sum_{k=0}^{p} \lambda_k^{-1} \mathbf{u}_k \mathbf{u}_k^T \mathbf{u}_1 \mathbf{u}_1^T = \lambda_1^{-1} \mathbf{u}_1 \mathbf{u}_1^T \mathbf{u}_1 \mathbf{u}_1^T = \lambda_1^{-1} \mathbf{u}_1 \mathbf{u}_1^T. \tag{I.23}$$

Second, the Sherman-Morrison identity yields the case $\mathcal{L} = \{1, 2\}$ as follows

$$\begin{aligned}
\left(\Sigma + \mu_1 \mathbf{u}_1 \mathbf{u}_1^T + \mu_1 \mathbf{u}_2 \mathbf{u}_2^T\right)^{-1} &= \left(\Sigma + \mu_1 \mathbf{u}_1 \mathbf{u}_1^T\right)^{-1} - \frac{\mu_2 \left(\Sigma + \mu_1 \mathbf{u}_1 \mathbf{u}_1^T\right)^{-1} \mathbf{u}_2 \mathbf{u}_2^T \left(\Sigma + \mu_1 \mathbf{u}_1 \mathbf{u}_1^T\right)^{-1}}{1 + \mu_2 \mathbf{u}_2^T \left(\Sigma + \mu_1 \mathbf{u}_1 \mathbf{u}_1^T\right)^{-1} \mathbf{u}_2} \\
&= \Sigma^{-1} - \frac{\mu_1 \lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}} \mathbf{u}_1 \mathbf{u}_1^T \Sigma^{-1} - \frac{\mu_2 \lambda_2^{-1}}{1+\mu_2\lambda_2^{-1}} \mathbf{u}_2 \mathbf{u}_2^T \Sigma^{-1},
\end{aligned} \tag{I.24}$$

which are consequences of the fact that $\mathbf{u}_l$'s are orthonormal vectors and (I.21). Specifically,

$$\begin{aligned}
\mathbf{u}_2^T \left(\Sigma + \mu_1 \mathbf{u}_1 \mathbf{u}_1^T\right)^{-1} \mathbf{u}_2 &= \mathbf{u}_2^T \left(\Sigma^{-1} - \frac{\mu_1 \lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}} \mathbf{u}_1 \mathbf{u}_1^T \Sigma^{-1}\right) \mathbf{u}_2 \\
&= \mathbf{u}_2^T \Sigma^{-1} \mathbf{u}_2 - \frac{\mu_1 \lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}} \mathbf{u}_2^T \mathbf{u}_1 \mathbf{u}_1^T \Sigma^{-1} \mathbf{u}_2,
\end{aligned}$$

$$= \quad \lambda_2^{-1} + 0$$

which is due to (I.21) and (I.22); further,

$$\begin{aligned}
&\left(\Sigma+\mu_1\mathbf{u}_1\mathbf{u}_1^T\right)^{-1}\mathbf{u}_2\mathbf{u}_2^T\left(\Sigma+\mu_1\mathbf{u}_1\mathbf{u}_1^T\right)^{-1} \\
&\quad = \left(\Sigma^{-1} - \frac{\mu_1\lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}}\mathbf{u}_1\mathbf{u}_1^T\Sigma^{-1}\right)\mathbf{u}_2\mathbf{u}_2^T\left(\Sigma^{-1} - \frac{\mu_1\lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}}\mathbf{u}_1\mathbf{u}_1^T\Sigma^{-1}\right), \qquad \text{(I.25)} \\
&\quad := T_1 + T_2 - T_3 - T_4 \\
&\quad = \lambda_2^{-1}\mathbf{u}_2\mathbf{u}_2^T\Sigma^{-1} + 0 - 0 - 0,
\end{aligned}$$

which is due to (I.21), where

$$T_1 := \Sigma^{-1}\mathbf{u}_2\mathbf{u}_2^T\Sigma^{-1} = \lambda_2^{-1}\mathbf{u}_2\mathbf{u}_2^T\Sigma^{-1},$$

holds due to (I.23),

$$T_2 := \left(\frac{\mu_1\lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}}\right)^2\mathbf{u}_1\mathbf{u}_1^T\Sigma^{-1}\mathbf{u}_2\mathbf{u}_2^T\mathbf{u}_1\mathbf{u}_1^T\Sigma^{-1} = 0,$$

since $\mathbf{u}_1^T\Sigma^{-1}\mathbf{u}_2 = \lambda_2^{-1}\mathbf{u}_1^T\mathbf{u}_2 = 0$ given that $(\lambda_2^{-1}, \mathbf{u}_2)$ is the paired eigenvalue-eigenvector for Σ^{-1},

$$T_3 := \frac{\mu_1\lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}}\mathbf{u}_1\mathbf{u}_1^T\Sigma^{-1}\mathbf{u}_2\mathbf{u}_2^T\Sigma^{-1},$$

since $\mathbf{u}_1^T\Sigma^{-1}\mathbf{u}_2 = 0$, and

$$T_4 := \frac{\mu_1\lambda_1^{-1}}{1+\mu_1\lambda_1^{-1}}\Sigma^{-1}\mathbf{u}_2\mathbf{u}_2^T\mathbf{u}_1\mathbf{u}_1^T\Sigma^{-1},$$

since $\mathbf{u}_2^T\mathbf{u}_1 = 0$. This concludes (I.25), and in turn, (I.24), (I.20) and (2.30) are justified.

It only remains to show (2.31) for Part i). Similar derivations to those used to show (2.22), (I.20), the Spectral Decomposition Theorem for Σ^{-1} and the fact that $\mathbf{u}_l$'s are orthonormal vectors would help to find that

$$\begin{aligned}
MSE\left(\widehat{\boldsymbol{\beta}}^{GSR}(\boldsymbol{\mu})\right) = \sigma^2\,\mathrm{Tr}(\Sigma^{-1}) &+ \sum_{l\in\mathcal{L}}\left(\frac{\mu_l\lambda_l^{-1}}{1+\mu_l\lambda_l^{-1}}\right)^2\left(\sigma^2\lambda_l^{-1} + \left(\mathbf{u}_l^T\boldsymbol{\beta}\right)^2\right) \\
&-2\sum_{l\in\mathcal{L}}\left(\frac{\mu_l\lambda_l^{-1}}{1+\mu_l\lambda_l^{-1}}\right)\sigma^2\lambda_l^{-1}.
\end{aligned}$$

The above is a separable function and is minimized when

$$\frac{\mu_l^* \lambda_l^{-1}}{1+\mu_l^* \lambda_l^{-1}} = \frac{\sigma^2 \lambda_l^{-1}}{\sigma^2 \lambda_l^{-1} + \left(\mathbf{u}_l^T \boldsymbol{\beta}\right)^2}, \text{ i.e., } \mu_l^* = \sigma^2/(\mathbf{u}_l^T \boldsymbol{\beta})^2 \text{ for all } l \in \mathcal{L}.$$

Consequently,

$$\begin{aligned} MSE\big(\widehat{\boldsymbol{\beta}}^{GSR}(\boldsymbol{\mu}^*)\big) &= \sigma^2 \left(\mathrm{Tr}(\Sigma^{-1}) - \sum_{l \in \mathcal{L}} \frac{\sigma^2 \lambda_l^{-2}}{\sigma^2 \lambda_l^{-1} + \left(\mathbf{u}_l^T \boldsymbol{\beta}\right)^2} \right) \\ &= \sigma^2 \sum_{l \notin \mathcal{L}} \lambda_l^{-1} + \sigma^2 \sum_{l \in \mathcal{L}} \lambda_l^{-1} \frac{\left(\mathbf{u}_l^T \boldsymbol{\beta}\right)^2}{\sigma^2 \lambda_l^{-1} + \left(\mathbf{u}_l^T \boldsymbol{\beta}\right)^2}, \end{aligned}$$

which concludes (2.31) and the proof for Part i).

Part ii) follows in a similar manner to the proof of Theorem 5 ii) in Appendix I.5, and thus, its proof is then omitted. The proof is now complete.

I.8 Proof of Theorem 8

We first prove (2.34a). Note first that $\frac{1}{n}\Sigma \to \Sigma_0$ implies

$$n^l a_l(\mathbf{1}) \to \tilde{a}_l(\mathbf{1}) := \mathbf{1}^T \Sigma_0^{-l} \mathbf{1} \quad \text{as } n \to \infty \text{ for all } l \in \mathbb{Z}. \tag{I.26}$$

Since $\mu^{**}(\mathbf{1}) < \infty$ due to Theorem 6 ii) as $\mathbf{1}$ is not an eigenvector of Σ, one may get from (I.26) that

$$\frac{\mu^{**}(\mathbf{1})}{1+\mu^{**}(\mathbf{1}) a_1(\mathbf{1})} \Sigma^{-1} \to \frac{\sigma^2 \tilde{a}_2(\mathbf{1})}{\sigma^2 \tilde{a}_0(\mathbf{1}) \tilde{a}_3(\mathbf{1}) + \tilde{a}_3(\mathbf{1}) \left(\mathbf{1}^T \boldsymbol{\beta}\right)^2} \Sigma_0^{-1} \quad \text{as } n \to \infty. \tag{I.27}$$

Similarly, (I.4) and (I.26) imply that

$$\frac{\widehat{\mu^{**}(\mathbf{1})}}{1+\widehat{\mu^{**}(\mathbf{1})} a_1(\mathbf{1})} \Sigma^{-1} \xrightarrow{p} \frac{\sigma^2 \tilde{a}_2(\mathbf{1})}{\sigma^2 \tilde{a}_0(\mathbf{1}) \tilde{a}_3(\mathbf{1}) + \tilde{a}_3(\mathbf{1}) \left(\mathbf{1}^T \boldsymbol{\beta}\right)^2} \Sigma_0^{-1}. \tag{I.28}$$

Thus, (I.4), (I.27) and (I.28) yield (2.34a).

We now prove (2.34b). A key ingredient of this proof is to note that the eigenvalues of Σ converge to the corresponding eigenvalues of Σ_0, since the eigenvalues are the roots of a polynomial which converge to the limit polynomial due to the Implicit Function Theorem. Thus, the eigenvalues

of Σ_0 are distinct as well. The same convergence property (up to a proportionality constant of ± 1) holds for the eigenvectors due to the Davis-Kahan Theorem as the eigenvalues of Σ and Σ_0 are distinct; e.g., see Theorem 1 of (Yu et al., 2015). This means that

$$\frac{1}{n}\lambda_k \to \lambda_k^{(0)} \quad \text{and} \quad \left(\mathbf{u}_k^T\boldsymbol{\beta}\right)^2 \to \left(\mathbf{u}_k^{(0)T}\boldsymbol{\beta}\right)^2 \quad \text{as } n \to \infty \text{ for any } 0 \le k \le p, \tag{I.29}$$

where $\left(\lambda_k^{(0)}, \mathbf{u}_k^{(0)}\right)$ is the k^{th} paired eigenvalue-eigenvector of Σ_0. Therefore,

$$\frac{\sigma^2\lambda_k^{-1}}{\sigma^2\lambda_k^{-1} + \left(\mathbf{u}_k^T\boldsymbol{\beta}\right)^2} \to 0 \quad \text{as } n \to \infty \text{ for any } 0 \le k \le p.$$

The latter and (I.4) imply that

$$\widehat{\boldsymbol{\beta}}^{GSR}(\boldsymbol{\mu}^*) \xrightarrow{p} \boldsymbol{\beta} \tag{I.30}$$

Similarly, one may show that

$$\widehat{\boldsymbol{\beta}}^{GSR}(\widehat{\boldsymbol{\mu}^*}) \xrightarrow{p} \boldsymbol{\beta}. \tag{I.31}$$

Finally, (I.30) and (I.31) imply (2.34b). We now prove part i), and we only show (2.35), since (2.36) can be proved in the same manner as its counterpart in Theorem 2, and thus, we skip the details. Clearly,

$$\left|\frac{\widehat{\mu^{**}(\mathbf{1})}a_1(\mathbf{1})}{1+\widehat{\mu^{**}(\mathbf{1})}a_1(\mathbf{1})}\right| \le 1 \quad \text{a.s. and} \quad \left|\frac{\mu^{**}(\mathbf{1})\,a_1(\mathbf{1})}{1+\mu^{**}(\mathbf{1})\,a_1(\mathbf{1})}\right| \le 1$$

and in turn,

$$\left(\frac{\widehat{\mu^{**}(\mathbf{1})}}{1+\widehat{\mu^{**}(\mathbf{1})}a_1(\mathbf{1})} - \frac{\mu^{**}(\mathbf{1})}{1+\mu^{**}(\mathbf{1})\,a_1(\mathbf{1})}\right)\Sigma^{-1} \quad \text{is uniformly bounded.}$$

Similar to the proof of Theorem 2, the latter implies (2.36), which concludes part i).

We now prove part ii) for which we only give the main steps. It is not difficult to conclude that

$$\sum_{l\in\mathcal{L}} \frac{\widehat{\mu_l^*}\lambda_l^{-1}}{1+\widehat{\mu_l^*}\lambda_l^{-1}}\mathbf{u}_l\mathbf{u}_l^T \quad \text{and} \quad \sum_{l\in\mathcal{L}} \frac{\mu_l\lambda_l^{-1}}{1+\mu_l\lambda_l^{-1}}\mathbf{u}_l\mathbf{u}_l^T$$

are uniformly bounded by keeping in mind that the eigenvectors are unitary vectors. The

remaining steps for proving part ii) are exactly the same as those used in part ii), and we skip the details. This concludes the proof of part ii), and the proof is now complete.

I.9 Proof of Proposition 1

We first derive the MSE of $\widehat{\boldsymbol{\beta}}^{SRR}(\rho)$, where the SRR estimator defined in (2.41). As before,

$$MSE\big(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big) \;=\; \mathrm{Tr}\left(\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big)\right) + \Big(bias\big(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big)\Big)^2. \tag{I.32}$$

One may find that

$$\begin{aligned}
\mathrm{Tr}\left(\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big)\right) &= \sigma^2\,\mathrm{Tr}\left(\left(\sum_{k=0}^{p}\frac{1}{(1-\rho)\lambda_k+\rho v}\mathbf{u}_k\mathbf{u}_k^T\mathbf{X}^T\right)^T\sum_{k=0}^{p}\frac{1}{(1-\rho)\lambda_k+\rho v}\mathbf{u}_k\mathbf{u}_k^T\mathbf{X}^T\right)\\
&= \sigma^2\sum_{k=0}^{p}\frac{1}{\big((1-\rho)\lambda_k+\rho v\big)^2}\,\mathrm{Tr}\left(\mathbf{X}\mathbf{u}_k\mathbf{u}_k^T\mathbf{X}^T\right)\\
&= \sigma^2\sum_{k=0}^{p}\frac{1}{\big((1-\rho)\lambda_k+\rho v\big)^2}\,\mathrm{Tr}\left(\mathbf{u}_k\mathbf{u}_k^T\sum_{l=1}^{p+1}\lambda_l\mathbf{u}_l\mathbf{u}_l^T\right)\\
&= \sigma^2\sum_{k=0}^{p}\frac{\lambda_k}{\big((1-\rho)\lambda_k+\rho v\big)^3}
\end{aligned} \tag{I.33}$$

and similar derivations yield that

$$\begin{aligned}
\Big(bias\big(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big)\Big)^2 &= \Big\|\mathbb{E}\big[(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big] - \boldsymbol{\beta}\Big\|_2^2\\
&= \left\|\sum_{k=0}^{p}\frac{1}{1-\rho)\lambda_k+\rho v}\mathbf{u}_k\mathbf{u}_k^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}\right\|_2^2\\
&= \left\|\sum_{k=0}^{p}\frac{\lambda_k}{(1-\rho)\lambda_k+\rho v}\mathbf{u}_k\mathbf{u}_k^T\boldsymbol{\beta}-\sum_{k=0}^{p}\mathbf{u}_k\mathbf{u}_k^T\boldsymbol{\beta}\right\|_2^2\\
&= \boldsymbol{\beta}^T\left(\sum_{k=0}^{p}\left(\frac{\lambda_k}{(1-\rho)\lambda_k+\rho v}-1\right)\mathbf{u}_k\mathbf{u}_k^T\right)^T\sum_{k=0}^{p}\left(\frac{\lambda_k}{(1-\rho)\lambda_k+\rho v}-1\right)\mathbf{u}_k\mathbf{u}_k^T\boldsymbol{\beta}\\
&= \sum_{k=0}^{p}\frac{\rho^2(\lambda_k-v)^2}{\big((1-\rho)\lambda_k+\rho v\big)^2}\big(\mathbf{u}_k^T\boldsymbol{\beta}\big)^2.
\end{aligned} \tag{I.34}$$

A bona fide estimator to $MSE\big(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\big)$ requires an estimator for σ^2 and $\big(\mathbf{u}_k^T\boldsymbol{\beta}\big)^2$. The plug-in

estimator for $\left(\mathbf{u}_k^T\boldsymbol{\beta}\right)^2$ is given by

$$\left(\mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{u}_k\right)^2=\left(\sum_{l=0}^{p}\frac{\mathbf{y}^T\mathbf{X}\mathbf{u}_l\mathbf{u}_l^T\mathbf{u}_k}{(1-\rho)\lambda_l+\rho v}\right)^2=\left(\frac{\mathbf{y}^T\mathbf{X}\mathbf{u}_k}{(1-\rho)\lambda_k+\rho v}\right)^2=\frac{v_k}{\left((1-\rho)\lambda_k+\rho v\right)^2}, \quad \text{(I.35)}$$

where $v_k := \mathbf{y}^T\mathbf{X}\mathbf{u}_k$ for all $0 \leq k \leq p$. The plug-in estimator for σ^2 could be similarly derived as follows:

$$\begin{aligned}&\frac{1}{n-p-1}\left(\mathbf{y}^T\mathbf{y}-2\mathbf{y}^T\mathbf{X}\left(\sum_{k=0}^{p}\frac{1}{1-\rho)\lambda_k+\rho v}\mathbf{u}_k\mathbf{u}_k^T\right)\mathbf{X}^T\mathbf{y}+\right.\\&\left.\mathbf{y}^T\mathbf{X}\left(\sum_{k=0}^{p}\frac{1}{1-\rho)\lambda_k+\rho v}\mathbf{u}_k\mathbf{u}_k^T\right)\sum_{k=0}^{p}\lambda_k\mathbf{u}_k\mathbf{u}_k^T\left(\sum_{k=0}^{p}\frac{1}{1-\rho)\lambda_k+\rho v}\mathbf{u}_k\mathbf{u}_k^T\right)\mathbf{X}^T\mathbf{y}\right)\\&=\frac{1}{n-p-1}\left(\mathbf{y}^T\mathbf{y}-2\sum_{k=0}^{p}\frac{v_k}{1-\rho)\lambda_k+\rho v}+\sum_{k=0}^{p}\frac{\lambda_k v_k}{\left((1-\rho)\lambda_k+\rho v\right)^2}\right)\end{aligned} \quad \text{(I.36)}$$

Putting together (I.32)–(I.36), we get that the final formula for $\widehat{MSE\left(\widehat{\boldsymbol{\beta}}^{SRR}(\rho)\right)}$ claimed through $H(\rho)$ that is given in Proposition 1. The proof is now complete.

I.10 Ancillary Results

Lemma 1. *Let $\boldsymbol{u}, \boldsymbol{v} \in \Re^m$ such that $u_k \geq 0$ and $v_k > 0$ for all $1 \leq k \leq m$. Then,*

$$\sum_{k=1}^{m}\frac{u_k v_k}{u_k+v_k} \leq \frac{\left(\boldsymbol{1}^T\boldsymbol{u}\right)\left(\boldsymbol{1}^T\boldsymbol{v}\right)}{\boldsymbol{1}^T\boldsymbol{u}+\boldsymbol{1}^T\boldsymbol{v}}, \quad \text{(I.37)}$$

where the above becomes an identity if and only if $\boldsymbol{u} = M\boldsymbol{v}$ for a given $M \geq 0$.

Proof. We first show that for any $\mathbf{p}, \mathbf{q} \in \Re^m$ such that $p_k \geq 0$ and $q_k > 0$ for all $1 \leq k \leq m$ with $\mathbf{1}^T\mathbf{p} = \mathbf{1}^T\mathbf{q} = 1$, the following holds

$$\sum_{k=1}^{m}\frac{p_k q_k}{a p_k+(1-a)q_k} \leq 1 \quad \text{for any } 0 < a < 1, \quad \text{(I.38)}$$

and the above becomes an identity if and only if $\mathbf{p} = \mathbf{q}$. Note that

$$1-\sum_{k=1}^{m}\frac{p_k q_k}{a p_k+(1-a)q_k} := H_{\phi_a}(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{m} q_k\phi_a\left(\frac{p_k}{q_k}\right), \quad \text{(I.39)}$$

where $\phi_a(t) := -\frac{t}{1-a+at}+(t-1)(1-a)+1$ for all $t \in \Re_+$. By definition, $H_{\phi_a}(\mathbf{p},\mathbf{q})$ is the

$\phi-$divergence between the probability distributions induced by $\mathbf{p}$ and $\mathbf{q}$ through the ϕ_a divergence function. Then, if $\phi_a(\cdot)$ is convex on $\Re_+$ and strictly convex in a neighborhood of 1 (which both are true), then $H_{\phi_a}(\mathbf{p},\mathbf{q}) \geq 0$ for any $\mathbf{p},\mathbf{q}$, and $H_{\phi_a}(\mathbf{p},\mathbf{q}) = 0$ if and only if $\mathbf{p} = \mathbf{q}$; for details, see (Pardo, 2005). This concludes (I.38). By taking $p_k = u_k/\mathbf{1}^T\mathbf{u}$ and $q_k = v_k/\mathbf{1}^T\mathbf{v}$ for all $1 \leq k \leq m$, and $a = \mathbf{1}^T\mathbf{u}/(\mathbf{1}^T\mathbf{u} + \mathbf{1}^T\mathbf{v})$ in (I.38), one may easily recover (I.37) whenever $\mathbf{1}^T\mathbf{u} > 0$; the case when $\mathbf{1}^T\mathbf{u} = 0$, which is equivalent to having $\mathbf{u} = \mathbf{0}$, is trivial. This completes the proof. ■

Lemma 2. *Let $A \succ 0$ be a symmetric matrix of size r. Then,* $\mathrm{Tr}(A^{-1}) \geq \mathrm{Tr}\left(\big(\,\mathrm{diag}(A)\big)^{-1}\right)$.

Proof. Let $\lambda_1, \ldots, \lambda_r$ be the eigenvalues of A and $\{\mathbf{u}_l, 1 \leq l \leq r\}$ be its orthonormal eigenvectors. Spectral decomposition tells us that

$$A = \sum_{k=1}^{r} \lambda_k \mathbf{u}_k \mathbf{u}_k^T \quad \text{and} \quad A^{-1} = \sum_{k=1}^{r} \lambda_k^{-1} \mathbf{u}_k \mathbf{u}_k^T.$$

The latter implies that

$$\mathrm{Tr}\left(\big(\,\mathrm{diag}(A)\big)^{-1}\right) = \sum_{k=1}^{r} (A_{kk})^{-1} = \sum_{k=1}^{r} \left(\sum_{s=1}^{r} \lambda_s u_{sk}^2\right)^{-1} \leq \sum_{k=1}^{r}\sum_{s=1}^{r} \lambda_s^{-1} u_{sk}^2 = \mathrm{Tr}(A^{-1}),$$

where the inequality is due to the Cauchy-Schwarz inequality and the last identity is true as the eigenvectors are orthonormal vectors. This completes the proof. ■

II Simulation Study

A vast synthetic data analysis is provided in this section. We start with explaining DGP in Section II.1, while the numerical experiments are provided in Section II.2 and further improved in Section II.3 through cross-validation. We conclude this section by expanding the discussion in Section 2.4 where we introduce the SRR estimator to improve the estimation error when the covariates exhibit an ill-conditioned covariance matrix when even RR is a suitable estimator.

II.1 Data Generating Process

The DGP is now specified. *First,* covariates, $\{\mathbf{X}_i\}_{i=1}^n$, are *independent and identically distributed (i.i.d.)* random variates from the following three parent distributions:

1. *Multivariate Gaussian covariates with Toeplitz covariance matrix*, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi}(\rho))$, where

$$\boldsymbol{\Psi}_{st}(\rho) = \rho^{|s-t|} \quad \text{for all } 1 \leq s, t \leq p. \tag{II.1}$$

   Here, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$ is the mean vector, $\boldsymbol{\Psi}(\rho)$ is the covariance matrix, and ρ represents the correlation coefficient that controls the dependence between covariates.

2. *Multivariate Gaussian dependence with Binomial covariates and Toeplitz covariance matrix*: that is, we first generate $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \Psi(\rho))$, and then, each marginal Z_{ik} is transformed to be binomially distributed through the following transformation:

$$X_{ik} = F^{-1}(\Phi(Z_{ik})), \quad \text{for } 1 \leq k \leq p, \tag{II.2}$$

   where Φ is the *cumulative distribution function (CDF)* of $N(0,1)$, and F^{-1} is the inverse CDF of the binomial distribution with parameters $N_q = 2$ number of trials and $q_0 \in [0.01, 0.25]$ success probability. That is, $\mathbf{X}_i$ has the Gaussian copula extracted from $\mathcal{N}(\mathbf{0}, \Psi(\rho))$ and Binomial(N_q, q_0) marginals.

3. *Latent Space Features:* Covariates are generated from a low-rank structure, which is similar to a setting from (Hastie et al., 2022). Specifically,

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{E}, \tag{II.3}$$

   where $\mathbf{A}$ is an $n \times f$ matrix of factor loadings with entries drawn independently from $\mathcal{N}(0,1)$, and $\mathbf{Z}$ is an $f \times p$ matrix of latent factors with entries drawn independently from $\mathcal{N}(0,1)$. The random matrix $\mathbf{E}$ is an $n \times p$ matrix of independent Gaussian noise with variance $\sigma^2 = 10^{-6}$, i.e., $\mathcal{N}(0, 10^{-6})$. Since $f < p$, the term $\mathbf{AZ}$ is a low-rank component with at most rank f, and the small noise $\mathbf{E}$ ensures that $\mathbf{X}$ remains close to the low-rank structure while also allowing an invertible covariance matrix even though is ill-conditioned. Thus, the covariance structure of $\mathbf{X}$ conditioned on $\mathbf{A}$ is:

$$\mathrm{Cov}\big(\mathbf{X} \mid \mathbf{A}\big) = \mathbf{A}^T\mathbf{A} + \sigma^2\mathbf{I}_p. \tag{II.4}$$

Second, the dependent variable is generated according to two different sampling distributions:

1. *Gaussian*

$$Y|\mathbf{X} = \mathbf{x} \sim \mathcal{N}\big(\boldsymbol{\beta}^T\widetilde{\mathbf{x}}, \sigma^2\big), \tag{II.5}$$

where σ^2 represents the variance of Y.

2. *t-Distributed* with ν degrees of freedom that controls the tail heaviness:

$$Y|\mathbf{X} = \mathbf{x} \sim t_\nu\big(\boldsymbol{\beta}^T\widetilde{\mathbf{x}}\big). \tag{II.6}$$

Third, the "true" regression parameters $\boldsymbol{\beta}$ are chosen in two ways:

1. *Alternating Sign Specification*

$$\beta_k = (-1)^{k+1}\left\lceil\frac{k}{2}\right\rceil, \quad \text{for } 1 \le k \le p+1, \text{ where } \lceil x\rceil \text{ is the ceiling function.} \tag{II.7}$$

2. *Uniformly Distributed* with zero intercept

$$\beta_0 = 0, \quad \beta_k \sim U(0.01, 0.3), \quad \text{for } 1 \le k \le p. \tag{II.8}$$

Binomially distributed covariates are common in genome-wide association studies (GWAS). In a GWAS, each genetic variant is tested for association with a health-related trait via a regression model that typically includes covariates such as age and gender. The genetic variant covariate is a genotype score that takes on values 0, 1, and 2.

II.2 Data Analyses

We compare the following six estimators: i) *OLS estimator* $\widehat{\boldsymbol{\beta}}^{OLS}$ by using the *lm* package in **R**, ii) *St estimator* $\widehat{a}^*\widehat{\boldsymbol{\beta}}^{OLS}$ as in (2.2), iii) *DSh estimator* $\widehat{\text{diag}(\mathbf{b}^*)}\widehat{\boldsymbol{\beta}}^{OLS}$ as in (2.3); iv) *Sh estimator* $\widehat{C^*}\widehat{\boldsymbol{\beta}}^{OLS}$ as in (2.4); v) *SR estimator* $\widehat{\boldsymbol{\beta}}^{SR}(\widehat{\mu^{**}(\mathbf{1})}; \mathbf{1}, 0)$ as in (2.27), vi) *GSR estimator* $\widehat{\boldsymbol{\beta}}^{GSR}(\widehat{\boldsymbol{\mu}^*})$ as in (2.30). Note that estimators ii)–vi) are proposed earlier in this paper and are implemented in our new **R** package, *savvySh*[†]. This simulation study does not include results for LSh as it performs similarly to OLS when using centered data. However, our package *savvySh* includes LSh as well.

[†]Available at: https://github.com/Ziwei-ChenChen/savvySh

We conduct the *first* simulation study with $N = 250$ replications to compare the performance of OLS, St, DSh, Sh, SR, and GSR. The sample sizes are set to $n = 500$, $1,000$, and $2,500$, with the number of covariates p varying with n except when $p = 1$. Specifically, the ratios of p to n are chosen as 1, 5%, 10%, 25%, 50%, and 75%. The covariance matrix is as in (II.1) with the mean fixed at $\mu = 0$ and $\rho = -0.75$, -0.5, -0.25, 0, 0.25, 0.5, and 0.75. The dependent variable is generated as in (II.5) with $\sigma = 1$ and $\sigma = 5$, and the corresponding results are presented in Tables 3 and 4. Additionally, the dependent variable is generated as in (II.5) with $\nu = 50/24$ degrees of freedom so that the variance of the t-distribution matches that of the normal distribution when $\sigma = 5$; the corresponding results are shown in Table 5. In all settings, the "true" regression coefficients $\boldsymbol{\beta}$ are specified as in (II.7). Each estimator is assigned a count of one in the tables if it achieves the minimum L_2 distance, which measures the closeness of the estimated coefficients to the "true" values. Smaller L_2 values indicate better accuracy.

In the *second* simulation study, we compare five estimators: OLS, St, DSh, SR, and GSR, based on insights from the *first* study but also to avoid the high computational cost for Sh estimation. This study focuses on simulating $N = 250$ replications with sample sizes $n = 1,000$, $2,500$, and $5,000$. The number of covariates is fixed at $p = 1$, 2, and 5, while the covariance matrix for the covariates is as before. Covariates are generated using (II.2), which transforms multivariate Gaussian random variates into multivariate binomial random variates with $N_q = 2$ trials. The probability q_0 varies across the 250 replications, starting from 0.01 and increasing incrementally to 0.25, ensuring equal spacing between values. The dependent variable is generated as in (II.5) with $\sigma = 1$ and $\sigma = 5$, while the "true" regression coefficients $\boldsymbol{\beta}$ are specified in (II.8). Similar to the *first* simulation study, each estimator is assigned a count of one in Table 6 if it achieves the minimum L_2 distance.

II.3 Further improvement of St, DSh and GSR

We could improve some of the shrinkage estimators by better balancing the variance and bias of some of the newly introduced shrinkage estimators, and we consider only St, DSh and GSR in this section as only those three estimators managed to outperform OLS in Section II.2. That is, let $\gamma > 0$ be the variance/bias balance parameter within the MSE of our shrinkage estimators,

Table 3: **Best performance regression model**

Normal Distribution: $\sigma = 1$																			
Panel A: $n = 500$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	1							1	3	2	8	16	14	25	29	32	33	40	35	30	25
St	8							5	7	16	4	10	24	36	42	45	32	34	38	33	23
DSh	11							3	7	10	20	25	39	50	43	66	67	82	74	66	45
Sh	0							3	0	0	0	1	0	0	32	10	7	1	3	1	1
SR	215							176	173	174	180	139	111	53	36	15	8	8	7	51	78
GSR	15							62	60	48	38	59	62	86	68	82	103	85	93	69	78
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	2	8	7	9	10	13	14	29	21	20	18	14	16	28	6	3	7	9	7	10	14
St	4	10	8	7	10	18	35	37	29	20	20	29	21	21	8	13	9	9	7	16	34
DSh	6	14	16	19	36	58	75	22	30	53	49	43	48	56	7	8	23	23	53	38	63
Sh	4	0	0	0	2	2	1	15	7	7	2	3	6	2	3	6	5	1	2	1	6
SR	182	163	167	165	133	79	34	83	83	79	98	98	81	53	157	131	135	137	119	85	44
GSR	52	55	52	50	59	80	91	64	80	71	63	63	78	90	69	89	71	71	62	100	89
Panel B: $n = 1{,}000$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	1							25	33	38	42	42	45	24	23	36	44	46	40	38	20
St	6							45	48	43	38	36	37	27	44	42	24	39	40	34	18
DSh	5							62	55	74	74	68	64	44	44	56	69	76	75	49	63
Sh	0							22	15	4	4	3	4	2	22	13	5	3	1	2	0
SR	225							25	4	0	0	0	11	71	34	17	6	3	7	48	78
GSR	13							71	95	91	92	101	89	82	83	86	102	83	87	79	71
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	18	28	29	30	24	26	20	19	18	11	21	12	17	27	28	14	12	15	10	13	23
St	32	34	30	37	32	21	18	24	23	31	27	18	18	16	25	17	26	24	16	21	25
DSh	39	53	54	55	62	48	48	21	34	27	47	43	50	69	18	34	29	37	41	35	69
Sh	18	6	2	9	3	1	1	13	4	6	0	4	2	2	13	6	10	4	2	4	2
SR	83	42	38	36	62	87	76	91	94	88	98	102	96	58	78	81	88	98	86	79	35
GSR	60	87	97	83	67	67	87	82	77	87	57	71	67	78	88	98	85	72	95	98	96
Panel C: $n = 2{,}500$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	0							2	6	3	4	7	8	12	30	36	35	37	27	30	25
St	6							8	4	4	1	7	14	31	39	41	34	39	45	37	16
DSh	3							4	5	10	19	15	48	76	52	56	78	73	79	57	42
Sh	0							2	2	0	0	0	0	1	14	11	1	0	3	0	2
SR	231							181	191	184	187	159	108	44	37	11	3	1	4	56	76
GSR	10							53	42	49	39	62	72	86	78	95	99	100	92	70	89
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	3	6	4	6	13	11	13	18	20	12	16	18	12	17	7	3	4	9	10	12	6
St	9	7	6	3	10	17	20	19	23	21	20	15	12	17	10	10	6	10	12	9	10
DSh	19	21	17	29	30	42	73	37	30	43	40	44	42	53	13	16	15	30	32	40	80
Sh	3	2	0	3	0	1	1	4	8	12	2	2	1	1	4	2	3	0	1	2	1
SR	170	171	169	171	152	117	50	84	96	84	93	106	110	79	136	147	150	135	123	108	53
GSR	46	43	54	38	45	62	93	88	73	78	79	65	73	83	80	72	72	66	72	79	100

Notes: Performance comparisons across different estimators are shown for Gaussian dependent variable with $\sigma = 1$ and Gaussian covariates as in (II.5) and (II.1), respectively. Estimators are ranked in L_2 distance (from the "true" regression parameters) across $N = 250$ replications, and the "best" estimator is highlighted in red for various choices of n, p/n and ρ; ρ is not required for simple linear regression ($p = 1$).

Table 4: **Best performance regression model**

Normal Distribution: $\sigma = 5$																			
Panel A: $n = 500$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	1							0	1	1	9	18	17	29	2	32	45	62	51	56	32
St	26							0	13	21	17	24	50	61	32	80	54	56	59	56	47
DSh	4							0	4	21	23	33	41	39	1	28	47	54	63	55	45
Sh	1							0	3	1	1	1	0	0	19	28	11	7	2	4	1
SR	194							34	125	155	153	112	80	57	2	0	0	0	0	15	68
GSR	24							216	104	51	47	62	62	64	194	82	93	71	75	64	57
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	0	12	8	12	16	16	15	27	40	35	41	34	24	39	7	9	12	15	13	16	21
St	12	21	15	17	22	33	56	78	62	67	48	52	32	35	25	24	19	24	23	45	67
DSh	0	4	12	16	44	49	59	1	16	27	33	29	47	42	9	0	12	20	44	28	23
Sh	7	5	4	0	2	3	1	26	20	17	8	6	7	3	4	8	7	1	6	1	6
SR	166	160	160	154	108	63	41	64	42	33	44	62	72	62	135	125	121	110	95	75	46
GSR	65	48	51	51	58	86	78	54	70	71	76	67	68	69	70	84	79	80	69	85	87
Panel B: $n = 1{,}000$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	1							1	30	47	49	49	53	38	7	45	62	50	50	56	29
St	17							27	70	61	54	49	58	49	72	64	49	53	51	49	31
DSh	3							2	38	53	55	62	67	38	10	45	50	64	59	63	60
Sh	0							25	23	8	6	4	4	3	50	34	5	6	2	3	1
SR	204							0	0	0	0	0	0	52	2	0	0	0	0	5	68
GSR	25							195	89	81	86	86	68	70	109	62	84	77	88	74	61
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	20	34	41	46	35	34	41	44	42	37	39	25	29	32	36	32	25	26	16	23	27
St	101	66	63	67	58	34	22	76	63	52	47	46	24	28	60	36	42	38	29	31	40
DSh	1	35	43	53	67	61	46	1	20	27	37	40	41	42	0	11	15	26	31	21	38
Sh	55	15	11	8	7	4	2	29	13	14	5	9	5	3	17	19	13	3	4	8	6
SR	13	3	0	0	0	41	69	44	36	22	45	65	91	69	87	83	75	82	85	84	55
GSR	60	97	92	76	83	76	70	56	76	98	77	65	60	76	50	69	80	75	85	83	84
Panel C: $n = 2{,}500$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	0							0	8	4	5	10	11	16	36	41	45	50	41	45	39
St	13							10	9	7	4	13	25	40	83	60	46	43	54	59	27
DSh	2							4	8	16	27	28	62	67	23	50	67	59	71	72	43
Sh	1							2	0	0	0	1	0	1	59	23	2	5	2	0	2
SR	218							170	178	167	164	133	75	37	0	0	0	0	0	1	71
GSR	16							64	47	56	50	65	77	89	49	76	90	93	82	73	68
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	6	5	6	9	17	14	16	57	32	31	42	34	22	22	15	8	6	11	12	14	12
St	17	10	3	11	13	22	36	68	53	33	38	28	27	24	22	13	13	11	17	11	30
DSh	7	19	27	25	38	54	74	10	31	52	55	48	46	38	3	16	15	27	32	48	56
Sh	7	4	1	3	0	3	2	22	23	13	6	4	5	3	6	8	4	1	1	2	1
SR	181	157	153	150	122	84	47	47	30	20	14	60	93	88	147	129	125	114	92	86	54
GSR	32	55	60	52	60	73	75	46	81	101	95	76	57	75	57	76	87	86	96	89	97

Notes: Notes: Performance comparisons across different estimators are shown for Gaussian dependent variable with $\sigma = 5$ and Gaussian covariates as in (II.5) and (II.1), respectively. Estimators are ranked in L_2 distance (from the "true" regression parameters) across $N = 250$ replications, and the "best" estimator is highlighted in red for various choices of n, p/n and ρ; ρ is not required for simple linear regression ($p = 1$).

Table 5: **Best performance regression model**

***t*-distribution: $v = 50/24$**																			
Panel A: $n = 500$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	3							1	5	3	2	7	13	18	10	40	41	54	44	49	34
St	14							8	8	7	18	17	34	46	60	53	51	47	46	51	27
DSh	6							1	11	24	30	36	34	64	16	42	65	57	56	67	50
Sh	0							2	1	1	2	0	1	1	37	28	7	9	4	3	4
SR	203							104	154	153	158	136	100	44	10	0	0	0	1	19	63
GSR	24							134	71	62	40	54	68	77	117	87	86	83	99	61	72
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	7	6	12	6	12	13	12	32	29	33	36	31	22	20	14	12	8	12	13	13	12
St	29	17	12	14	22	44	58	64	34	44	37	26	30	25	19	11	12	15	16	31	50
DSh	3	16	26	17	40	55	53	7	21	26	40	23	37	55	3	7	18	30	34	50	49
Sh	5	5	0	1	2	1	4	18	12	10	11	3	4	4	6	1	4	6	2	3	4
SR	151	163	157	153	112	71	39	73	81	68	51	89	81	63	144	142	135	120	98	83	47
GSR	55	43	43	59	62	66	84	56	73	69	75	78	76	83	64	77	73	67	87	70	88
Panel B: $n = 1{,}000$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	1							13	27	36	41	49	48	27	27	49	42	45	44	38	29
St	13							65	50	39	51	47	47	23	74	48	44	36	37	46	33
DSh	4							21	56	53	74	71	62	57	26	62	64	72	75	76	51
Sh	0							35	27	3	2	5	5	2	45	12	11	2	3	2	3
SR	217							0	0	0	0	0	0	67	8	2	0	0	0	14	66
GSR	15							116	90	119	82	78	88	74	70	77	89	95	91	74	68
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	34	40	56	37	39	32	25	29	26	30	25	20	24	23	25	31	27	21	32	22	20
St	72	56	43	39	46	35	29	60	38	29	26	30	26	30	57	29	29	28	24	31	22
DSh	23	42	57	46	66	48	45	12	35	44	50	44	47	49	5	14	37	32	41	41	40
Sh	35	20	9	10	4	7	6	28	11	11	9	8	4	4	20	6	3	5	3	6	3
SR	37	10	6	7	18	58	67	71	70	55	58	75	91	71	80	95	97	80	76	70	63
GSR	49	82	79	111	77	70	78	50	70	81	82	73	58	73	63	75	57	84	74	80	102
Panel C: $n = 2{,}500$																			
p/n	**1**							**5%**							**10%**						
ρ	**NA**							**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	1							1	1	4	6	7	15	22	35	39	41	46	36	44	30
St	19							5	5	10	11	11	23	37	57	56	48	38	45	46	21
DSh	5							9	9	13	18	32	44	58	46	47	61	70	66	64	55
Sh	0							3	1	0	0	0	0	0	44	19	8	2	3	4	4
SR	205							174	189	175	164	142	93	38	2	0	0	0	0	5	70
GSR	20							58	45	48	51	58	75	95	66	89	92	94	100	87	70
p/n	**25%**							**50%**							**75%**						
ρ	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**	**-0.75**	**-0.5**	**-0.25**	**0**	**0.25**	**0.5**	**0.75**
OLS	8	6	9	6	9	10	13	34	26	25	21	26	29	15	7	10	7	11	7	10	14
St	15	12	7	17	7	21	37	50	47	30	37	27	15	14	15	11	11	14	15	21	29
DSh	7	15	21	25	39	52	64	26	50	50	59	51	37	48	9	14	22	35	36	54	57
Sh	7	3	0	4	1	4	2	12	10	9	8	7	2	3	7	4	5	2	4	4	2
SR	160	162	164	146	137	95	40	57	47	53	49	74	99	82	142	143	132	125	107	71	53
GSR	53	52	49	52	57	68	94	71	70	83	76	65	68	88	70	68	73	63	81	90	95

Notes: Notes: Performance comparisons across different estimators are shown for t-distribution dependent variable with $v = 50/24$ and Gaussian covariates as in (II.6) and (II.1), respectively. Estimators are ranked in L_2 distance (from the "true" regression parameters) across $N = 250$ replications, and the "best" estimator is highlighted in red for various choices of n, p/n and ρ; ρ is not required for simple linear regression ($p = 1$).

Table 6: **Best performance regression model**

$\sigma = 1$															
Panel A: $n = 1{,}000$															
p	1	2							5						
ρ	NA	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	-0.75	-0.5	-0.25	0	0.25	0.5	0.75
OLS	113	81	81	78	77	74	74	60	46	47	51	45	44	42	32
St	22	34	29	31	50	30	19	27	57	61	59	79	69	59	45
DSh	47	62	56	56	48	63	55	57	51	39	44	48	49	33	25
SR	56	37	37	39	28	25	14	12	16	27	24	18	12	5	8
GSR	14	36	47	46	47	58	88	94	80	76	72	60	74	111	140
Panel B: $n = 2{,}500$															
p	1	2							5						
ρ	NA	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	-0.75	-0.5	-0.25	0	0.25	0.5	0.75
OLS	95	89	79	71	76	74	64	63	67	62	68	53	48	64	31
St	21	25	27	25	32	20	15	11	45	42	46	63	47	48	42
DSh	80	65	66	68	70	73	72	69	49	54	57	60	60	39	26
SR	35	26	35	28	19	19	13	5	14	21	20	10	11	4	4
GSR	19	45	43	58	53	64	86	102	75	71	59	64	84	95	147
Panel C: $n = 5{,}000$															
p	1	2							5						
ρ	NA	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	-0.75	-0.5	-0.25	0	0.25	0.5	0.75
OLS	90	81	75	69	63	66	58	61	65	64	62	60	65	39	46
St	25	20	29	26	21	20	13	13	37	43	38	37	36	41	46
DSh	93	90	92	83	82	86	78	66	62	70	74	66	58	60	46
SR	28	16	13	18	12	11	8	3	5	9	7	3	1	2	0
GSR	14	43	41	54	72	67	93	107	81	64	69	84	90	108	112
$\sigma = 5$															
Panel A: $n = 1{,}000$															
p	1	2							5						
ρ	NA	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	-0.75	-0.5	-0.25	0	0.25	0.5	0.75
OLS	29	19	16	13	18	13	17	18	4	2	2	1	3	0	0
St	110	125	139	130	135	122	122	139	202	199	195	204	195	206	199
DSh	40	34	37	41	38	35	49	41	10	17	15	14	18	11	13
SR	47	24	25	23	26	21	10	6	0	4	4	0	2	0	0
GSR	24	48	33	43	33	59	52	46	34	28	34	31	32	33	38
Panel B: $n = 2{,}500$															
p	1	2							5						
ρ	NA	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	-0.75	-0.5	-0.25	0	0.25	0.5	0.75
OLS	58	36	30	29	33	35	27	21	7	13	9	7	7	11	2
St	68	83	93	106	87	86	84	92	167	149	147	159	155	154	159
DSh	49	52	39	36	58	45	54	53	22	35	33	31	27	22	12
SR	48	31	33	26	23	25	20	11	6	5	10	2	4	0	0
GSR	27	48	55	53	49	59	65	73	48	48	51	51	57	63	77
Panel C: $n = 5{,}000$															
p	1	2							5						
ρ	NA	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	-0.75	-0.5	-0.25	0	0.25	0.5	0.75
OLS	59	51	53	50	34	56	51	35	22	28	24	22	28	12	6
St	61	68	64	62	74	67	57	64	111	109	108	121	114	118	103
DSh	45	42	50	56	47	54	53	52	30	30	31	41	33	22	20
SR	53	39	27	30	21	28	20	12	13	21	22	4	10	3	3
GSR	32	50	56	52	74	45	69	87	74	62	65	62	65	95	118

Notes: Performance comparisons across different estimators are shown for Gaussian dependent variable with $\sigma = \{1, 5\}$ and multivariate binomial covariates as in (II.5) and (II.2), respectively. Estimators are ranked in L_2 distance (from the "true" regression parameters) across $N = 250$ replications, and the "best" estimator is highlighted in red for various choices of n, p/n and ρ; ρ is not required for simple linear regression ($p = 1$).

and define the slightly modified St and DSh bona fide estimators in (2.7) as follows:

$$\widehat{a^*(\gamma)}\widehat{\boldsymbol{\beta}}^{OLS} \quad \text{and} \quad \text{diag}(\widehat{\mathbf{b}^*(\gamma)})\widehat{\boldsymbol{\beta}}^{OLS}, \quad \text{where}$$

$$\widehat{a^*(\gamma)} := \frac{\gamma\big(\widehat{\boldsymbol{\beta}}^{OLS}\big)^T\widehat{\boldsymbol{\beta}}^{OLS}}{\gamma\big(\widehat{\boldsymbol{\beta}}^{OLS}\big)^T\widehat{\boldsymbol{\beta}}^{OLS} + \widehat{M_0^*}} \quad \text{and} \quad \widehat{b_k^*(\gamma)} = \frac{\gamma\Big(\widehat{\boldsymbol{\beta}}_k^{OLS}\Big)^2}{\gamma\Big(\widehat{\boldsymbol{\beta}}_k^{OLS}\Big)^2 + \widehat{\sigma^2}\sigma_k} \quad \text{with } 0 \leq k \leq p.$$

Similarly, the slightly changed GSR bona fide estimator in (2.33b) is as follows:

$$\widehat{\boldsymbol{\beta}}^{GSR}\left(\widehat{\boldsymbol{\mu}^*(\gamma)}\right) = \left(\mathbf{I}_{p+1} - \sum_{l\in\mathcal{L}} \frac{\widehat{\mu_l^*(\gamma)}\lambda_l^{-1}}{1+\widehat{\mu_l^*(\gamma)}\lambda_l^{-1}}\mathbf{u}_l\mathbf{u}_l^T\right)\widehat{\boldsymbol{\beta}}^{OLS}, \quad \text{where} \tag{II.9a}$$

$$\widehat{\mu_l^*(\gamma)} = \frac{\widehat{\sigma^2}}{\gamma\left(\mathbf{u}_l^T\widehat{\boldsymbol{\beta}}^{OLS}\right)^2} \quad \text{for all } 0 \leq l \leq p. \tag{II.9b}$$

The previous asymptotic theory clearly holds for any given $\gamma > 0$, and $\gamma = 1$ reverts to the previously defined St, DSh and GSR estimators. The parameter γ creates an uneven contribution to MSE, meaning that the bias is now weighted through γ, which may be beneficial in finite sample estimation to adjust our shrinkage estimators by performing cross-validation on γ. Our numerical experiments – not shown here but available upon request – show that GSR may benefit from such an adjustment, which has had a neutral effect on St and DSh.

II.4 Further Comparative Analysis between RR and SRR

Recall that we outlined the superior performance of SRR over RR and OLS at the end of Section 2.4, and we now provide more granular evidence in that respect by using synthetic data generated as explained in Section 2.4. Table 1 illustrates that the L_2-based estimation error of SRR is uniformly lower than that of RR and OLS, but we do not know by how much. In fact, the differences are very large between OLS and SRR, and significancy different when comparing RR to SRR; the pictorial representations in Figure 1 visually support these findings by plotting the L_2 errors of RR against those of SRR for two representative lower-dimensional cases though all other examples exhibit a similar pattern: i) $p/n = 5\%$, $f/p = 75\%$ and ii) $p/n = 10\%$, $f/p = 50\%$. These plots show how effective SRR is as compared to RR since many points are below the bisection line, which reinforces our findings in Table 1.

Further, we visualize in Figure 2 coefficient-specific errors using boxplots of L_1 error. Note that $\|x\|_p$ decreases as p increases, and thus, the L_1 error would capture any eventual outliers more

effectively than the L_2 errors. The dashed black horizontal line in Figure 2 indicates the lowest median L_1 error observed among all estimated coefficients from both RR and SRR, providing a clear benchmark point for comparing the RR and SRR performance. These boxplots illustrate that SRR consistently produces smaller median errors and exhibits less variability in estimation errors as compared to RR. In a nutshell, SRR is extremely effective in low-dimensional covariance subspace cases.

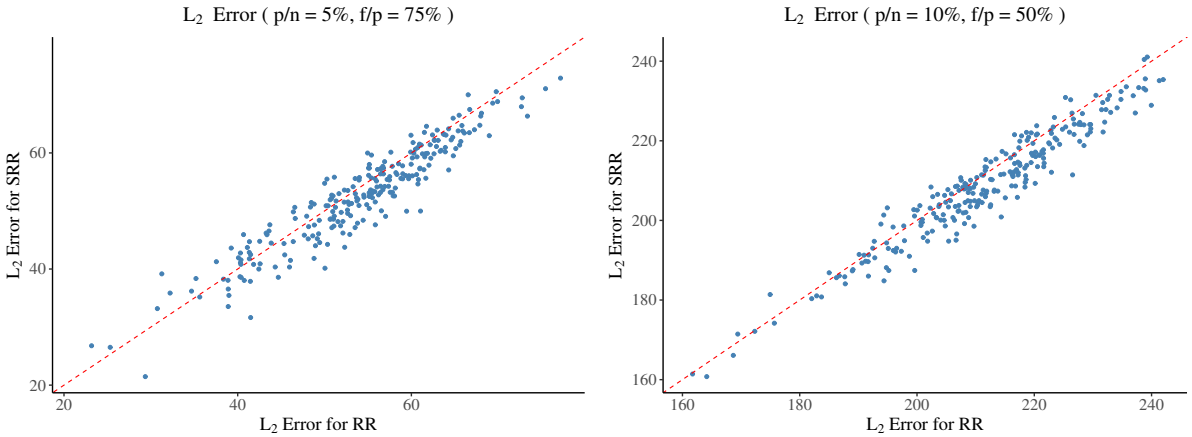


Figure 1: Scatter plots of L_2-distance (from the "true" regression parameters) for RR and SRR are provided by plotting $N = 250$ replications of samples of size $n = 1,000$ for two representative choices of p/n and f/p considered in Table 1. The red dashed line is a bisection indicating that SRR has a lower/higher L_2 error than RR for points below/above this line.

III Discussion about Eigenvalues and Kolmogorov Setting

This section provides extended information about the conclusions summarized in Section 2.5 as Results 9 and 10. *First,* we provide numerical evidence to support some interesting empirical evidence about the empirical eigenvalues. Such evidence would explain the behavior of MLR shrinkage estimators that directly or indirectly rely on the covariates' eigenvalues. Specifically, Section III.1 contains a series of pictorial representations confirming the statements in Result 9. *Second,* we illustrate at the beginning of Section III.2 how crucial the eigenvalues are in MLR estimation. *Third,* the second part of Section III.2 provides empirical evidence that St and GSR significantly improve the estimation error of OLS under the Kolmogorov setting. The last two sets of conclusions have been stated as Result 10.

The DGP in this section has two variants. One variant is as in Section II.1 where covariates are $\mathcal{N}(\mathbf{0}, \mathbf{\Psi}(\rho))$ distributed with a Toeplitz covariance matrix as in (II.1) and Gaussian depen-

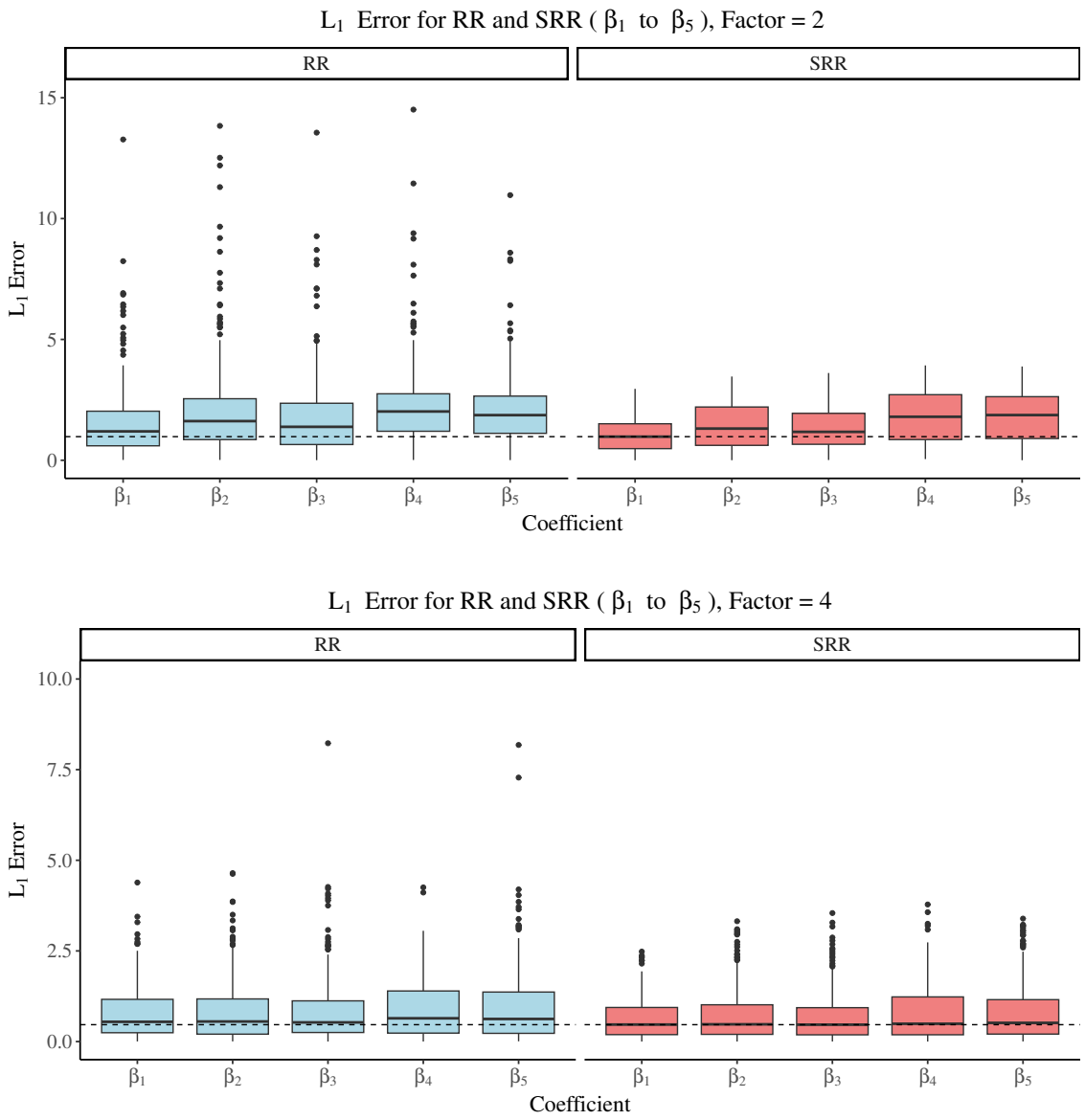


Figure 2: Boxplots of L_1 errors for the estimated coefficients of RR and SRR with $p = 5$, $n = 1{,}000$, and $N = 250$. The first-row plot shows results for $f = 2$, and the second-row plot for $f = 4$. The left panels show the distribution of L_1 errors for RR, and the right panels show the distribution for SRR. The dashed black horizontal line indicates the lowest median L_1 error among all coefficients for RR and SRR, which serves as a reference for comparison.

dent variable distributed as in (II.5). While this DGP controls the strength of dependence, we introduce a second DGP that is designed to control the eigenvalues of the "true" covariance matrix corresponding to the covariates. That is, Gaussian covariates are generated from $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}\,\mathrm{diag}(\boldsymbol{\lambda})\,\mathbf{Q}^T)$, while the response is Gaussian as in (II.5). Two data generating processes are considered for the covariance matrix, but both use the same "true" eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ with $\lambda_i \sim U(0, \lambda_{\max})$ for $i = 1, \ldots, p$. *First,* independent covariates are con-

sidered, i.e., $\mathbf{Q} = \mathbf{I}_p$. *Second,* dependent covariates are considered by preserving the eigenvalues $\boldsymbol{\lambda}$ and randomly generated eigenvectors, i.e. $\mathbf{Q}$ is a random orthogonal matrix. The latter is possible by taking $\widetilde{\mathbf{A}} = \frac{1}{2}\left(\mathbf{A} + \mathbf{A}^T\right)$, where $\mathbf{A}$ is a random square matrix , e.g., entries are i.i.d. $N(0,1)$; further, define $\widetilde{\widetilde{\mathbf{A}}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^T$ and its QR decomposition $\widetilde{\widetilde{\mathbf{A}}} = \mathbf{QR}$ gives $\mathbf{Q}$ as the required orthonormal matrix.

III.1 Bias of Empirical Eigenvalues

We generate multivariate Gaussian random samples and summarize our results from this section in Figures 3 and 4. The main conclusions are stated as Result 9, which complements the existing literature about the empirical eigenvalues' behavior. Result 9 i) is not new, but Result 9 ii) is slightly surprising, though something related has appeared in (Muirhead, 1987) – see p.278 – in the context of the Wishart random matrix. A somehow related mathematical argument is that

$$\mathbf{I}_p = \underset{\mathbf{Q}:\, \mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p}{\operatorname{argmax}} \mathbf{1}^T \mathbf{Q}\, \operatorname{diag}(\boldsymbol{\lambda})\, \mathbf{Q}^T \mathbf{1} \qquad \text{for any fixed } \boldsymbol{\lambda} > \mathbf{0},$$

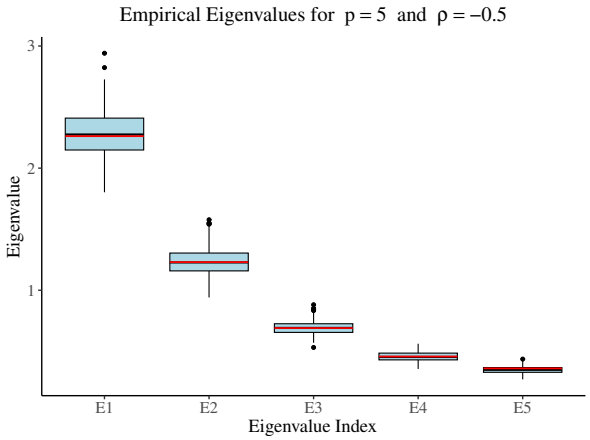
which one may obtain via the Rayleigh quotient result or use similar arguments to those used in the proof of Principal Component Analysis. Result 9 iii) is not surprising since the sum of the empirical eigenvalues is an unbiased estimator of the sum of the "true" eigenvalues due to the fact that the sample covariance is an unbiased estimator of the "true" covariance matrix.

III.2 Heuristics About Covariates' Dependence and Kolmogorov Setting

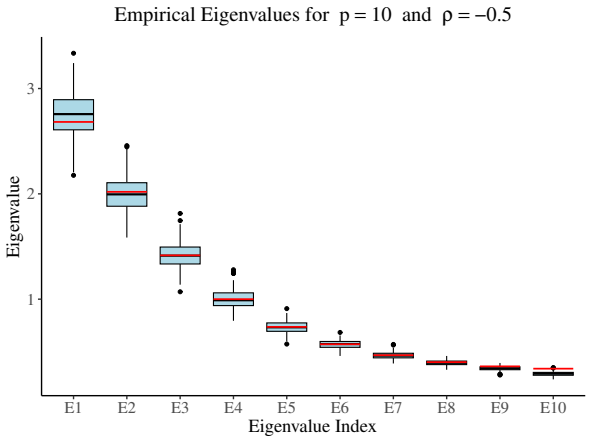
It has been noted in (El Karoui et al., 2013; El Karoui, 2013; Donoho and Montanari, 2016) that the OLS estimator has a non-zero asymptotic MSE under the Kolmogorov setting when both p and n get large. Under the assumptions that the covariates are Gaussian distributed with $\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$ and the error terms ϵ are i.i.d. with zero mean and variance σ^2, which is different than our setting given in Assumption 2.1, the OLS estimator – denoted as $\widehat{\boldsymbol{\beta}}^{OLS}(\boldsymbol{\beta}, \Sigma)$ satisfies

$$\widehat{\boldsymbol{\beta}}^{OLS}(\boldsymbol{\beta}, \Sigma) - \boldsymbol{\beta} \overset{d}{=} \left|\left|\widehat{\boldsymbol{\beta}}^{OLS}\left(\boldsymbol{\beta}, \mathbf{I}_p\right) - \boldsymbol{\beta}\right|\right|_2 \Sigma^{-1/2}\mathbf{U}, \tag{III.1}$$

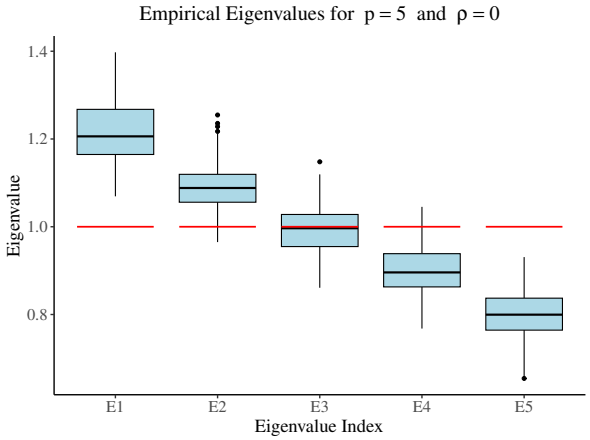
where $\boldsymbol{\beta} \in \Re^p$ is the "true" parameter vector, $\overset{d}{=}$ means equal in distribution and $\mathbf{U}$ is an p-dimensional uniformly (on sphere of radius 1) distributed random vector that is independent of $\widehat{\boldsymbol{\beta}}^{OLS}\left(\boldsymbol{\beta}, \mathbf{I}_p\right)$. Note that this setting assumes a zero intercept which simplifies the exposition. It



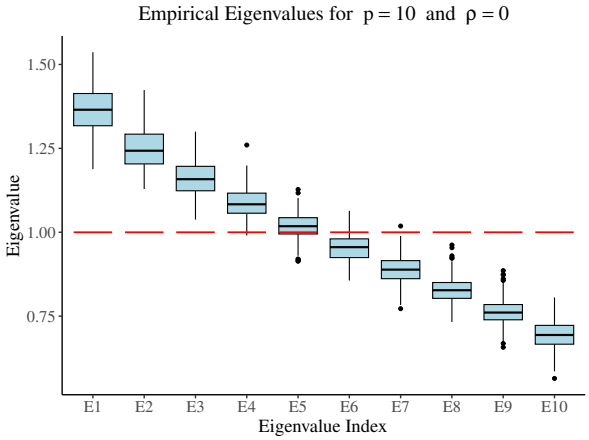
(a) $p = 5,\ \rho = -0.5$



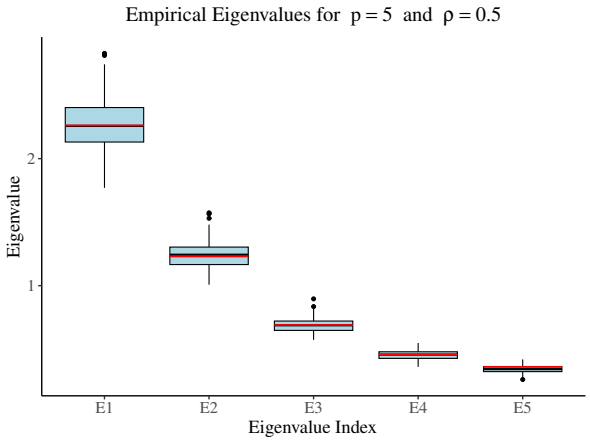
(b) $p = 10,\ \rho = -0.5$



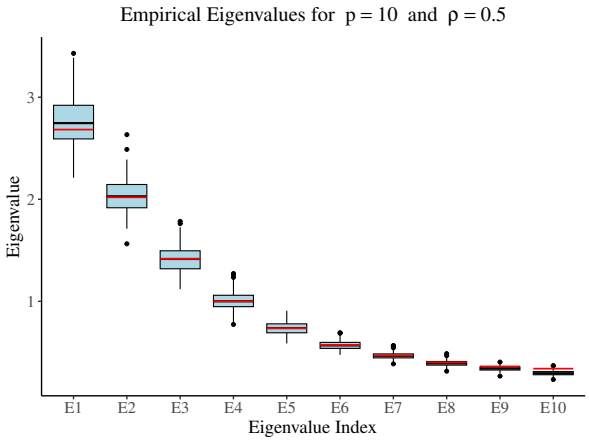
(c) $p = 5,\ \rho = 0$



(d) $p = 10,\ \rho = 0$

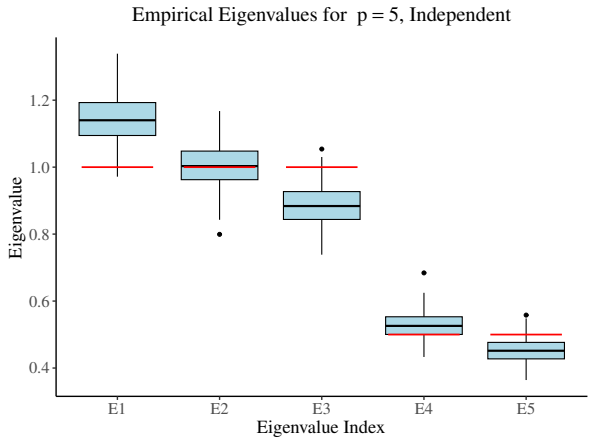


(e) $p = 5,\ \rho = 0.5$

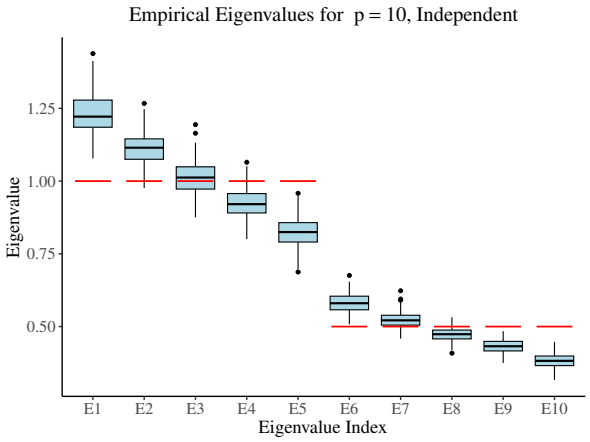


(f) $p = 10,\ \rho = 0.5$

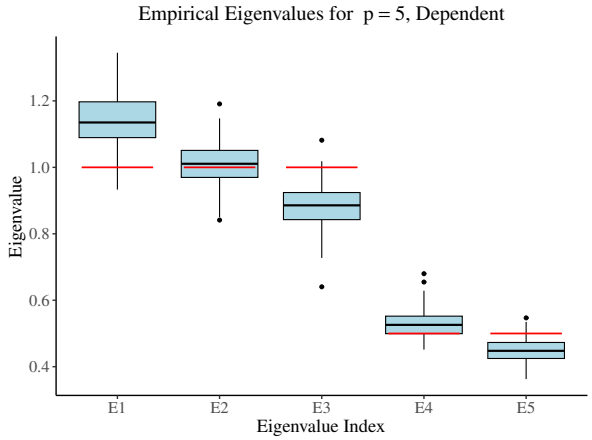
Figure 3: Boxplots of the empirical eigenvalues from sample covariance matrices computed from data generated as $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}(\rho))$, where $\mathbf{\Psi}(\rho)$ is a Toeplitz covariance matrix as in (II.1); results are based on $N = 250$ replicates of samples with sample size of $n = 250$. Each row shows results for a fixed correlation ρ of -0.5, 0 and 0.5 at the top, middle and bottom, respectively, and each column compares two settings when $p = 5$ (left) and $p = 10$ (right). The red horizontal segments indicate the "true" eigenvalues of $\mathbf{\Psi}(\rho)$.



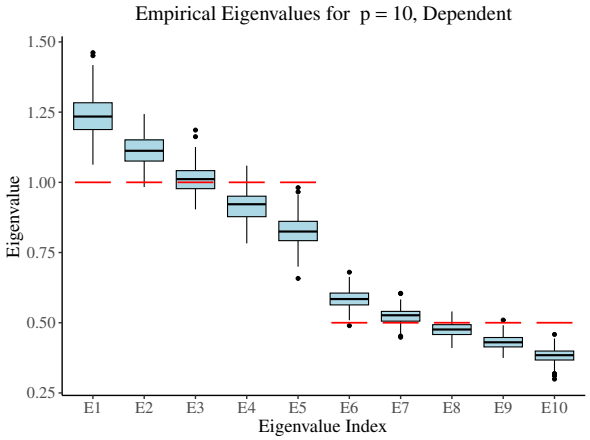
(a) $p = 5$, independent covariates



(b) $p = 10$, independent covariates



(c) $p = 5$, dependent covariates



(d) $p = 10$, dependent covariates

Figure 4: Boxplots of the empirical eigenvalues from sample covariance matrices computed from data generated as $\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}\,\mathrm{diag}(\boldsymbol{\lambda})\,\mathbf{Q}^{\mathbf{T}}\right)$, where $\mathbf{Q} = \mathbf{I}_p$ (top) and $\mathbf{Q}$ is a $p \times p$ randomly generated orthogonal matrix (bottom); results are based on $N = 250$ replicates of samples with sample size of $n = 250$. The red horizontal segments indicate the "true" eigenvalues, which are 1 and $1/2$; specifically, the first $\lceil p/2 \rceil$ eigenvalues are equal to 1 and the remaining $p - \lceil p/2 \rceil$ eigenvalues are set to $1/2$.

is not difficult to find that

$$\mathbb{E}[\mathbf{U}^T \Sigma^{-1} \mathbf{U}] = \frac{1}{p} \operatorname{Tr}\left(\Sigma^{-1}\right). \tag{III.2}$$

It is argued in (El Karoui et al., 2013; El Karoui, 2013; Donoho and Montanari, 2016) that

$$\left|\left|\widehat{\boldsymbol{\beta}}^{OLS}\left(\boldsymbol{\beta}, \mathbf{I}_p\right) - \boldsymbol{\beta}\right|\right|_2 \to \frac{\kappa}{1-\kappa}\sigma^2 \quad \text{as } n \to \infty, \text{ where } p/n \to \kappa \in (0,1) \text{ as } n \to \infty. \tag{III.3}$$

Putting (III.1)–(III.3) together, one may expect the asymptotic MSE of $\widehat{\boldsymbol{\beta}}^{OLS}(\boldsymbol{\beta}, \Sigma)$ and asymptotic MSE of $\widehat{\boldsymbol{\beta}}^{OLS}\left(\boldsymbol{\beta}, \mathrm{diag}(\boldsymbol{\lambda})\right)$ to be equivalent if Σ has eigenvalues given by $\boldsymbol{\lambda}$. Thus, one would

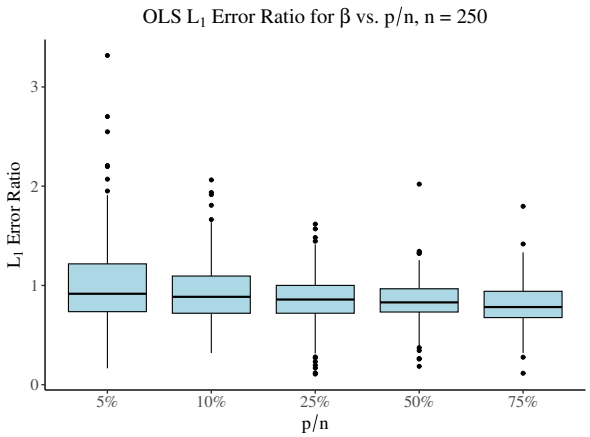
expect that the OLS estimator L_2 errors might be similar when comparing the dependence and independence cases.

The above possible conjecture is not exactly what we have found in Figures 5 and 6, where the population eigenvalues are identical in the dependence and independence cases, a property that is preserved by the empirical eigenvalues (see Figures 5 and 6, subplots (b) and (d)). This invariance empirical property is not present for the OLS estimator where L_1 errors are dissimilar (dependence vs. independence) with ratios smaller than 1, meaning that the independence case is closer to the ground truth. Similar to Figure 2, we choose to display the L_1 ratios instead of L_2 ratios as the L_1 distance is more sensitive to outliers.

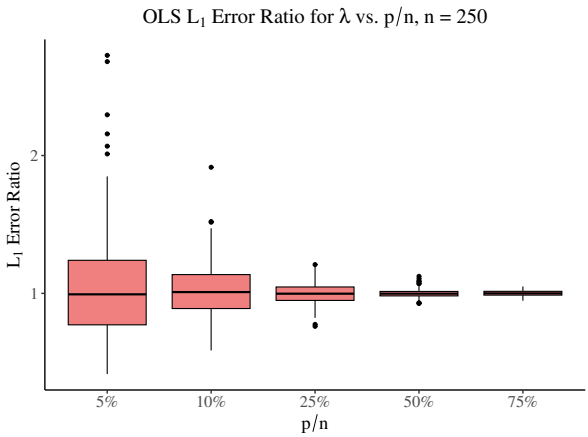
We also check the OLS L_1 equivalence between the dependence and independence cases when the "true" eigenvalues are preserved, but for the Kolmogorov setting, i.e., when both n and p get large so that $p/n \approx \kappa \in (0, 1)$. Our conclusion is stated as Result 10 i) where the pattern for small samples ($n = 250$) obtained in Figures 5 and 6 is observed for larger samples and various p/n ratios as it can be seen in Figure 7. Note that the invariance property for the empirical eigenvalues is preserved for large sample sizes as we have noted in Figures 5 and 6.

We further analyze if the St, DSh and GSR estimators exhibit the same property as OLS when comparing the possible L_1 equivalence between the dependence and independence cases when the "true" eigenvalues are preserved for the Kolmogorov setting. As in the OLS case, we show in Figures 8 – 10 that the estimation error of St/DSh/GSR is lower for independent Gaussian covariates as compared to the dependent Gaussian covariates case, which is stated as Result 10 ii). In fact, we found that the ratios for OLS and DSh are similar (with ratios less than 1), while St and GSR exhibit a similar pattern that is different than OLS, i.e., St and GSR ratios are closer to 1 than OLS' ratios.

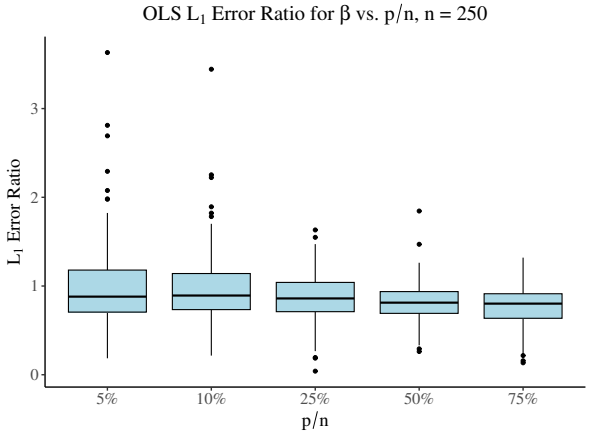
We now compare the estimation error of three shrinkage estimators (St, DSh and GSR) to OLS under the Kolmogorov setting, and as before, we choose L_1 errors to make such comparisons. For simplicity, we assumed independent standard Gaussian covariates given the previous empirical evidence. We found in Figure 11 that OLS clearly outperforms DSh for the high-dimensional settings that we considered here, but St outperforms OLS in all possible settings. Further, GSR outperforms OLS in almost all settings, and when it does not, the differences are within 1% on average median; for large ratios (such as $p/n \in \{90\%, 95\%\}$), the average median is improved by 7% to 23%. Furthermore, GSR outperforms St in almost all settings, and when it does not, the



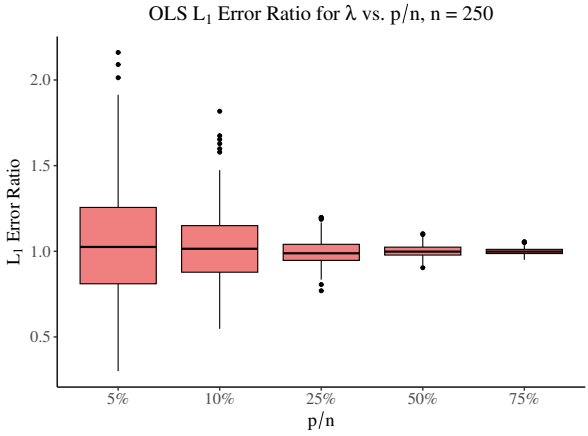
(a) $\boldsymbol{\beta} = (1, -1, 1 \ldots),\ \lambda_i \sim U(0, 10)$



(b) $\boldsymbol{\beta} = (1, -1, 1 \ldots),\ \lambda_i \sim U(0, 10)$



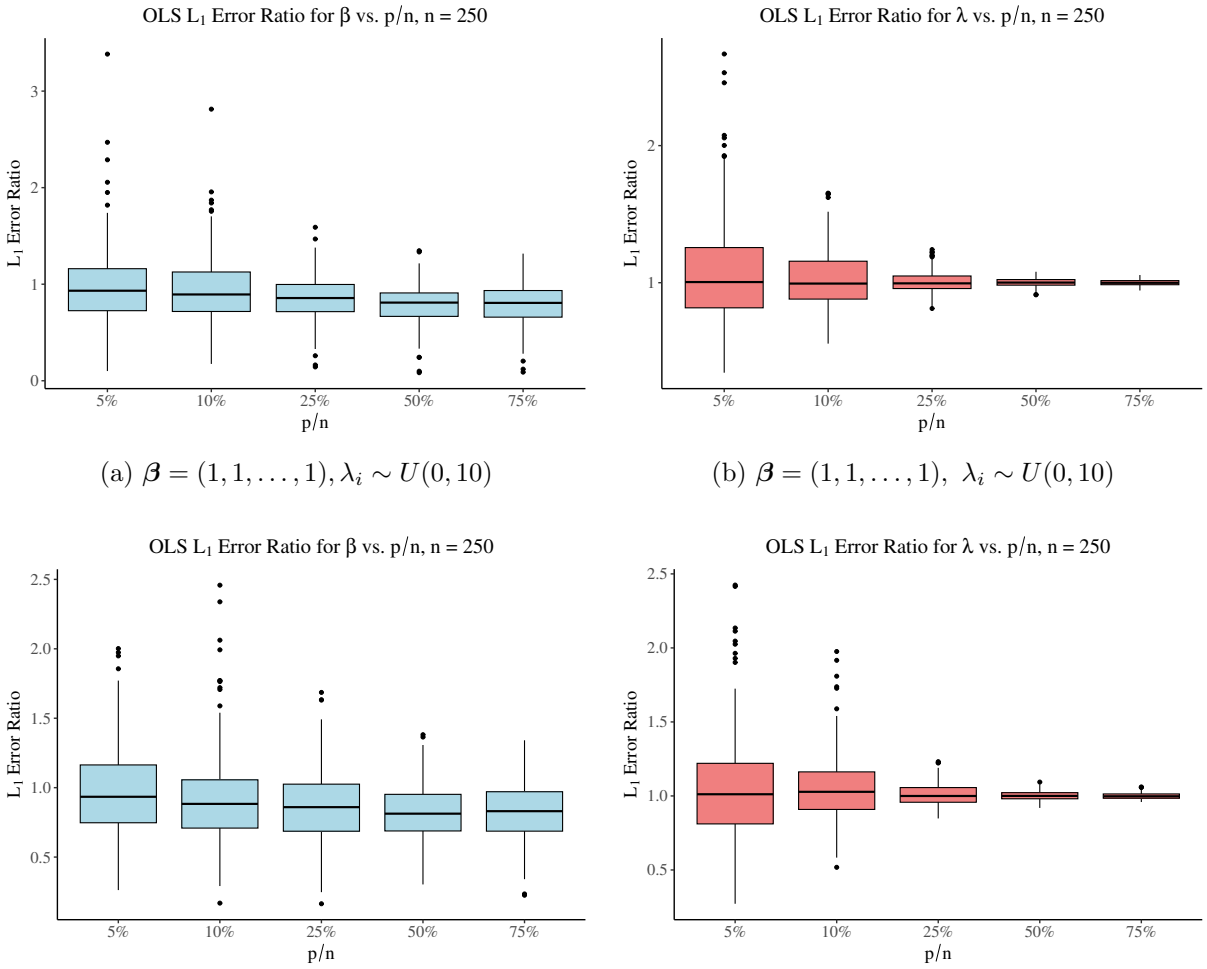
(c) $\boldsymbol{\beta} = (1, -1, 1 \ldots),\ \lambda_i \sim U(0, 1)$



(d) $\boldsymbol{\beta} = (1, -1, 1 \ldots),\ \lambda_i \sim U(0, 1)$

Figure 5: Boxplots of the L_1 error ratios – independent covariates (with $\mathbf{Q} = \mathbf{I}_p$) divided by dependent covariates (with $\mathbf{Q}$ being a random orthogonal matrix) – for OLS estimates (left) and empirical eigenvalues (right) for various p/n. A *ratio* < 1 indicates that the model fitted with independent covariates yields a lower L_1 error than that with dependent covariates. Each L_1 error ratio is based on two samples of size $n = 250$ drawn from populations with independent and dependent Gaussian covariates; both covariance matrices have the same eigenvalues $\boldsymbol{\lambda}$ that are randomly generated from $U(0, \lambda_{\max})$ with $\lambda_{\max} = 10$ (top) and $\lambda_{\max} = 1$ (bottom). Each boxplot is based on $N = 250$ replications and in all cases, the "true" regression coefficients are set as $\boldsymbol{\beta} = (1, -1, 1, \ldots)$, i.e., all are equal to 1 but with alternate signs.

differences are within 2% to 3% on average median; for large ratios (such as $p/n \in \{90\%, 95\%\}$), the average median is improved by 8% to 13%. We summarize this pattern as Result 10 iii).



(a) $\boldsymbol{\beta} = (1, 1, \ldots, 1), \lambda_i \sim U(0, 10)$

(b) $\boldsymbol{\beta} = (1, 1, \ldots, 1),\ \lambda_i \sim U(0, 10)$

(c) $\boldsymbol{\beta} = (1, 1, \ldots, 1),\ \lambda_i \sim U(0, 1)$

(d) $\boldsymbol{\beta} = (1, 1, \ldots, 1),\ \lambda_i \sim U(0, 1)$

Figure 6: Boxplots of the L_1 error ratios – independent covariates (with $\mathbf{Q} = \mathbf{I}_p$) divided by dependent covariates (with $\mathbf{Q}$ being a random orthogonal matrix) – for OLS estimates (left) and empirical eigenvalues (right) for various p/n. A *ratio* < 1 indicates that the model fitted with independent covariates yields a lower L_1 error than that with dependent covariates. Each L_1 error ratio is based on two samples of size $n = 250$ drawn from populations with independent and dependent Gaussian covariates; both covariance matrices have the same eigenvalues $\boldsymbol{\lambda}$ that are randomly generated from $U(0, \lambda_{\max})$ with $\lambda_{\max} = 10$ (top) and $\lambda_{\max} = 1$ (bottom). Each boxplot is based on $N = 250$ replications and in all cases, the "true" regression coefficients are all equal to 1.

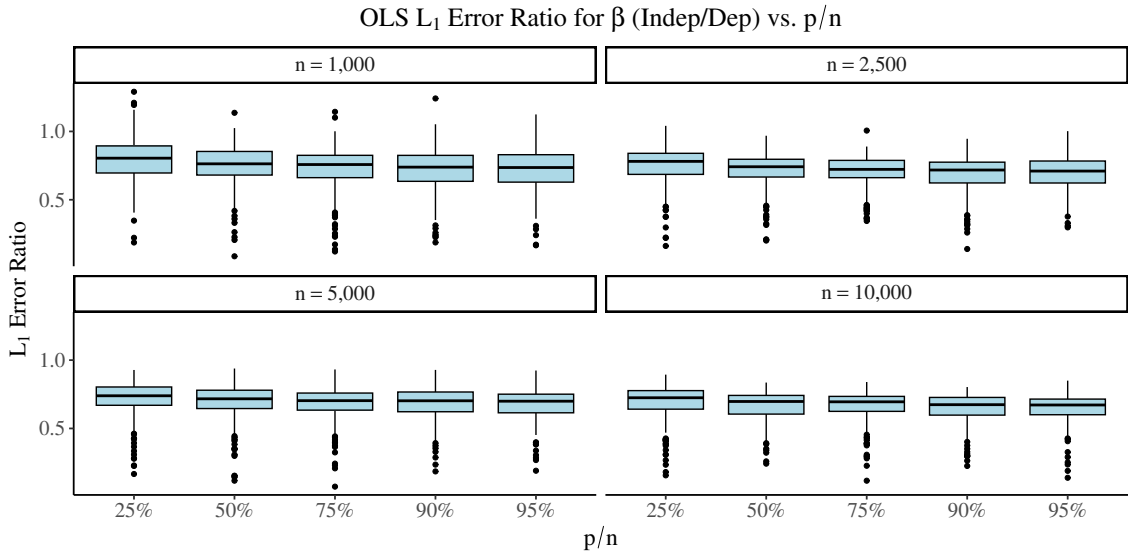


Figure 7: Boxplots of the L_1 error ratios – independent covariates (with $\mathbf{Q} = \mathbf{I}_p$) divided by dependent covariates (with $\mathbf{Q}$ being a random orthogonal matrix) – for OLS estimates based on samples of sizes $n = 1,000$ and $n = 10,000$ for various p/n. A $ratio < 1$ indicates that the OLS model fitted with independent covariates yields a lower L_1 error than the OLS model with dependent covariates. Each L_1 error ratio is based on samples drawn from populations with independent and dependent Gaussian covariates; both covariance matrices have the same eigenvalues $\boldsymbol{\lambda}$ that are randomly generated from $U(0,1)$. Each boxplot is based on $N = 250$ replications and in all cases, the "true" regression coefficients are all equal to 1.

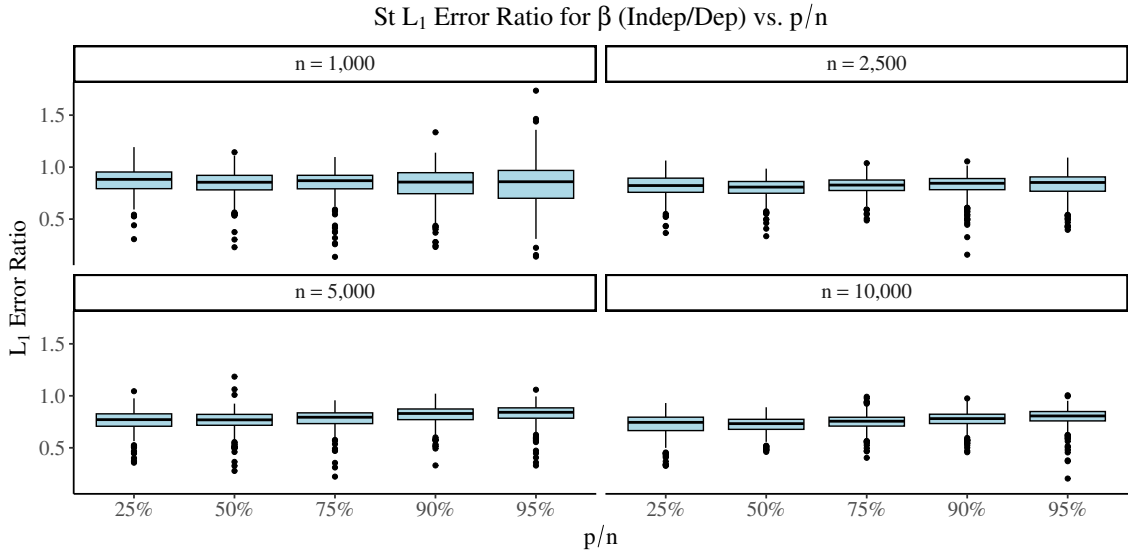


Figure 8: Boxplots of the L_1 error ratios – independent covariates (with $\mathbf{Q} = \mathbf{I}_p$) divided by dependent covariates (with $\mathbf{Q}$ being a random orthogonal matrix) – for St estimates based on samples of sizes $n = 1,000$ and $n = 10,000$ for various p/n. A $ratio < 1$ indicates that the St model fitted with independent covariates yields a lower L_1 error than the St model with dependent covariates. Each L_1 error ratio is based on samples drawn from populations with independent and dependent Gaussian covariates; both covariance matrices have the same eigenvalues $\boldsymbol{\lambda}$ that are randomly generated from $U(0,1)$. Each boxplot is based on $N = 250$ replications and in all cases, the "true" regression coefficients are all equal to 1.

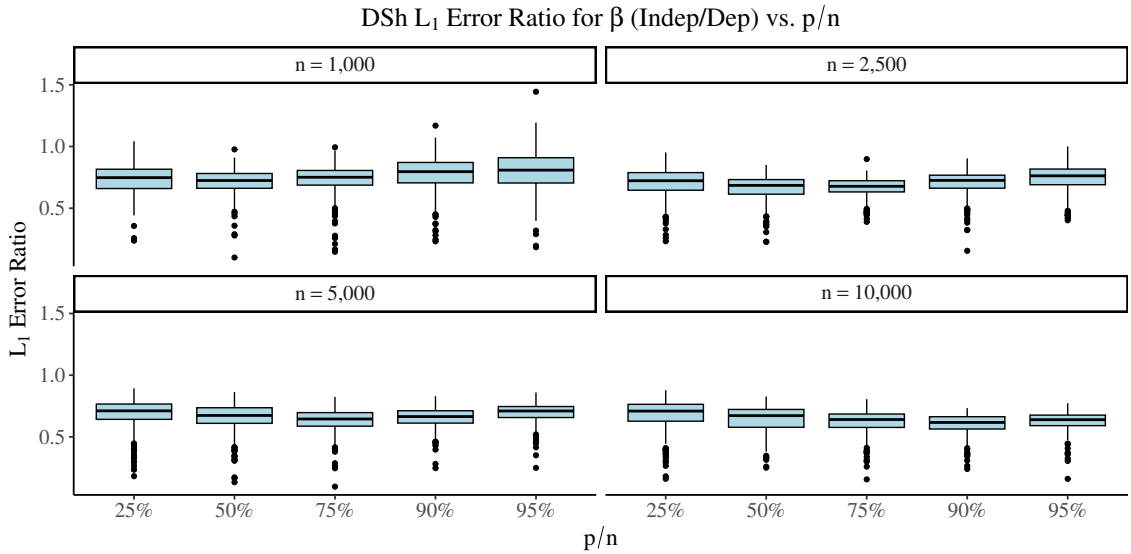


Figure 9: Boxplots of the L_1 error ratios – independent covariates (with $\mathbf{Q} = \mathbf{I}_p$) divided by dependent covariates (with $\mathbf{Q}$ being a random orthogonal matrix) – for DSh estimates based on samples of sizes $n = 1,000$ and $n = 10,000$ for various p/n. A $ratio < 1$ indicates that the DSh model fitted with independent covariates yields a lower L_1 error than the DSh model with dependent covariates. Each L_1 error ratio is based on samples drawn from populations with independent and dependent Gaussian covariates; both covariance matrices have the same eigenvalues $\boldsymbol{\lambda}$ that are randomly generated from $U(0,1)$. Each boxplot is based on $N = 250$ replications and in all cases, and the "true" regression coefficients are all equal to 1.

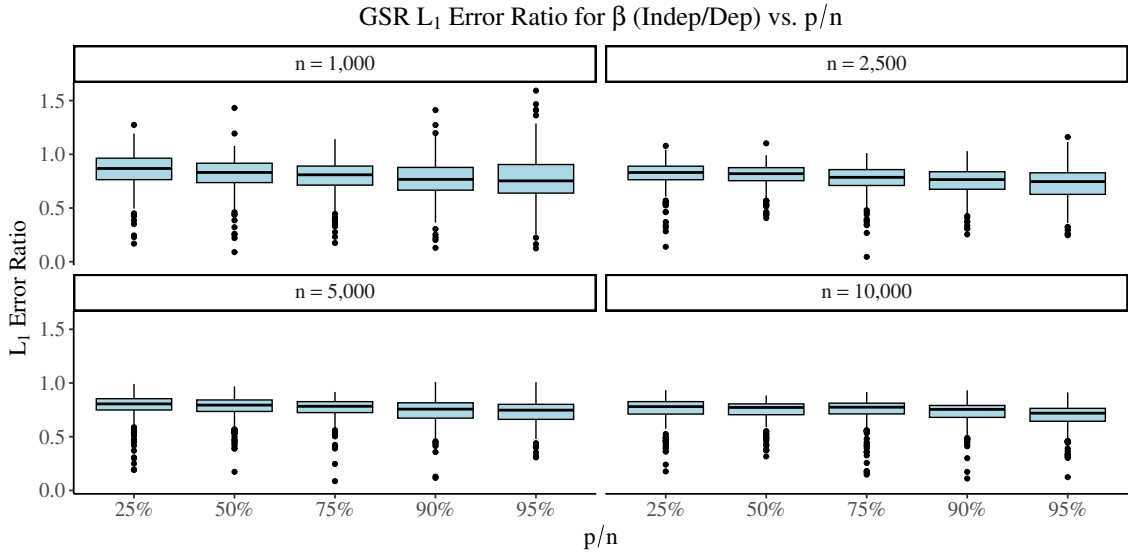


Figure 10: Boxplots of the L_1 error ratios – independent covariates (with $\mathbf{Q} = \mathbf{I}_p$) divided by dependent covariates (with $\mathbf{Q}$ being a random orthogonal matrix) – for GSR estimates based on samples of sizes $n = 1,000$ and $n = 10,000$ for various p/n. A $ratio < 1$ indicates that the GSR model fitted with independent covariates yields a lower L_1 error than the GSR model with dependent covariates. Each L_1 error ratio is based on samples drawn from populations with independent and dependent Gaussian covariates; both covariance matrices have the same eigenvalues $\boldsymbol{\lambda}$ that are randomly generated from $U(0,1)$. Each boxplot is based on $N = 250$ replications and in all cases, and the "true" regression coefficients are all equal to 1.

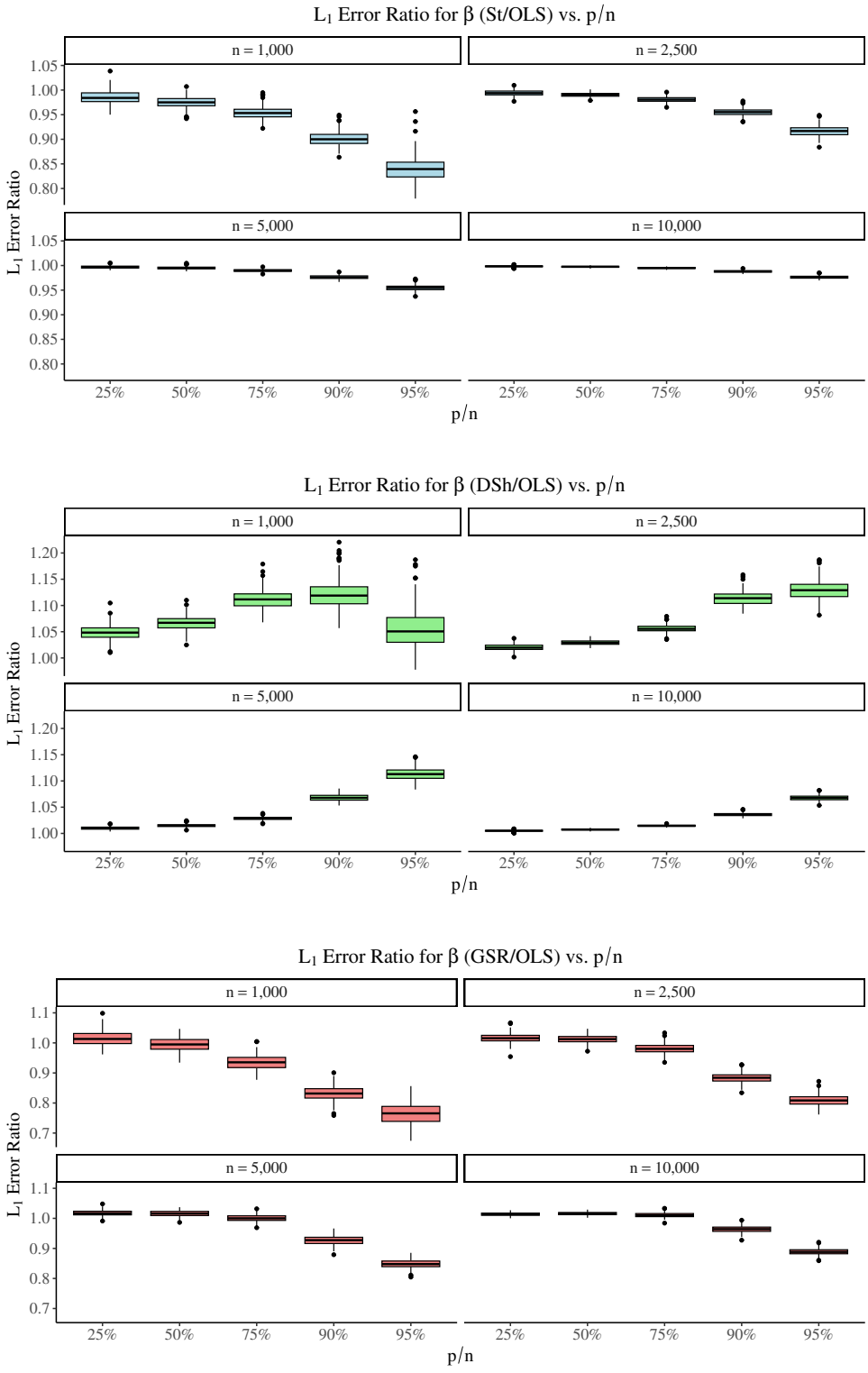


Figure 11: Boxplots of the L_1 error ratios of St vs. OLS (top), DSh vs. OLS (middle) and GSR vs. OLS (bottom) for various p/n; only independent covariates (with $\mathbf{Q} = \mathbf{I}_p$) are considered with samples of sizes $n = 1,000$ and $n = 10,000$. A $ratio < 1$ indicates that the shrinkage model yields a lower L_1 error than the OLS model. Each L_1 error ratio is based on samples drawn from populations with independent and dependent Gaussian covariates; both covariance matrices have the same eigenvalues $\boldsymbol{\lambda}$ that are randomly generated from $U(0, 1)$. Each boxplot is based on $N = 250$ replications and in all cases, the "true" regression coefficients are all equal to 1.

IV Statistical Fine-Mapping Application

IV.1 Fine-mapping algorithm modification

We integrated our new estimation methods within single and multi-trait fine-mapping methods JAM (expanded) and flashfm at several stages. First, within the genetic region of interest, we calculated GWAS summary statistics by fitting a single-SNP regression model at each genetic variant in the region, using one of OLS, GSR, St, and DSh. The resulting effect estimates ($\hat{\beta}$) at each variant were input into the JAM algorithm, together with the SNP correlation matrix (thinned so that no variants had squared correlation greater than 0.99) to identify initial multi-SNP models (multiple regression models) of potential causal variants. Within the JAM (expanded) algorithm, we used the selected estimation method (one of OLS, GSR, St, and DSh) to re-fit all multi-SNP models (initial multi-SNP models and tag multi-SNP models, where each variant in the initial model is replaced by variants with correlation at least 0.99 with the initial variant Hernández et al. (2021)). Finally, we substituted the multi-SNP effect estimates based on the selected method into previously derived estimates of the log *approximate Bayes' factor (ABF)* for single trait models and the joint log(ABF) for multiple traits, as derived in flashfm (Hernández et al., 2021). An **R** script[‡] with functions for our new implementation is also available for download.

IV.2 Data Generation

The data generation has been carried out under a realistic scenario that mimics the MAF and genetic variant correlation structure in a region containing the *IL2RA* gene (345 SNPs in chromosome 10p-6030000-6220000 (genomic build GRCh37/hg19)), which has genetic associations with autoimmune diseases such as multiple sclerosis (MS). This region has been previously shown to exhibit a tagging behaviour for causal variants making it more difficult to fine-map genetic associations at these variants; when there are two causal variants (C_1=rs61839660 and C_2=rs62626317), sometimes a different variant (D_1=rs2104286), that is correlated with both causal variants, is detected as a single causal variant (Asimit et al., 2019); in this region this tagging behaviour was also observed for two causal variants, C_1=rs61839660 and C_3=rs11594656, jointly tagged by D_2=rs706779.

[‡]Available at: https://github.com/jennasimit/flashfm-savvySh

For the *IL2RA* region, we generated a population of 10,000 individuals based on the CEU 1000 Genomes Phase 3 data (Consortium et al., 2015) using HapGen2 (Su et al., 2011). Each variant has a genotype score that takes on values 0, 1, and 2 and is the count of one of the two alleles at the variant. Under Hardy-Weinberg Equilibrium, the genotype score at each variant follows a binomial distribution with $N_q = 2$ number of trials and q_0 success probability, where q_0 is the frequency of the allele that is counted in the score, the effect allele frequency. The minor allele frequency (MAF) is the frequency of the allele that occurs with lower frequency (i.e, < 0.5). Only variants with $MAF > 0.005$ were included in our simulations. This aligns with our previous simulations involving counting covariates with Gaussian dependence.

For each of the 100 replications, a random sample of 5,000 individuals was selected from the population of 10,000. Quantitative traits were simulated to each have two causal variants, of which one (C_1) was shared between the traits; trait 1 had causal variants C_1 and C_2 and trait 2 had causal variants C_1 and C_3. Within each replication, the SNP effects for the causal variants were random uniformly generated to be between 0.15 and 0.4. Then for our two traits, the measurement for trait k of individual j, y_{kj}, is obtained from $y_{kj} = \sum_{i=1}^{m_k} \beta_{ik} x_{ij} + \varepsilon_{kj}$, where x_{ij} is the number of effect alleles of variant i for individual j (i.e. genotype score), β_{ik} is the effect of causal variant i for trait k, m_k is the number of causal variants for trait k (here, $m_k = 2, k = 1, 2$), and ε_{kj} is the k^{th} element of the j^{th} multivariate Normal distributed error variable with mean zero and covariance Σ, which is the covariance matrix of the traits. We set the variance of each trait to 0.20 and their correlation to 0.40.

For fine-mapping, the power of a method is estimated by the mean proportion of causal variants that are prioritized using a particular threshold for the MPP of causality (e.g., $MPP > 0.9$).

V Improve GLM Prediction – Cyber-sickness Data Example

This section provides further details of the summary of our cyber-sickness data analysis summarized in Section 3.3. We first provide a data description and an exploratory data analysis that would prepare the reader for the two GLM models we deploy here, namely, Logistic GLM (see *SI Appendix* V.1) and Poisson GLM (*SI Appendix* V.2).

We consider a *physiological* dataset with recordings from 23 participants while immersed in a VR roller coaster simulation, which can be found at GitHub repository and consists of 23 folders,

each containing the raw recordings of the 23 participants in the VR experiment. There are four groups of features in this dataset: *heart-rate (HR)*, *breath-rate (BR)*, *galvanic skin response (GaSR)*, and *heart-rate variability (HRV)*. In addition to the sensors' measurements, each of the four group of features includes the percentage of change from the *resting baseline (PC)*, *minimum inside the* $3s$ *rolling window (MIN)*, and *maximum value of the* $3s$ *rolling window (MAX)*; note that "s" refers to seconds in this dataset. The data are sampled at a time step of $1s$, while the length of the recordings for all participants vary between $567s$ and $1745s$.

At each measurement time t, we rely on 13 features $\boldsymbol{X}(t)$ that include HR and three sub-features (PC, MIN, MAX) for each group of features (BR, GaSR, HRV) as in (Kundu et al., 2022). The dependent/target variable at t is denoted as $\boldsymbol{Y}(t)$, which is the cyber-sickness FMS score, provided as verbal feedback during VR simulation. Thus, the original dataset is labeled on a scale from 0 (no cyber-sickness) to 10 (high cyber-sickness), using self-reported sickness feedback from the participants in the experiment; the sample distribution of the FMS scores is displayed in Figure 12 (left). The latter figure shows the skewed distribution of the raw labels that would lead to poor classifiers, and therefore, (Kundu et al., 2022) suggested to regroup the FMS scores into four different severity classes: i) "class 0" of no cyber-sickness when $FMS = 0$, ii) "class 1" of low cyber-sickness when $FMS = \{1, 2, 3\}$, iii) "class 2" of moderate cyber-sickness when $FMS = \{4, 5, 6\}$, and iv) "class 3" of acute cyber-sickness when $FMS = \{7, 8, 9\}$.

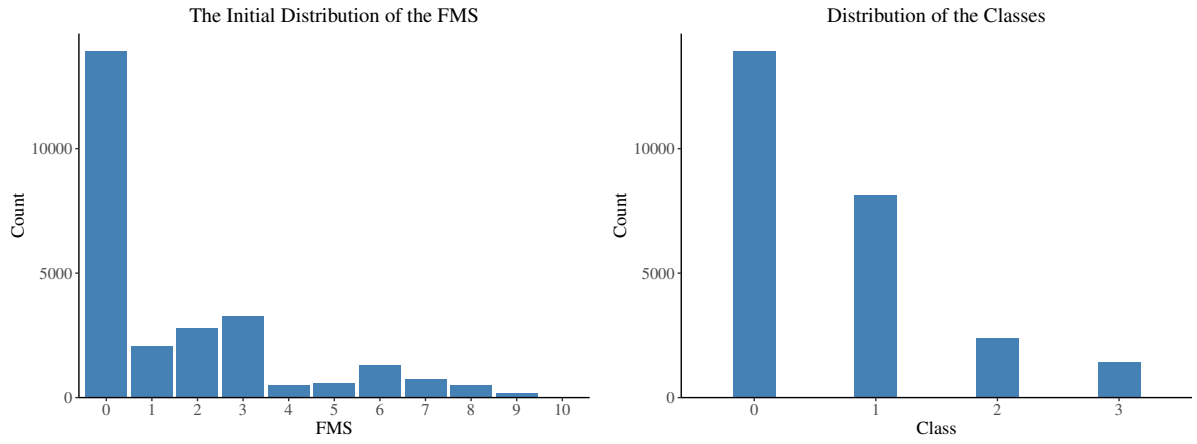


Figure 12: The distribution of the labels: original labels (left) and its regrouping in four classes (right)

Using a modified version of the provided Python *process_data.py* script, the data were sampled at $1s$ intervals and concatenated into a single file, *raw_data.csv* which contains $25,893$ rows and 15 columns. The first column ("Feedback") represents the FMS score, the last column

("Individual") is the id $(0-22)$ of the participants, while the columns (2 to 14) are the 13 features described before.

Note that there are 67 and 74 examples in the dataset recording $FMS = 2.5$ and $FMS = 3.5$, respectively, which deviate from the general scoring (integers from 0 to 10), and thus, we rounded down these scores to 2 and 3, respectively. We further process the data to feed into our prediction models by considering the $\boldsymbol{X}(t-M), \ldots, \boldsymbol{X}(t-1)$ feature space, where M is the number of prior time steps used for prediction. The choice of M would affect the prediction error and therefore, our prediction models are deployed for $M = 1$ (with 13 features $\boldsymbol{X}(t-1)$ and for $M = 10$ (with 130 features for $\boldsymbol{X}(t-10), \ldots, \boldsymbol{X}(t-1)$) to predict $\boldsymbol{Y}(t)$, denoted as $\widehat{\boldsymbol{Y}}(t)$. The features are used in two forms: i) raw data and ii) standardized form with zero mean and unit variance.

The upper correlation heatmap in Figure 13 indicates significant blocks of high correlations among the features for $M = 1$, which is not surprising given how some of them were generated (e.g., HR measurements, but the same happens for the HRV and GaSR blocks). The level of multicollinearity within the feature space increases even more for $M = 10$ because there is not much change in the physiological data when measured at a time step of $1s$. The fact that the features for $M = 10$ have more high correlations compared to $M = 1$ case can be observed in Figure 13 (b), where for a better readability, we displayed only the correlations between $\boldsymbol{X}(t-3), \boldsymbol{X}(t-2)$ and $\boldsymbol{X}(t-1)$.

For each M, the data are randomly split into 70% training and 30% testing sets, which is repeated $N = 50$ replications. Logistic and Poisson GLMs are fitted via IRLS – for details, see *SI Appendix* VII – which is a for-loop operation that solves a multiple linear regression model at each loop, and rely on six estimators, namely, OLS, SR, GSR, St, DSh, and Sh. Note that the standard statistical packages implement IRLS via OLS, and thus, we have modified the **R** function `glm.fit2` from the `glm2` package[§] to incorporate our shrinkage regression estimators into the classical IRLS implementation. To compare the performance of six models, the estimated MSE is evaluated and we report their average value for $N = 50$ replications in Tables 7 - 9; for each replication, the estimated $MSE = \frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{Y}(t) - \widehat{\boldsymbol{Y}}(t)\right)^2$, where T is the size of the dataset; note that $T = 25,870$ for $M = 1$ and $T = 25,663$ for $M = 10$.

The next two sections contain the two GLMs (Logistic and Poisson) that are deployed to the

[§]The `glm2` package is available at https://cran.r-project.org/web/packages/glm2/index.html.

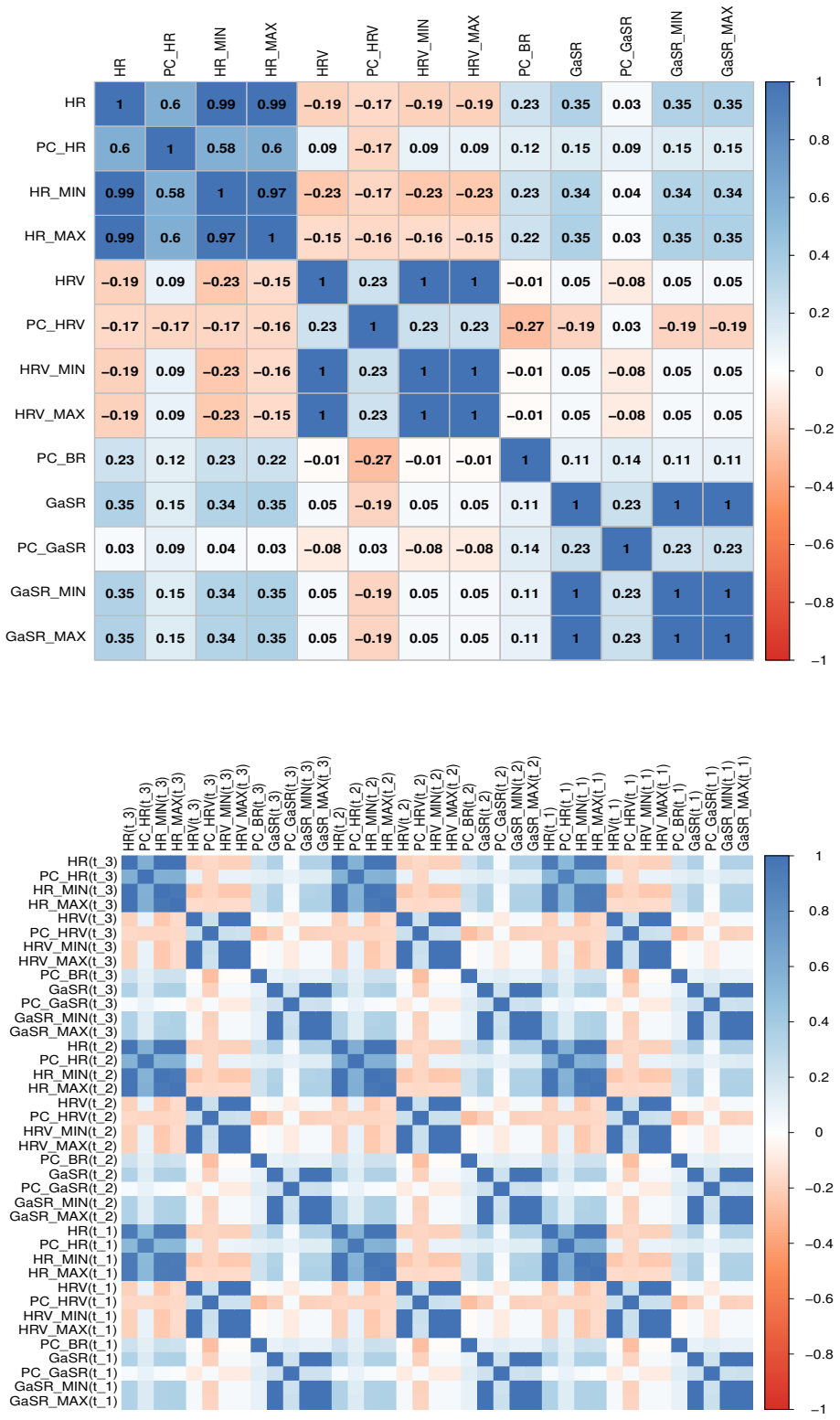


Figure 13: Correlation heatmaps for the feature space with $M = 1$ (upper) 39 and $M = 10$ (lower).

cyber-sickness data. Logistic GLM is considered in *SI Appendix* V.1 with its canonical link function (*logit* link function) for various pairs of classes defined by the raw FMS scores. Further, we employ a Poisson GLM in *SI Appendix* V.2 with its canonical link function (*log* link function) to incorporate the ordinal structure of the cyber-sickness severity levels for the data grouped in the four classes ("class 0" to "class 3") explained before; this mimics the purpose of a multi-classification model and we check its effectiveness. Note that the link function defines the response variable GLM estimator and its choice is crucial on deploying GLM models; for details about the link function, see (VII.2), though *SI Appendix* VII provides a more thorough discussion on GLM modeling.

V.1 Logistic GLM with *logit* Link Function

Logistic GLM is commonly known as the logistic regression and it is a special GLM case when the response variable is binomially distributed. Logistic GLM is useful in binary and multi-class classification problems, and the GLM methodology could be deployed as the binomial belongs to the exponential family defined in (VII.1). The GLM estimator is as in (VII.2) with the (canonical) *logit* link function given with $h(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$, where $\eta = \mathbf{x}^{\top}\boldsymbol{\beta}$. Binary classifications are considered for different pairs of FMS scores as follows: i) $FMS = 0$ vs. $FMS = 1$, ii) $FMS = 0$ vs. $FMS = 2$, iii) $FMS = 0$ vs. $FMS = 3$, and iv) $FMS = 0$ vs. $FMS = 6$. Note that $FMS = 0$ is the dominant label, while the other four labels ($FMS = \{1, 2, 3, 6\}$) have more than $1,000$ examples which avoids having extremely unbalanced binary classification exercises; for details, see Figure 12 (left).

Table 7: **Estimated MSE for binary classifications with raw input data**

Model	$FMS = 0$ **vs.** $FMS = 1$		$FMS = 0$ **vs.** $FMS = 2$		$FMS = 0$ **vs.** $FMS = 3$		$FMS = 0$ **vs.** $FMS = 6$	
	$M = 1$	$M = 10$	$M = 1$	$M = 10$	$M = 1$	$M = 10$	$M = 1$	$M = 10$
OLS	0.1261	0.1224	0.1399	0.1342	0.1270	0.1292	0.0839	0.0773
St	0.1261	0.1224	0.1404	0.1342	0.1270	0.1292	0.0840	0.0773
DSh	0.1318	0.4513	0.1536	0.1892	0.1637	0.3052	0.1141	0.1013
Sh	0.1261	0.1224	0.1399	0.1342	0.1270	0.1292	0.0839	0.0773
SR	0.1262	0.1224	0.1398	0.1342	0.1270	0.1292	0.0841	0.0774
GSR	0.1264	0.1228	0.1401	0.1329	0.1271	0.1275	0.0851	0.0790

Tables 7 and 8 summarize the results in this section and we draw two null conclusions. *First,* feature standardization does not significantly improve the model performance, and *second,* some shrinkage regressions (SR, GSR, and Sh) are no worse than OLS, but the performance differences

Table 8: **Estimated MSE for binary classifications with standardized input data**

Model	$FMS = 0$ **vs.** $FMS = 1$		$FMS = 0$ **vs.** $FMS = 2$		$FMS = 0$ **vs.** $FMS = 3$		$FMS = 0$ **vs.** $FMS = 6$	
	$M = 1$	$M = 10$	$M = 1$	$M = 10$	$M = 1$	$M = 10$	$M = 1$	$M = 10$
OLS	0.1266	0.1216	0.1406	0.1339	0.1271	0.1280	0.0842	0.0763
St	0.1278	0.1225	0.1436	0.1401	0.1278	0.1283	0.0878	0.0822
DSh	0.1355	0.3810	0.1523	0.3938	0.1408	0.3546	0.0870	0.3734
Sh	0.1266	0.1217	0.1406	0.1338	0.1271	0.1280	0.0842	0.0763
SR	0.1266	0.1216	0.1406	0.1339	0.1271	0.1280	0.0842	0.0763
GSR	0.1265	0.1219	0.1413	0.1321	0.1273	0.1260	0.0844	0.0766

are not significant. These imply that using shrinkage regressions in logistic regression would not significantly outperform the classical IRLS deployment that relies on OLS. While this is a disappointing result, the next section shows a very different picture where some shrinkage regressions are very effective for Poisson GLM deployments.

V.2 Poisson GLM with *log* Link Function

Poisson GLMs are deployed in this section by exploiting the ordinal type of FMS scores observed in Figure 12 (left), and we relabel the data as shown in Figure 12 (right) which keeps the ordinality trend. The sampling distribution is assumed to be Poisson, which is a member of the exponential family defined in (VII.1), and thus, the GLM machinery could be deployed. The GLM estimator in (VII.2) is with the Poisson canonical link function, known as *log* link function, and is given with $h(\eta) = e^{\eta}$, where $\eta = \mathbf{x}^{\top}\boldsymbol{\beta}$.

Table 9: **Estimated MSE for Poisson GLM with raw and standardized input data**

Model	**Raw Data**		**Standardized Data**	
	$M = 1$	$M = 10$	$M = 1$	$M = 10$
OLS	0.9174	0.8898	0.9233	0.8913
St	0.9172	0.8896	0.8940	0.8270
DSh	1.1498	3.0269	0.9604	1.9795
Sh	0.9175	0.8898	0.9232	0.8909
SR	0.9197	0.8898	0.9233	0.8913
GSR	0.9161	0.8871	0.9231	0.8896

Table 9 summarizes the Poisson GLM results and we draw some interesting conclusions. *First,* feature standardization improves the model performance for some GLM implementations. *Second,* St estimators significantly improve OLS implementations by approximately 3% and 7% for $M = 1$ and $M = 10$, respectively, while other shrinkage estimators (SR, GSR, and Sh) perform

at the same level as OLS.

We conclude this small data analysis by inferring that when deploying GLM estimation through IRLS, one may need to consider replacing OLS with shrinkage estimators in order to enhance the model performance. Even though the empirical evidence is limited, this conclusion is validated by a recent work of (Asimit et al., 2025a) that provides ample evidence in that respect via extensive simulated and real-data analyses. This is a viable, effective and computationally efficient method – since OLS and our shrinkage estimators, except for Sh which become computationally expensive in large-scale problems, are computationally equivalent – for reducing the notoriously high estimation error in GLM estimation.

VI Portfolio Investment Application

This section provides the additional pieces of information about the finance application briefly discussed in Section 3.4. A brief data description and technical details about the portfolio construction are given in *SI Appendix* VI.1. Numerical experiments are made across various market conditions in *SI Appendix* VI.2.

VI.1 Data Description and Methodology

The S&P500 is a stock market index that includes the 500 large-cap U.S. companies across various industries – e.g., technology, healthcare, finance, consumer goods, and industrials – and represents a market benchmark, meaning that investors and fund managers compare their portfolio performance against S&P500. The index is weighted by market capitalization, which implies that companies with larger market values have a bigger influence on the S&P500 movements.

We had collected the S&P500 daily returns of 1,070 companies that were part of the S&P500 at least once within the observation period that starts on January 1, 2000 and ends on December 31, 2023. The index's constituents change every three months, though firms could exit S&P500 due to mergers and acquisitions, poor financial results or failure to meet the eligibility criteria. Among the 1,070 firms, we have selected the 441 companies that remained listed on the US stock exchanges without interruption; we call this dataset as *DA441* for which we have data about 6,037 trading days per company.

The portfolio strategies considered in this section are risk minimization GMV portfolios, and

their very definition is as follows:

$$\mathbf{w}^* = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}} = \underset{\mathbf{w}:\, \mathbf{1}^T\mathbf{w}=1}{\operatorname{argmin}} \ \mathbf{w}^T\Sigma\mathbf{w}, \tag{VI.1}$$

where Σ is the covariance matrix of asset returns and $\mathbf{w}$ is the portfolio weight vector (the proportion of each asset in the entire portfolio). Note that the optimal solution in (VI.1) is the fully invested portfolio (since $\mathbf{1}^T\mathbf{w} = 1$); if the risk-free asset is included in the portfolio, then we have a non-fully invested portfolio and the equality constrain ($\mathbf{1}^T\mathbf{w} = 1$) is removed, and in turn, the optimal solution in (VI.1) becomes $\mathbf{w}^* = \Sigma^{-1}\mathbf{1}$. These explain why one could recast the unconstrained variant of (VI.1) as a standard regression problem. Specific to our very own setting – see also Section 3.1 of Fan et al. (2012) – we solve

$$\min_{\mathbf{w},b} \Big(\mathbf{y} - \mathbf{w}^T\mathbf{X} - b\mathbf{1}\Big)^T \Big(\mathbf{y} - \mathbf{w}^T\mathbf{X} - b\mathbf{1}\Big),$$

where $\mathbf{y} = R_{441}$ represents the return of the "target" asset (i.e., the last company in our *DA441* dataset) and $\mathbf{X} = R_{441} - R_j$ for $j = 1, \ldots, 440$. Any regression model can now be used to estimate for the regression coefficients $\boldsymbol{w}^* = (w_1, \ldots, w_{440})^T$ and intercept b; the weight for the remaining asset is then calculated as $w^*_{441} = 1 - \sum_{j=1}^{440} w^*_j$, so that the portfolio is fully invested.

We perform the previous estimations by using all eight regression methods, and a rolling window approach is employed to construct and evaluate portfolios over time. For each window, we use a fixed historical period (five or ten years) of daily returns for training and a subsequent three-month period (assuming 21 trading days a month, 252 days a year) for testing; after each three-month test, the window advances by three months. This design mimics a dynamic rebalanced portfolio, where investors regularly update the portfolio weights.

Once we obtain the portfolio weights based on the training data, we evaluate its future performance over the three-month testing period by using the **R** function `Return.portfolio` that is implemented without rebalancing. For each day in the testing period, the **R** function tracks the portfolio's value and we compute the following performance measures: i) average the daily returns and annualize them under the assumption of 252 trading days per year, ii) standard deviation of these daily returns and annualized equivalent values, and iii) Sharpe ratio that is the ratio between i) and ii).

VI.2 Out-of-Sample Performance

The most important test for a prediction model dealing with time series data are to evaluate its out-of-sample performance for which we employ a rolling window approach. After running all eight regression methods for each rolling window, we summarize the performance through three key metrics. *First,* the mean annual return is found by averaging the annualized returns from each window, which evaluates the overall profit of the eight investment strategies. *Second,* the mean annualized volatility is calculated by averaging each window's standard deviation of annualized returns, which evaluates the overall risk of the eight investment strategies. *Third,* we average the Sharpe ratios across all windows to evaluate overall risk-adjusted performance that measures the profit per unit of risk. Note that investors and fund managers are looking for high annualized returns and high Sharpe ratios. Besides such three overall performance measurements, we count the number of windows in which each method achieves the highest performance and convert this counting measure to the frequency of success.

Table 10 summarizes various performance measures under different rolling window settings. OLS shows very poor performance irrespective of the market conditions. Moreover, the shrinkage estimators have very good performance with St being the overall best estimator among all possible choices, but Sh also performs very well in risk-adjusted performance. Furthermore, eigenvalue-driven methods (RR, GSR and SRR) are useful to stabilize the risk, but are not effective from the point of view of investors that give their low expected returns and low Sharpe ratios.

To further analyze how our shrinkage estimators perform in adverse market conditions, we provide Table 11. Two time periods affected by extreme market conditions are chosen for this analysis; the first period is vastly influenced by the Financial Crisis (July 10, 2008 – March 8, 2011) and the second period coincides with the COVID-19 Pandemic (March 1, 2020 – January 14, 2022). Performance metrics include annual returns and annual Sharpe ratios computed over five-year and ten-year training windows. Given our three-month rolling window design, the Financial Crisis period includes eleven testing windows, while the COVID-19 Pandemic period includes eight testing windows, and thus, the counting and frequency of success are not computed for this analysis. The overall picture in Table 11 is not very different than what we have found in Table 10, and we conclude that some of our shrinkage estimators (namely, St, DSh, and Sh) are suitable for constructing portfolios during turbulent market periods.

Table 10: **Portfolio Performance Metrics Across All Rolling Windows**

Panel A: 5-years Training over 75 Rolling Windows

Models	Return			Standard Deviation			Sharpe Ratio		
	Mean	Counts	Freq (%)	Mean	Counts	Freq (%)	Mean	Counts	Freq (%)
OLS	12.02%	3	4.00%	11.10%	4	5.33%	1.456	3	4.00%
RR	11.57%	1	1.33%	9.95%	33	44.00%	1.654	8	10.67%
St	15.64%	24	32.00%	13.04%	0	0.00%	1.590	15	20.00%
DSh	14.17%	10	13.33%	11.99%	2	2.67%	1.536	9	12.00%
Sh	12.61%	20	26.67%	12.11%	15	20.00%	1.669	21	28.00%
SR	12.02%	6	8.00%	11.10%	3	4.00%	1.456	6	8.00%
GSR	12.04%	1	1.33%	10.57%	10	13.33%	1.569	1	1.33%
SRR	10.48%	10	13.33%	10.33%	8	10.67%	1.489	12	16.00%

Panel B: 10-years Training over 55 Rolling Windows

Models	Return			Standard Deviation			Sharpe Ratio		
	Mean	Counts	Freq (%)	Mean	Counts	Freq (%)	Mean	Counts	Freq (%)
OLS	14.87%	3	5.45%	10.79%	4	7.27%	1.794	3	5.45%
RR	14.79%	2	3.64%	10.28%	16	29.09%	1.894	4	7.27%
St	17.60%	17	30.91%	11.38%	4	7.27%	1.994	15	27.27%
DSh	16.71%	7	12.73%	11.03%	6	10.91%	1.973	8	14.55%
Sh	15.32%	16	29.09%	11.11%	12	21.82%	1.899	17	30.91%
SR	14.87%	3	5.45%	10.79%	1	1.82%	1.794	3	5.45%
GSR	15.13%	1	1.82%	10.61%	2	3.64%	1.854	1	1.82%
SRR	14.31%	6	10.91%	10.39%	10	18.18%	1.778	4	7.27%

Notes: Three performance measures (mean, risk measured via standard deviation, and Sharpe ratio) are computed for every rolling window and its summary results are tabulated. The "Mean" columns report the average of this annualized metric across all windows; the "Counts" columns indicate the number of windows in which a method achieves the highest performance for that metric among all methods, while the "Freq (%)" columns provide the corresponding percentage. Panel A assumes a 5-year training period that results in 75 rolling windows over the observation period, while Panel B assumes a 10-year training period over 55 rolling windows. Values highlighted in red denote the "best" performance for that metric within each panel.

VII Generalized Linear Model and its IRLS Implementation

A GLM assumes that the response variable Y, defined on $\mathcal{Y} \subseteq \Re$, is related to covariates $\mathbf{X}$ defined on $\mathcal{X} \subseteq \Re^d$, where $d = p + 1$ in this paper. The conditional distribution of Y belongs to the exponential dispersion family, with the following probability density or mass function

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\}. \tag{VII.1}$$

Here, θ is the canonical parameter, ϕ is the dispersion parameter, and a, b, and c are known functions. The function $b(\theta)$ determines the mean-variance relationship of the response variable.

Table 11: **Portfolio Performance Metrics under Extreme Market Conditions**

Model	Financial Crisis				Pandemic			
	5-year training		10-year training		5-year training		10-year training	
	Return	Sharpe Ratio	Return	Sharpe Ratio	Return	Sharpe Ratio	Return	Sharpe Ratio
OLS	2.93%	0.798	16.37%	1.9366	17.23%	1.908	14.87%	1.878
RR	8.77%	1.471	18.21%	2.2240	10.28%	1.578	13.50%	1.707
St	18.33%	1.535	20.91%	2.1902	22.98%	2.174	19.13%	2.124
DSh	14.42%	1.501	18.78%	1.9865	24.69%	2.248	19.66%	2.162
Sh	16.69%	1.833	22.24%	2.4382	11.32%	1.433	13.51%	1.389
SR	2.93%	0.798	16.37%	1.9367	17.23%	1.908	14.87%	1.878
GSR	6.06%	1.147	17.40%	2.0668	14.44%	1.782	15.08%	1.881
SRR	6.23%	1.183	17.16%	2.0803	7.10%	1.188	12.53%	1.418

Notes: Two performance measures (mean and Sharpe ratio) are computed for every rolling window during two major economic downturns: *Financial Crisis (July 10, 2008 - March 8, 2011)* and *COVID-19 Pandemic (March 1, 2020 - January 14, 2022)*; these results are averaged and tabulated. Because the *DA441* dataset has a starting date on January 3, 2000, the Financial Crisis period has eleven valid windows for the 5-year training and only six windows for the 10-year training, as the earliest 10-year window's testing phase starts on January 3, 2010. The Pandemic period yields eight rolling windows for both training lengths. Values highlighted in red denote the "best" performance for that metric within each panel.

The mean of Y is linked to a linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ through a link function g, so that

$$\mathrm{E}[Y_i|\mathbf{X}_i = \mathbf{x}_i] = h(\mathbf{x}_i^\top \boldsymbol{\beta}), \tag{VII.2}$$

where $h = g^{-1}$ is the inverse of the link function. The most natural choice (if this is possible) is the *canonical link function* where $h(\cdot) = b'(\cdot)$ on $\Re$.

The MLE estimator of $\boldsymbol{\beta}$ is chosen for GLM modeling, and the log-likelihood function for an independent sample of size n is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi), \quad \text{where } \theta_i = \left(b'\right)^{-1} \circ h\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right),$$

where $\circ$ is the composition operator. Maximizing the above is equivalent to minimizing the following objective function

$$\mathcal{C}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \big(\theta_i y_i - b(\theta_i)\big). \tag{VII.3}$$

Taking the derivative of $\mathcal{C}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and setting it to zero yields the *normal equations*

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad \forall j, \tag{VII.4}$$

where $\mu_i = h(\mathbf{x}_i^\top \boldsymbol{\beta})$ is the conditional mean, and $V(\mu_i)$ is the variance function determined by the exponential dispersion model.

The IRLS algorithm is used to solve the non-linear system of equations in (VII.4) by approximating (VII.3) as a WLS instance. This equivalence arises because solving (VII.4) is equivalent to minimizing the following WLS instance

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}. \tag{VII.5}$$

Since μ_i depends non-linearly on $\boldsymbol{\beta}$, the above is iteratively linearized using a Taylor expansion around the current parameter estimate $\widehat{\boldsymbol{\beta}}^{(t)}$ at each iteration t. Specifically,

$$\widehat{\boldsymbol{\beta}}^{(t+1)} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\mathbf{z}^{(t)} - \mathbf{X}\boldsymbol{\beta}\right)^\top \mathbf{W}^{(t)} \left(\mathbf{z}^{(t)} - \mathbf{X}\boldsymbol{\beta}\right), \tag{VII.6}$$

where $\mathbf{W}^{(t)}$ is the weight matrix and $\mathbf{z}^{(t)}$ is the pseudo-response, updated at each iteration as

$$\mathbf{W}^{(t)} = \operatorname{diag}\left(\frac{\left(h'\left(\eta_i^{(t)}\right)\right)^2}{V\left(\mu_i^{(t)}\right)}\right), \quad \mathbf{z}_i^{(t)} = \eta_i^{(t)} + \frac{y_i - \mu_i^{(t)}}{h'\left(\eta_i^{(t)}\right)},$$

with $\mu_i^{(t)} = h(\eta_i^{(t)})$ and $\eta_i^{(t)} = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{(t)}$. The weight matrix $\mathbf{W}^{(t)}$ reflects the curvature of the objective function at the current parameter estimates, while the pseudo-response $\mathbf{z}^{(t)}$ incorporates the linearized adjustments based on residuals.

In summary, (VII.6) is repeatedly solved until convergence is achieved within a specified threshold for the change in the objective function – from (VII.3) – between two consecutive iterations though a maximal number of iterations may be imposed but the scale of GLM implementations in this paper do not require such imposition. Note that the IRLS algorithm effectively links the WLS formulation in (VII.5) with the iterative updates in (VII.6), reducing the GLM estimation problem to solving a sequence of multiple linear regression problems.