Journal of Machine Learning Research 24 (2023) 1-57 Submitted 11/22; Revised 10/23; Published 11/23

Robust High-Dimensional Low-Rank Matrix Estimation: Optimal Rate and Data-Adaptive Tuning

Xiaolong Cui NK.XLCUI@GMAIL.COM
School of Statistics and Data Science, Nankai University, China

Lei Shi LEISHI@BERKELEY.EDU
Department of Biostatistics, University of California, Berkeley, U.S.A.

Wei Zhong WZHONG@XMU.EDU.CN
WISE and Department of Statistics and Data Science, SOE, Xiamen University, China

Changliang Zou ZOUCL@NANKAI.EDU.CN
School of Statistics and Data Sciences, LPMC, KLMDASR and LEBPS, Nankai University, China

Editor: Ambuj Tewari

Abstract

The matrix lasso, which minimizes a least-squared loss function with the nuclear-norm regularization, offers a generally applicable paradigm for high-dimensional low-rank matrix estimation, but its efficiency is adversely affected by heavy-tailed distributions. This paper introduces a robust procedure by incorporating a Wilcoxon-type rank-based loss function with the nuclear-norm penalty for a unified high-dimensional low-rank matrix estimation framework. It includes matrix regression, multivariate regression and matrix completion as special examples. This procedure enjoys several appealing features. First, it relaxes the distributional conditions on random errors from sub-exponential or sub-Gaussian to more general distributions and thus it is robust with substantial efficiency gain for heavy-tailed random errors. Second, as the gradient function of the rank-based loss function is completely pivotal, it overcomes the challenge of tuning parameter selection and substantially saves the computation time by using an easily simulated tuning parameter. Third, we theoretically establish non-asymptotic error bounds with a nearly-oracle rate for the new estimator. Numerical results indicate that the new estimator can be highly competitive among existing methods, especially for heavy-tailed or skewed errors.

Keywords: heavy-tailed error, high dimension, low-rank matrix, non-asymptotic bounds, robustness, tuning parameter selection

1. Introduction

The estimation of low-rank matrices under high-dimensional settings has received extensive attention and in-depth research in the past decade. Its applications include recommendation systems (Ramlatchan et al., 2018), image inpainting (Zheng et al., 2018), compressed sensing (Golbabaee and Vandergheynst, 2012), sensor localization (Nguyen et al., 2019) and so on. The most popular model for low-rank matrix recovery is the linear operator model

$$\mathbf{y} = \mathfrak{X}(\mathbf{X}; \mathbf{A}_0) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{A}_0 \in \mathbb{R}^{m_1 \times m_2}$ is the matrix of interest, which is usually assumed to have a low-dimensional intrinsic structure, $\mathbf{y} \in \mathbb{R}^p$ is the response, $\mathbf{X}$ is the covariate vector/matrix

which belongs to some linear space $\mathcal{L}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is the random error and $\mathfrak{X} : \mathcal{L} \times \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^p$ is a bilinear operator with respect to each argument. We assume that $\mathbf{X}$ is independent of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}$ has independent elements. This model allows us to deal with several important problems in a unified manner, including matrix regression (matrix compressed sensing)(Recht et al., 2010), multivariate linear regression (Yuan et al., 2007) and matrix completion (Candès and Recht, 2009; Gross, 2011), among others. For example, when $p = 1$, $\mathfrak{X}(\mathbf{X}; \mathbf{A}) = \operatorname{tr}\left(\mathbf{A}^\top \mathbf{X}\right)$ and $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$ is a matrix of explanatory variables, the linear operator model (1) becomes the well-studied trace regression model (Negahban and Wainwright, 2011). See Section 2.1 for a detailed discussion.

One of the most successful estimation methods is the regularization approach based on the trade-off between fitting the target matrix to the data and minimizing the model complexity, i.e., solving

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A} \in \mathcal{S}} \left\{ Q_n(\mathbf{A}) + \lambda P(\mathbf{A}) \right\}, \tag{2}$$

where $\mathcal{S}$ is a convex parameter space in $\mathbb{R}^{m_1 \times m_2}$, $Q_n(\mathbf{A})$ is an empirical loss function, λ is a tuning parameter and $P(\mathbf{A})$ is an appropriate penalization function. Under this paradigm, the most popular one may be the matrix lasso, which considers a least-squared loss with the nuclear-norm penalization or its variants. The literature in this area is vast. To name a few, for the trace regression model, Negahban and Wainwright (2011) derived non-asymptotic Frobenius norm estimation bounds under the sub-Gaussian assumption on the noise. Law et al. (2021) established a nearly optimal in-sample prediction risk bound for the rank-constrained least-squares estimator under no assumptions on the design matrix. For matrix regression, Fazel et al. (2008) and Recht et al. (2010) used the matrix lasso to explore the possibility of recovering a target matrix by observing its linear projection onto chosen dictionaries. For multivariate regression, similar ideas can be found in Yuan et al. (2007), Bunea et al. (2011, 2012), Bing et al. (2019), Kong et al. (2020) and the references therein. For the problem of noisy matrix completion, Koltchinskii et al. (2011), Negahban and Wainwright (2012), Rohde and Tsybakov (2011), among others, investigated the properties of the nuclear-norm penalized least-squares. They derived estimation error bounds which are shown to match the information-theoretic lower bound up to logarithmic factors. Other related works base on paradigm (2) include Hu et al. (2020), Hu et al. (2021) and Yu et al. (2022). Another important method for low-rank matrix estimation are based on matrix factorization framework (Sun and Luo, 2016; Ma et al., 2018; Tong et al., 2021a; Zhang et al., 2022), which stand beyond the consideration of the penalized estimation framework of the current paper.

Although the quadratic loss based methods can recover the target matrix with optimal rate under sub-Gaussian random errors, it is extremely sensitive to heavy-tailed or skewed errors. Many literature adopt robust loss instead of quadratic loss to deal with heavy-tailed random errors, and also establish the optimal estimation rate, see She and Chen (2017), Elsener and van de Geer (2018) and Tan et al. (2022), among others. However, these papers typically focus on one model and can not handle aforementioned important low-rank recovery problem in a unified manner. Fan et al. (2021) introduced the shrinkage principle to deal with heavy-tailed random errors in several important low-rank matrix recovery problems through the trace regression model. Although their estimators achieve the same estimation error rate as Negahban and Wainwright (2011) when the random error

has bounded second moments, an additional tuning parameter, the truncation level, needs to be determined. Consequently, a cross-validation method is inevitably required to select the regularization parameter λ and the truncation level together, which is time-consuming and lacks theoretical guarantee. More critically, the challenges of heavy-tailed error and tuning parameter selection are usually intertwined. Solutions specifically designed for only one aspect of the two challenges could leave the other aspect more unsatisfactory. To the best of our knowledge, there is no method that solves both problems simultaneously in the low-rank matrix estimation.

In this paper, we propose a robust procedure, termed as rank matrix lasso, in a unified high-dimensional low-rank matrix estimation framework. It enjoys both robustness for heavy-tailed error distributions and computational efficiency of the tuning parameter selection. Our major contributions are listed from the following three aspects. **(1)** From the methodology aspect, we propose a new tuning-easy robust method incorporating a Wilcoxon-type rank-based loss function with the nuclear-norm penalty for the low-rank matrix estimation model (1). **(2)** From the computation aspect, as the gradient function of the rank-based loss function is completely pivotal, it overcomes the challenge of tuning parameter selection and substantially saves the computation time by using an easily simulated tuning parameter. Thus, the rank matrix lasso is tuning-easy. **(3)** From the theory aspect, we establish non-asymptotic error bounds with a nearly-oracle rate for the new estimator in a unified high-dimensional low-rank matrix recovery framework under much weaker assumptions on covariates and random errors. We largely overcome the bounded and centralized assumption on covariates in Wang et al. (2020) which is restrictive for low-rank matrix recovery. Technical arguments for extending the existing linear regression model to our more general linear operator model are nontrivial and may be also interesting in their own rights.

The remainder of our paper is structured as follows. In Section 2, we present the proposed rank matrix lasso procedure. Section 3 studies the theoretical non-asymptotic properties of the new estimator. Numerical studies, including simulations and a real-data application, are presented in Section 4 and Section 5 respectively. Section 6 concludes the paper. Technical proofs and additional simulation results are provided in the Appendix.

Notations. Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ be a rectangular matrix. We use $\|\mathbf{A}\|_\infty = \max_{i,j} |a_{ij}|$ to denote the ℓ_∞ norm, $\|\mathbf{A}\|_F$ for Frobenius norm, $\|\mathbf{A}\|_{\text{op}}$ for operator norm, and $\|\mathbf{A}\|_1$ for nuclear norm. For square matrix $\mathbf{A}$, denote the smallest eigenvalues of $\mathbf{A}$ by $\lambda_{\min}(\mathbf{A})$. For vectors, we use $\|\cdot\|_1$ and $\|\cdot\|_2$ for the ℓ_1 and ℓ_2 norms, respectively. Let $\psi_p(x) = e^{x^p} - 1, p \geq 1$, then the ψ_p-Orlicz norm of a random variable X is defined as: $\|X\|_{\psi_p} = \inf\{t > 0 : \mathbb{E}\{\psi_p(|X|/t)\} \leq 1\}$. For a random vector $\mathbf{x} \in \mathbb{R}^d$, we define its ψ_p-Orlicz norm $\|\mathbf{x}\|_{\psi_p} := \sup_{\mathbf{v} \in \mathcal{D}^{d-1}} \|\mathbf{v}^\top \mathbf{x}\|_{\psi_p}$, where $\mathcal{D}^{d-1}$ is the d-dimensional unit sphere. Let $\mathbf{e}_k(m)$ be the k-th m-dimensional unit vector.

2. Methodology

In this section, we first introduce the model and several examples in Section 2.1. Then, in Section 2.2, we propose a new robust estimation method. Finally, in Section 2.3, we present a data-driven approach for tuning parameter selection.

2.1 Model

Consider n independent observations collected from the linear operator model (1)

$$\mathbf{y}_i = \mathfrak{X}(\mathbf{X}_i; \mathbf{A}_0) + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{y}_i$, $\mathbf{X}_i$ and $\boldsymbol{\varepsilon}_i$ are the response, covariate matrix and random error for the ith observation, respectively. We assume that $\mathbf{A}_0$ is nearly low-rank by requiring that its singular value sequence $\{\sigma_i(\mathbf{A}_0)\}_{i=1}^m$ decays quickly enough, where $\sigma_i(\mathbf{A}_0)$ is the i-th largest singular value of $\mathbf{A}_0$ and $m = \min\{m_1, m_2\}$. This assumption on $\mathbf{A}_0$ is less stringent and more natural to model the real-world problems than the exact low-rank assumption. In particular, for a parameter $q \in [0, 1]$ and a positive radius R_q, we consider $\mathbf{A}_0$ coming from the set

$$\mathcal{B}_q(R_q) := \left\{ \mathbf{A} \in \mathbb{R}^{m_1 \times m_2} \mid \sum_{i=1}^{m} \sigma_i^q(\mathbf{A}) \leq R_q \right\}.$$

Note that when $q = 0$, the set $\mathcal{B}_0(R_0)$ corresponds to the set of matrices with rank at most R_0. This model provides a unified high-dimensional low-rank matrix recovery framework including various cases of interest.

Example 1 (Matrix regression) *The matrix regression model is a setup in which one observes random linear projections of the unknown matrix $\mathbf{A}_0$. Concretely speaking, we have trace inner products*

$$y_i = \langle \mathbf{X}_i, \mathbf{A}_0 \rangle + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where $\langle \mathbf{X}_i, \mathbf{A}_0 \rangle = \operatorname{tr}\left(\mathbf{X}_i^\top \mathbf{A}_0\right)$, $\mathbf{X}_i \in \mathcal{L} = \mathbb{R}^{m_1 \times m_2}$ is a random matrix so that $\langle \mathbf{X}_i, \mathbf{A}_0 \rangle$ is a linear projection. In the typical form of matrix regression, which is called the compressed sensing, the observation matrix $\mathbf{X}_i$ has independent identically distributed (i.i.d.) standard normal entries. Here we relax this restriction to general sub-Gaussian ensembles. In this case, $\mathfrak{X}(\mathbf{X}_i; \mathbf{A}_0) = \langle \mathbf{X}_i, \mathbf{A}_0 \rangle = \operatorname{tr}\left(\mathbf{X}_i^\top \mathbf{A}_0\right)$ and $p = 1$. Moreover, model (2) includes the high-dimensional linear regression model as a special case. Let $m_1 = m_2 = m$, and take $\{\mathbf{X}_i\}_{i=1}^n$ and $\mathbf{A}_0$ to be diagonal, then $\langle \mathbf{X}_i, \mathbf{A}_0 \rangle = \mathbf{x}_i^\top \boldsymbol{\theta}_0$, where $\mathbf{x}_i$ and $\boldsymbol{\theta}_0$ denote the vectors of diagonal elements of $\mathbf{X}_i$ and $\mathbf{A}_0$, respectively. In this special case, having a low-rank $\mathbf{A}_0$ is equivalent to having a sparse $\boldsymbol{\theta}_0$.

Example 2 (Multivariate regression) *The goal of multivariate regression is to estimate a prediction function that maps covariates $\mathbf{x}_i \in \mathbb{R}^{m_2}$ to multidimensional output vectors $\mathbf{y}_i \in \mathbb{R}^{m_1}$. More specifically, consider the linear model*

$$\mathbf{y}_i = \mathbf{A}_0 \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{3}$$

where $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{m_1}$. We can write the multivariate regression as an instance of the linear operator model (1) with $\mathbf{X}_i = \mathbf{x}_i \in \mathcal{L} = \mathbb{R}^{m_2}$ and $\mathfrak{X}(\mathbf{x}_i; \mathbf{A}_0) = \mathbf{A}_0 \mathbf{x}_i$.

Example 3 (Matrix completion) *The matrix completion problem can be formulated into the trace inner products model (2), where the matrices $\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}$ are so-called masks. They are assumed to lie in*

$$\mathcal{X} = \left\{ \mathbf{e}_k(m_1)\, \mathbf{e}_l^\top(m_2) : 1 \leq k \leq m_1, 1 \leq l \leq m_2 \right\}.$$

We will assume that $\{\mathbf{X}_i\}_{i=1}^n$ *are i.i.d. samples with some underlying distribution* Π *on* $\mathcal{X}$. *The goal is to reconstruct all the entries of* $\mathbf{A}_0$.

2.2 New Estimation Method: Rank Matrix Lasso

For the unified high-dimensional low-rank matrix recovery model (1), we consider a new estimator of $\mathbf{A}_0$ by minimizing the following penalized loss function

$$\widehat{\mathbf{A}} = \underset{\mathbf{A}\in\mathcal{S}}{\arg\min} \left\{ Q_n\left(\mathbf{A}\right) + \lambda \left\|\mathbf{A}\right\|_1 \right\}, \tag{4}$$

where $\left\|\mathbf{A}\right\|_1$ denotes the nuclear norm of $\mathbf{A}$ and λ denotes the tuning parameter. The loss function is defined as

$$Q_n\left(\mathbf{A}\right) = \frac{2(n+1)}{\sqrt{3}n(n-1)} \sum_{k=1}^{p}\sum_{i=1}^{n} \phi\left(R\left(\varepsilon_{ik}\left(\mathbf{A}\right)\right)\right) \cdot \varepsilon_{ik}\left(\mathbf{A}\right), \tag{5}$$

where $\varepsilon_{ik}\left(\mathbf{A}\right) = y_{ik} - \mathbf{e}_k^\top\left(p\right)\mathfrak{X}\left(\mathbf{X}_i;\mathbf{A}\right)$, $R\left(\varepsilon_{ik}\left(\mathbf{A}\right)\right)$ denotes the rank of $\varepsilon_{ik}\left(\mathbf{A}\right)$ among $\varepsilon_{1k}\left(\mathbf{A}\right),\ldots,\varepsilon_{nk}\left(\mathbf{A}\right)$, and $\phi(1) \le \phi(2) \le \cdots \le \phi(n)$ is a set of scores generated as $\phi(i) = \varphi(i/(n+1))$ for some nondecreasing score function $\varphi(u)$ defined on the interval $(0,1)$ and standardized such that $\int_0^1 \varphi(u)du = 0$ and $\int_0^1 \varphi^2(u)du = 1$. Hereafter we denote the population version of the loss function $\mathbb{E}\{Q_n\left(\mathbf{A}\right)\}$ by $Q\left(\mathbf{A}\right)$. We will show that $\mathbf{A}_0$ is the minimizer of $Q\left(\mathbf{A}\right)$ under some weak conditions.

The multivariate rank-based loss function $Q_n\left(\mathbf{A}\right)$ was first proposed by Davis and McKean (1993) in low-dimensional multivariate linear model. In this paper, we consider the rank-based loss with Wilcoxon score

$$\varphi(u) = \sqrt{12}(u - 1/2) \tag{6}$$

to achieve both robustness and efficiency in the low-rank matrix recovery problem. As pointed out by She and Chen (2017), changing the squared error loss to a robust loss amounts to designing a set of multiplicative weights for $\varepsilon_{ik}\left(\mathbf{A}\right)$. We can regard the rank loss as using $R\left(\varepsilon_{ik}\left(\mathbf{A}\right)\right)/(n+1) - 1/2$ to weight $\varepsilon_{ik}\left(\mathbf{A}\right)$, while ℓ_2 loss and ℓ_1 loss use $\varepsilon_{ik}\left(\mathbf{A}\right)$ and $\mathrm{sign}(\varepsilon_{ik}\left(\mathbf{A}\right))$ as weights respectively. In light of this observation, the weight $R\left(\varepsilon_{ik}\left(\mathbf{A}\right)\right)/(n+1) - 1/2$ can be regarded as a balance between weight $\varepsilon_{ik}\left(\mathbf{A}\right)$ and weight $\mathrm{sign}(\varepsilon_{ik}\left(\mathbf{A}\right))$. Intuitively, outliers will not have as much impact on $R\left(\varepsilon_{ik}\left(\mathbf{A}\right)\right)/(n+1) - 1/2$ as they do for weight $\varepsilon_{ik}\left(\mathbf{A}\right)$, and at the same time, information on the relative magnitude of errors can still be utilized to improve the performance of estimator comparing to the ℓ_1 loss. Similar ideas also appear on the commonly used Huber loss which is a combination of ℓ_2 and ℓ_1 loss, but the truncation level needs to be determined.

Wang and Li (2009) considered the weighted Wilcoxon-type univariate rank-based loss with the SCAD penalty (Fan and Li, 2001) for low-dimensional linear regression. Furthermore, Wang et al. (2020) investigated the appealing features of this Wilcoxon-type univariate rank-based loss function for high-dimensional linear regression. A natural question is whether the new estimator $\widehat{\mathbf{A}}$ in (4) for high-dimensional low-rank matrix recovery problems can still inherit the merits of the rank lasso estimator for linear regressions (Wang et al., 2020). In the later sections, we name our new robust method via the penalized multivariate rank-based loss optimization with score function (6) for high-dimensional low-rank

matrix recovery problems as rank matrix lasso. We will show that the new estimator $\widehat{\mathbf{A}}$ behaves very similarly as matrix lasso for normal random errors and remains robust under heavy-tailed errors.

2.3 The Choice of the Tuning Parameter

It is critical to select the tuning parameter λ for the regularization methods in a computationally efficient way as different λ's may produce quite different models. Traditional cross-validation (CV) or information criteria techniques are computationally inefficient to exhaustively search an appropriate value of λ. Fortunately, it was noted in Wang et al. (2020) that for high-dimensional linear regression model, the gradient function of the rank based loss function is completely pivotal (Belloni et al., 2011; Parzen et al., 1994), leading to an appealing tuning-easy property. This inspires us to consider whether the similar property can be achieved by our rank matrix lasso method. Pivotal tuning would be especially interesting in matrix cases, since it allows us to circumvent the difficulty of tuning parameter selection for high-dimensional matrix estimation problems, which are typically very time-consuming if we apply conventional selection criteria such like cross-validation.

Let $R_{ik} = \operatorname{rank}(\varepsilon_{ik})$ be the rank of ε_{ik} among $\{\varepsilon_{1k}, \ldots, \varepsilon_{nk}\}$. Write $\xi_{ik} = 2R_{ik} - (n+1)$ for $i = 1, \ldots, n, k = 1, \ldots, p$. By the definition of $Q_n(\mathbf{A})$, direct computation yields the gradient of $Q_n(\mathbf{A})$ evaluated at $\mathbf{A}_0$,

$$\nabla Q_n(\mathbf{A}_0) = -2\{n(n-1)\}^{-1} \sum_{k=1}^{p} \sum_{i=1}^{n} \mathbf{H}_{ik} \xi_{ik},$$

where $\mathbf{H}_{ik} \in \mathbb{R}^{m_1 \times m_2}$ and the (a,b)-th element of $\mathbf{H}_{ik}$ is $\mathbf{e}_k^\top(p)\, \mathfrak{X}(\mathbf{X}_i; \mathbf{E}_{ab})$, here $\mathbf{E}_{ab} = \mathbf{e}_a(m_1)\, \mathbf{e}_b^\top(m_2) \in \mathbb{R}^{m_1 \times m_2}$.

It is important to observe that $\{R_{1k}, R_{2k}, \ldots, R_{nk}\}$ follows the uniform distribution on the permutations of the integers $\{1, 2, \ldots, n\}$. Therefore, $\nabla Q_n(\mathbf{A}_0)$ has a known distribution conditional on covariates $\mathbf{X}_1, \cdots, \mathbf{X}_n$.

By the theoretical analysis given in Section 3, conditional on the event that

$$\lambda \geq 2 \left\| \nabla Q_n(\mathbf{A}_0) \right\|_{\mathrm{op}}, \tag{7}$$

the rank matrix lasso estimator enjoys the nearly-oracle error bound, $\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F \propto \lambda^{1-q/2}$. Larger λ increases the probability of that event but will have an adverse effect on estimation accuracy. This suggests that it is desirable to choose a small λ such that the event (7) holds with high probability. In the same spirit of Wang et al. (2020), we introduce a new variable $S_n = \left\| \nabla Q_n(\mathbf{A}_0) \right\|_{\mathrm{op}}$ and recommend to take λ equal to

$$\lambda^* = 2G_{S_n}^{-1}(1 - \alpha_0), \tag{8}$$

where $G_{S_n}^{-1}(1-\alpha_0)$ denotes the $(1-\alpha_0)$th quantile of the distribution of S_n conditional on covariates $\mathbf{X}_1, \cdots, \mathbf{X}_n$. Because S_n is distribution-free as discussed above, the λ^* does not depend on the estimation of any unknown population quantity and thus can be obtained via a simulation method given $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

3. Non-asymptotic Properties

We first present Theorem 1 that serves as a roadmap to establish the convergence rate for $\widehat{\mathbf{A}}$. Then we apply this theorem to the three specific problems, matrix regression, multivariate regression and matrix completion, in Sections 3.1-3.3, respectively, and derive explicit non-asymptotic error bounds which allow us to compare with existing works.

Our main result is given as follows. The set $\mathcal{S}$ is a convex parameter space which is determined based on the concrete settings.

Theorem 1 *Suppose $\mathbf{A}_0 \in \mathcal{B}_q(R_q) \cap \mathcal{S}$ and that the regularization parameter λ is chosen such that $\lambda \geq 2\|\nabla Q_n(\mathbf{A}_0)\|_{\text{op}}$. Suppose further that for $\lambda_\varepsilon \geq 0$, $\widetilde{\lambda} \geq 0$ and all $\mathbf{A} \in \mathcal{S}$,*

$$|\{Q_n(\mathbf{A}) - Q(\mathbf{A})\} - \{Q_n(\mathbf{A}_0) - Q(\mathbf{A}_0)\}| \leq \lambda_\varepsilon \|\mathbf{A} - \mathbf{A}_0\|_1 + \widetilde{\lambda}.$$

Then for each integer $r \in \{1, 2, \ldots, m\}$, the estimator $\widehat{\mathbf{A}}$ satisfies

$$Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \leq \max\left\{12\sqrt{2r}(\lambda_\varepsilon + \lambda)\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F, 12(\lambda_\varepsilon + \lambda)\sum_{j=r+1}^{m}\sigma_j(\mathbf{A}_0), 3\widetilde{\lambda}\right\}.$$

Define the restricted set

$$\mathcal{C} := \left\{\boldsymbol{\Delta} \in \mathbb{R}^{m_1 \times m_2} \mid \|\boldsymbol{\Delta}''\|_1 \leq 3\|\boldsymbol{\Delta}'\|_1 + 4\sum_{j=r+1}^{m}\sigma_j(\mathbf{A}_0)\right\}.$$

If $Q(\mathbf{A}) - Q(\mathbf{A}_0) \geq \kappa\|\mathbf{A} - \mathbf{A}_0\|_F^2$ for some positive number κ and all $\mathbf{A} - \mathbf{A}_0 \in \mathcal{C}$, then we have

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F \leq \max\left\{24\sqrt{R_q}\left(\frac{\lambda + \lambda_\varepsilon}{\kappa}\right)^{1-q/2}, \sqrt{\frac{3\widetilde{\lambda}}{\kappa}}\right\}.$$

This theorem reveals that three major conditions are required to yield a convergence rate of $\widehat{\mathbf{A}}$. First, we need λ to be greater than $2\|\nabla Q_n(\mathbf{A}_0)\|_{\text{op}}$. Second, λ_ε and $\widetilde{\lambda}$ are two nonrandom constants which depend on the model parameters n, m_1 and m_2. They bound the quantity $Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0)$ by controlling the empirical process. We need to control an empirical process to specify a proper rate of λ_ε and $\widetilde{\lambda}$ by advanced empirical process techniques. At last, we need to verify that $Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \geq \kappa\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2$ which controls the quality of minoration of the loss function by a quadratic function and strongly relates to the shape of the density function of the errors. In the restricted set $\mathcal{C}$, we take $r = \#\{j \in \{1, 2, \ldots, m\} \mid \sigma_j(\mathbf{A}_0) \geq (\lambda + \lambda_\varepsilon)/\kappa\}$ in our proof, which is known as the "effective rank" of a near low-rank matrix (Negahban and Wainwright, 2011). In specific example, it can be shown that $\lambda + \lambda_\varepsilon$ decay to 0 as the sample size n increases, this effective rank will increase, reflecting the fact that as we obtain more samples, we can afford to estimate more of the smaller singular values of the matrix $\mathbf{A}_0$. The definition of $\boldsymbol{\Delta}'$ and $\boldsymbol{\Delta}''$ are given in the Appendix A to save space.

Next we will show that, when specialized to the three examples in Section 2.1, under much weaker assumptions on random errors, the proposed robust rank matrix lasso estimators achieve the same rates as those presented in Negahban and Wainwright (2011), Negahban and Wainwright (2012), Klopp (2014). For technical convenience, we assume $\mathcal{S} = \{\mathbf{A} \in \mathbb{R}^{m_1 \times m_2} \mid \|\mathbf{A}\|_\infty \leq \eta\}$ for some large positive constant η.

3.1 Matrix Regression

For the matrix regression model (2), we impose the following conditions on observation matrices and random errors.

Assumption 2 *The random errors ε_i's are i.i.d. with density function $f(\cdot)$. Let $\zeta_{ij} = \varepsilon_i - \varepsilon_j, 1 \le i \ne j \le n$. Denote $f^*(\cdot)$ as the probability density function of ζ_{ij}. There exists positive constants b_1 and b_2 such that $f^*(0) \ge b_1$ and $|\partial f^*(t)/\partial t| \le b_2$ for all t.*

Assumption 3 *The $\{\mathrm{vec}(\mathbf{X}_i)\}_{i=1}^n$ are i.i.d. sub-Gaussian vectors with $\|\mathrm{vec}(\mathbf{X}_i)\|_{\psi_2} \le \kappa_0 < \infty$. Denote the population covariance matrix $\mathbf{J} = \mathrm{Cov}(\mathrm{vec}(\mathbf{X}_i))$. Define the restricted eigenvalue*

$$\rho = \inf_{\mathbf{\Delta} \in \mathcal{C}, \mathbf{\Delta} \ne \mathbf{0}} \frac{\mathrm{vec}(\mathbf{\Delta})^\top \mathbf{J}\, \mathrm{vec}(\mathbf{\Delta})}{\mathrm{vec}(\mathbf{\Delta})^\top \mathrm{vec}(\mathbf{\Delta})}.$$

There exists constant b_3 such that $\rho \ge b_3 > 0$.

Assumption 4 *Let $\mathbf{\Delta} \in \mathbb{R}^{m_1 \times m_2}$, there exists a positive constant b_4 such that*

$$\frac{3b_1}{2\sqrt{2}b_2} \inf_{\mathbf{\Delta} \ne \mathbf{0}} \frac{\left(\mathbb{E}\langle \mathbf{X}_1 - \mathbf{X}_2, \mathbf{\Delta}\rangle^2\right)^{3/2}}{\mathbb{E}\left|\langle \mathbf{X}_1 - \mathbf{X}_2, \mathbf{\Delta}\rangle\right|^3} \ge b_4 > 0.$$

Remark 5 *Existing works on low-rank matrix estimation usually impose sub-Gaussian distribution (Negahban and Wainwright, 2011, 2012) or bounded moment condition (Fan et al., 2021) on random errors which excludes many heavy-tailed distributions and skewed distributions such as Cauchy distribution, log-normal distribution and χ^2 distribution. Assumption 2 relaxes such requirement to a large degree. Assumption 3 is standard to studying the error bound for matrix lasso estimators. It allows the covariance matrix $\mathbf{J}$ to be rank-degenerate when $\mathbf{A}_0$ is exact low rank such that $\sum_{j=r+1}^m \sigma_j(\mathbf{A}_0) = 0$. In such case, the restricted set becomes a cone $\mathcal{C} := \{\mathbf{\Delta} \in \mathbb{R}^{m_1 \times m_2} \mid \|\mathbf{\Delta}''\|_1 \le 3\|\mathbf{\Delta}'\|_1\}$. This cone completely excludes certain directions, and thus it is possible that the covariance matrix $\mathbf{J}$, while being rank-degenerate, can satisfy $\rho > 0$ over the cone. A simple sufficient condition of Assumption 3 is that the smallest eigenvalue of $\mathbf{J}$ is bounded away from zero. Similar assumptions are made in recent literature on high-dimensional linear regression and low-rank matrix estimation to establish error bounds, for example, Negahban and Wainwright (2011), Fan et al. (2018), Wang et al. (2020), and Tan et al. (2022). Assumption 4 controls the quality of minoration of the loss function by a quadratic function. From our theory, we have $Q(\mathbf{A}) - Q(\mathbf{A}_0) \ge \frac{b_1 b_3}{2}\|\mathbf{A} - \mathbf{A}_0\|_F^2$ for all $\mathbf{A} - \mathbf{A}_0 \in \mathcal{C}$ and $\|\mathbf{A} - \mathbf{A}_0\|_F \le b_4$. To guarantee that $Q(\mathbf{A})$ is sharply curved around the ground truth $\mathbf{A}_0$, we make the assumption $b_4 > 0$. It helps us transform the small loss difference $Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0)$ into small estimation error $\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F$. Assumption 4 also appears in Belloni and Chernozhukov (2011) as part of the "restricted identifiability and nonlinearity" condition. Similar conditions are used in recent literature (Gu and Zou, 2020; Zhou et al., 2023). Indeed, if the vectorized covariates $\mathrm{vec}(\mathbf{X})$ have a log-concave density, which includes many interesting distributions such as multivariate normal distributions and uniform distribution, then $b_4 \ge 3b_1/\left(2\sqrt{2}b_2K\right)$ for a universal constant K. This follows from the fact that $\mathbb{E}|\langle \mathbf{X}, \mathbf{\Delta}\rangle|^3 \le K(\mathbb{E}|\langle \mathbf{X}, \mathbf{\Delta}\rangle|^2)^{3/2}$*

holds for log-concave $\mathrm{vec}(\mathbf{X})$ *with some universal constant* K *by Theorem 5.22 of Lovász and Vempala (2007) and log-concavity is preserved under affine transformations and convolution; see Saumard and Wellner (2014) for a nice review of log-concavity.*

It is worth noting that the above conditions can ensure that $\mathbf{A}_0$ is the minimizer of population version loss function $Q(\mathbf{A})$. See Lemma 22 in the Appendix. The following theorem gives the estimation error rate of $\widehat{\mathbf{A}}$.

Theorem 6 *Suppose that* $\mathbf{A}_0 \in \mathcal{B}_q(R_q) \cap \mathcal{S}$ *and Assumptions 2-4 hold. The regularization parameter* λ *is chosen as* λ^**. Then there exists constant* $c > 0$ *and* $C > 0$ *such that when* $n > c(m_1+m_2)R_q^{2/(2-q)}$ *and* $m_1 + m_2 > \ln(2/\alpha_0)$*, the estimator* $\widehat{\mathbf{A}}$ *satisfies*

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2 \leq C\left(\frac{\kappa_0}{b_1 b_3}\right)^{2-q} R_q\left(\frac{m_1+m_2}{n}\right)^{1-q/2} \tag{9}$$

with probability at least $1 - \alpha_0 - 2\exp\{-(m_1+m_2)\}$.

The Frobenius norm rate here is identical to the rate established by Negahban and Wainwright (2011) under sub-Gaussian random error assumptions and also matches the minimax optimal rate of Frobenius norm established by Rohde and Tsybakov (2011). The condition $m_1 + m_2 > \ln(2/\alpha_0)$ is very weak. For a small $\alpha_0 = 10^{-4}$, this amounts to requiring $m_1 + m_2 > 10$. In our assumption, b_1 is kind of related to the dispersion measure of error. The smaller the dispersion of error, the larger the value that b_1 can take, resulting in smaller error bounds. This can be seen more clearly by using Gaussian error $\varepsilon \sim \mathcal{N}(\mu, \sigma^2)$. Now $\varepsilon_i - \varepsilon_j \sim \mathcal{N}(0, 2\sigma^2)$, $b_1 = f^*(0) = 1/(2\sqrt{\pi}\sigma)$ and the corresponding estimation error bound is

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2 \leq C\left(\frac{\kappa_0\sigma}{b_3}\right)^{2-q} R_q\left(\frac{m_1+m_2}{n}\right)^{1-q/2},$$

which indicates that our results are sharp for Gaussian errors like the result in Negahban and Wainwright (2011).

Comparing our result to the result in Fan et al. (2021). Our Assumption 2 cannot be directly compared with the moment condition in Fan et al. (2021) for that the two conditions apply to different settings. Our assumption can work well for the heavy-tailed distribution without moments, like the Cauchy distribution. This is also reflected by our simulation that our method performs better than Fan et al. (2021) under the Cauchy distribution. Fan et al. (2021) do not assume the existence of error's density function and include independent but not identical distributions. In Theorem 6, as b_1 approaches infinity, our error bound approaches zero. In contrast, Fan et al. (2021) works with the moment condition $\forall i = 1, \ldots, n, \mathbb{E}|y_i|^{2k} \leq M < \infty$ for some $k > 1$, and their estimation error bound is

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2 \leq C R_q\left(\frac{M^{1/k}(m_1+m_2)}{n}\right)^{1-q/2},$$

in which the values of (k, M) quantify the effect of error dispersions.

Remark 7 *In this theorem, we require that $m_1 + m_2$ tends to infinity so that the optimal rate hold with high probability. When applied to $m_1 + m_2$ is fixed, our theory also holds with $m_1 + m_2$ replaced by $\log n$, i.e.,*

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2 \leq C \left(\frac{\kappa_0}{b_1 b_3}\right)^{2-q} R_q \left(\frac{\log n}{n}\right)^{1-q/2}$$

with probability at least $1-\alpha_0-2n^{-1}$. If m_1+m_2 diverges faster than n in practical settings, some additional assumptions about the structure of $\mathbf{A}_0$, such as sparsity, are necessary since the Frobenius norm rate here matches the minimax optimal rate established by Rohde and Tsybakov (2011). Recent literature Tan et al. (2022) considers the sparse reduced rank regression, they addressed this problem by adding an additional ℓ_1 penalty to induce a sparse low-rank solution. Their theory implies that the $m_1 + m_2$ can diverge much faster than n with the sparsity condition. However, this setting is beyond the scope of our paper, and extending our approach to this case deserves future research.

Remark 8 *In our theory, the number $1 - \alpha_0$ can be regarded as the "confidence level" in the sense that our nonasymptotic bounds on the estimation error will be controlled at the optimal rate with probability close to $1 - \alpha_0$. In Appendix H3, we show the performance of our estimator under different values of α_0. The estimation error is not so sensitive to the choice of α_0 for different random errors and the confidence level $1-\alpha_0 \in [0.8, 0.9]$ would give good performance results in terms of balancing regularization bias with estimation variation. Our concrete recommendation for practice is to set $1 - \alpha_0 = 0.8$. After the α_0 is given, our estimator enjoys the minimax optimal convergence rate and our method is tuning-easy in the sense that the proposed penalization parameter is independent of the random error distribution and easier to obtain compared with other state-of-the-art methodologies.*

Remark 9 *In robust statistics, vertical outliers and leverage points are also worthy of attention. Our method is applicable for the vertical outliers. Consider that the vertical outliers are modeled by the Huber's ϵ-contamination model (Huber, 1992), specifically, the error ε_i follows a mixture distribution of the form $P_\epsilon = (1 - \epsilon)P + \epsilon Q$, where P is usually assumed to be some light-tailed distribution, Q is an arbitrary noise distribution, and ϵ measures the strength of contamination. As long as the error ε_i are i.i.d. from P_ϵ, the Assumption 2 holds for a large class of P_ϵ that includes the distributions and their mixtures that we consider in our simulations. But our method is not applicable to situations where leverage points exist. In our method, the selection of regularization parameter λ is related to the covariates. The existence of leverage points may cause λ to be very large, affecting the estimation error. In the Appendix H4, we conduct simulations to compare the performance of different methods in these two cases.*

For the special case, say the linear regression problem in the form of $\langle \mathbf{X}_i, \mathbf{A}_0 \rangle = \mathbf{x}_i^\top \boldsymbol{\theta}_0$, where $\mathbf{x}_i \in \mathbb{R}^m$ is the covariate vector and $\boldsymbol{\theta}_0 \in \mathbb{R}^m$ is the parameter of interest, we have the following result.

Corollary 10 *Suppose that Assumptions 2-4 hold and $\sum_{i=1}^m |\theta_{0i}|^q \leq R_q$, where $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0m})^\top$ and $0 \leq q \leq 1$. The regularization parameter λ is chosen as $\lambda = \lambda^*$.*

Then there exists constant $c > 0$ and $C > 0$ such that when $n > cR_q^{2/(2-q)} \log m$ and $m > (3/\alpha_0)^{1/3}$, the rank estimator, denoted as $\widehat{\boldsymbol{\theta}}$, satisfies

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2 \leq C \left(\frac{\kappa_0}{b_1 b_3}\right)^{2-q} R_q \left(\frac{\log m}{n}\right)^{1-q/2}$$

with probability at least $1 - \alpha_0 - 3m^{-3}$.

Due to special structures in the linear regression, the estimator $\widehat{\boldsymbol{\theta}}$ achieves a faster estimation rate than (9) in Theorem 6. This corollary is an extension of Wang et al. (2020)'s results to the sub-Gaussian design. The estimator $\widehat{\boldsymbol{\theta}}$ attains the minimax optimal rate of ℓ_2 norm established by Raskutti et al. (2011).

3.2 Multivariate Regression

For the multivariate regression model (3), we need the following conditions.

Assumption 11 *The random errors $\boldsymbol{\varepsilon}_i$'s are i.i.d. with marginal density function $f_k(\cdot)$ for ε_{ik}. Let $\zeta_{ijk} = \varepsilon_{ik} - \varepsilon_{jk}, 1 \leq i \neq j \leq n, 1 \leq k \leq m_1$. Let $f_k^*(\cdot)$ denote the probability density function of ζ_{ijk}. There exists positive constants b_1 and b_2 such that $f_k^*(0) \geq b_1$ and $|\partial f_k^*(t)/\partial t| \leq b_2$ for all t, uniformly in k.*

Assumption 12 *The covariates $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. sub-Gaussian vectors with $\|\mathbf{x}_i\|_{\psi_2} \leq \kappa_0 < \infty$. Denote $\mathbf{J} = \mathrm{Cov}(\mathbf{x}_i)$. Define the restricted eigenvalue*

$$\rho = \inf_{\boldsymbol{\Delta} \in \mathcal{C}, \boldsymbol{\Delta} \neq \mathbf{0}} \frac{\mathrm{tr}(\boldsymbol{\Delta}\mathbf{J}\boldsymbol{\Delta}^\top)}{\|\boldsymbol{\Delta}\|_F^2}.$$

There exists constant b_3 such that $\rho \geq b_3 > 0$.

Assumption 13 *Let $\boldsymbol{\Delta} \in \mathbb{R}^{m_1 \times m_2}$, there exists a positive constant b_4 such that*

$$\frac{3b_1}{2\sqrt{2}b_2} \inf_{\boldsymbol{\Delta} \neq \mathbf{0}} \frac{\left\{\sum_{k=1}^{m_1} \mathbb{E}\left|(\mathbf{x}_1 - \mathbf{x}_2)^\top \boldsymbol{\Delta}_k\right|^2\right\}^{3/2}}{\sum_{k=1}^{m_1} \mathbb{E}\left|(\mathbf{x}_1 - \mathbf{x}_2)^\top \boldsymbol{\Delta}_k\right|^3} \geq b_4 > 0$$

where $\boldsymbol{\Delta}_k^\top$ is the k-th row of $\boldsymbol{\Delta}$.

Theorem 14 *Suppose $\mathbf{A}_0 \in \mathcal{B}_q(R_q) \cap \mathcal{S}$ and Assumptions 11-13 hold. The regularization parameter λ is chosen such that $\lambda = \lambda^*$. Then there exists constant $c > 0$ and $C > 0$ such that when $n > c(m_1 + m_2)R_q^{2/(2-q)}$ and $m_1 + m_2 > \ln(2/\alpha_0)$, the estimator $\widehat{\mathbf{A}}$ satisfies*

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2 \leq C \left(\frac{\kappa_0}{b_1 b_3}\right)^{2-q} R_q \left(\frac{m_1 + m_2}{n}\right)^{1-q/2}$$

with probability at least $1 - \alpha_0 - 2\exp\{-(m_1 + m_2)\}$.

Once again, this estimation rate is identical to the rate established under sub-Gaussian random error in Negahban and Wainwright (2011).

3.3 Matrix Completion

For the matrix completion problem (Example 3), the trace regression formulation (2) will induce an identifiability issue if we try to apply the Wilcoxon-type loss directly, due to the special structure of the masks $\mathbf{X}_i$. It is well-understood in classic nonparametric literature that the Wilcoxon loss is unable to extract the intercept term from a linear model (Hettmansperger and McKean, 1998). To wit, note that we have the following fact holds almost surely: $\langle \mathbf{X}_i\,,\, c\mathcal{I}\rangle = c,$ for any $c \in \mathbb{R}$, where $\mathcal{I} = \mathbf{1}\mathbf{1}^\top$. Consequently, we can always reformulate (2) by offsetting the ground truth and introducing an intercept term

$$y_i = c + \langle \mathbf{X}_i\,,\, \mathbf{A}_0 - c\mathcal{I}\rangle + \varepsilon_i.$$

Hence, the failure for the estimation of c will cause the problem of identifiability between $\mathbf{A}_0$ and $\mathbf{A}_0 - c\mathcal{I}$.

However, there is a feasible solution by introducing the Rademacher sequence $a_i \in \{-1, +1\}$ and cope with the model

$$a_i y_i = \langle a_i \mathbf{X}_i\,,\, \mathbf{A}_0\rangle + a_i\varepsilon_i, \tag{10}$$

where $\{a_i\}_{i=1}^n$ is independent of $\{\mathbf{X}_i, y_i\}_{i=1}^n$. With this simple manipulation, we can overcome the intercept issue. To see this, suppose for some constant $\mathbf{C} \in \mathbb{R}^{m_1\times m_2}$ and $c \in \mathbb{R}$, we have almost surely $\langle a_i\mathbf{X}_i\,,\, \mathbf{C}\rangle = c$. It can be easily verified that this holds only when $\mathbf{C} = \mathbf{0}$ and $c = 0$, implying that no alternative offsetting formulation involving an intercept term exists for (10). Our proposed procedure is directly applicable with the pseudo observations $\{a_i y_i, a_i\mathbf{X}_i\}_{i=1}^n$. Thus for the matrix matrix completion problem, $Q_n(\mathbf{A})$ is

$$Q_n(A) = \frac{2(n+1)}{\sqrt{3}n(n-1)}\sum_{i=1}^n \phi\left(R\left(a_i\varepsilon_i(\mathbf{A})\right)\right)\cdot a_i\varepsilon_i(\mathbf{A}),$$

where $\varepsilon_i(\mathbf{A}) = y_i - \langle \mathbf{X}_i, \mathbf{A}\rangle$. We show that in Appendix E, (10) can guarantee that $\mathbf{A}_0$ is the minimizer of $Q(\mathbf{A})$ under appropriate condition.

In literature, a conventional setting for the $\mathbf{X}_i$'s is that they are i.i.d sampled from the uniform distribution Π. See Rohde and Tsybakov (2011), Koltchinskii et al. (2011) and Elsener and van de Geer (2018). We consider here a more general sampling model as formulated by Klopp (2014). More precisely, let $\pi_{jk} = \mathbb{P}\left(\mathbf{X} = \mathbf{e}_j(m_1)\,\mathbf{e}_k^\top(m_2)\right)$ be the probability to observe the (j,k)-th entry. Denote by $C_k = \sum_{j=1}^{m_1}\pi_{jk}$ the probability to observe an element from the k-th column and by $R_j = \sum_{k=1}^{m_2}\pi_{jk}$ the probability to observe an element from the j-th row. Note that $\max_{i,j}(C_i, R_j) \geq 1/\min(m_1, m_2) = 1/m_2$, where we assume $m_1 \geq m_2$ without loss of generality. We impose the following assumptions on the sampling distribution and error distribution.

Assumption 15 *There exists a positive constant $L \geq 1$ such that $\max_{i,j}(C_i, R_j) \leq L/m_2$.*

Assumption 16 *There exists a positive constant $\mu \geq 1$ such that $\pi_{jk} \geq (\mu m_1 m_2)^{-1}$.*

Then for any $\mathbf{\Delta} \in \mathbb{R}^{m_1\times m_2}$, $\mathbb{E}\left(\langle\mathbf{\Delta}, \mathbf{X}_i\rangle^2\right) \geq (\mu m_1 m_2)^{-1}\|\mathbf{\Delta}\|_F^2$.

Assumption 17 *The random errors ε_i's are i.i.d with density function $f(\cdot)$. Let $\zeta_{ij}^- = \varepsilon_i - \varepsilon_j$, $\zeta_{ij}^+ = \varepsilon_i + \varepsilon_j, 1 \le i \ne j \le n$. Let $f^-(\cdot)$ and $f^+(\cdot)$ denote probability density function of ζ_{ij}^- and ζ_{ij}^+ respectively. We assume the median of ζ_{ij}^+ is 0 and there exists a positive constant c_1 such that $f^-(t) \ge 1/(2c_1^2)$ and $f^+(t) \ge 1/(2c_1^2)$ for all $|t| \le 4\eta$.*

Write

$$S_n = 2\left\{n(n-1)\right\}^{-1} \left\| \sum_{i=1}^{n} a_i \mathbf{X}_i \xi_i \right\|_{\mathrm{op}},$$

where $\xi_i = 2R_i - (n+1), i = 1, \ldots, n$ and $\{R_1, R_2, \ldots, R_n\}$ follows the uniform distribution on the permutations of the integers $\{1, 2, \ldots, n\}$. We recommend to take λ equal to

$$\lambda^* = 2G_{S_n}^{-1}(1-\alpha_0),$$

where $G_{S_n}^{-1}(1-\alpha_0)$ denotes the $(1-\alpha_0)$-quantile of the distribution of S_n conditional on pseudo covariates $\{a_1\mathbf{X}_1, \ldots, a_n\mathbf{X}_n\}$. We have the following theorem.

Theorem 18 *Suppose $\mathbf{A}_0 \in \mathcal{B}_q(R_q) \cap \mathcal{S}$ and Assumptions 15-17 hold. Consider the regularization parameter $\lambda = \lambda^*$. If $m_1 + m_2 \ge (3/\alpha_0)^{1/3}$, then there exist a numerical constant C such that*

$$\begin{aligned} &\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2 \\ \le & C \max\left\{ R_q \left(c_1^2 \mu m_1 m_2 \sqrt{\frac{L\log(m_1+m_2)}{nm_2}} \right)^{2-q},\ c_1^2 \eta \mu m_1 m_2 \sqrt{\frac{\log(m_1+m_2)}{n}} \right\} \end{aligned} \tag{11}$$

with probability greater than $1 - \beta_0 - 2(m_1+m_2)^{-2}$, where $\beta_0 = \mathbb{P}\left(\lambda^* \le 2\|\nabla Q_n(\mathbf{A}_0)\|_{\mathrm{op}}\right)$.

Remark 19 *When the random errors ε_i's are symmetric, $a\mathbf{X}$ is independent of $a\varepsilon$ (see Lemma 25 in the Appendix), which implies that $\|\nabla Q_n(\mathbf{A}_0)\|_{\mathrm{op}}$ has the same distribution as S_n conditional on the pseudo covariates $\{a_1\mathbf{X}_1, \ldots, a_n\mathbf{X}_n\}$. This is in accordance with previous results. However, in general cases, the distribution of $\|\nabla Q_n(\mathbf{A}_0)\|_{\mathrm{op}}$ is unknown conditional on all the pseudo covariates $\{a_1\mathbf{X}_1, \ldots, a_n\mathbf{X}_n\}$ due to the dependence between $a\mathbf{X}$ and $a\varepsilon$, so the pivotal tuning property is no longer valid in an exact sense. Nevertheless, we conjecture that β_0 in the above theorem is close to α_0, say G_{S_n} is an approximation of $\|\nabla Q_n(\mathbf{A}_0)\|_{\mathrm{op}}$. Our simulation shows that our method is still superior to other methods with such a choice of λ in general cases.*

Remark 20 *When the random errors ε_i's are symmetric, we have $\beta_0 = \alpha_0$. If $q = 0$, R_0 becomes the rank of $\mathbf{A}_0$, and the error bound reduces to*

$$\frac{\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2}{m_1 m_2} \le C \max\left\{ c_1^4 \mu^2 L \frac{R_0 m_1 \log(m_1+m_2)}{n},\ c_1^2 \eta\mu \sqrt{\frac{\log(m_1+m_2)}{n}} \right\}.$$

This bound is of the same order as the one given in Theorems 7 and 10 of Klopp (2014), established under sub-Gaussian error assumptions. For the nearly low-rank case, if we

consider the matrix completion setting (i.e.,$n \ll m_1 m_2$), then the first term dominate the maximum in (11)*, and this rate is the same as the statistical rate of the Huber estimator and least absolute deviation estimator in Elsener and van de Geer (2018) and is only a logarithmic factor different from the minimax optimal rate established in Koltchinskii et al. (2011).*

4. Simulation

In this section, we investigate the performance of our proposed rank matrix lasso (RML) estimator through Monte Carlo simulations. The simulation results are evaluated through 100 Monte Carlo replications. Our implementation is based on a proximal gradient algorithm which can be found in Appendix G.

Example 4.1 [Matrix regression] We firstly study the matrix regression model (2). We consider two dimensions: $m_1 = m_2 = m = 40$ and $m_1 = m_2 = m = 80$. The ground truth is generated by $\mathbf{A}_0 = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U}$ is the first five eigenvector from the sample covariance matrix of 100 i.i.d $\mathcal{N}_{m_1}(0, \mathbf{I}_{m_1})$ samples, $\mathbf{V}$ is the first five eigenvectors from another sample covariance of 100 i.i.d. $\mathcal{N}_{m_2}(0, \mathbf{I}_{m_2})$ data points. The covariates are i.i.d. copies of a generic random matrix $\mathbf{X}$, which is also composed of $\mathcal{N}(0, 1)$ entries(we also consider the correlated designs with varying strengths of correlation in the Appendix H5). The random errors ε_i are sampled independently from each of the following distributions: Gaussian $\mathcal{N}(0, 0.25)$, scaled Cauchy $\mathcal{C}(0, 1)/64$, and scaled and centered log-normal $\{\mathcal{LN}(0, 9) - \exp(9/2)\}/400$. The sample size grows from 3200 to 6400.

We compare four nuclear norm penalization estimators: matrix lasso (Negahban and Wainwright, 2011), Robustified matrix lasso (Fan et al., 2021), regularized LAD (Elsener and van de Geer, 2018), and our rank matrix lasso. The tuning parameter of matrix lasso and regularized LAD are given by (8). In practice, (8) cannot be applied to the calculation of λ^* for matrix lasso and regularized LAD for that we don't know the distribution of error. For the convenience of calculation, we assume the distribution of error is known for matrix lasso and regularized LAD. The tuning parameter λ^* given by (8) is obtained by simulation based on 100 repetitions with $\alpha_0 = 0.2$. The tuning parameter of Robustified matrix lasso is given by RCV introduced in Fan et al. (2021). We use "ℓ_2", "Robust ℓ_2", "ℓ_1" and "RML" to denote the four methods, respectively. Note that though theoretical guarantee for the regularized LAD estimator was investigated only under the matrix completion model, we still take it as a benchmark for comparison. The logarithm values of the Frobenius norm $\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F$ for those estimators are presented in Figure 1.

All the robust estimators have much smaller statistical errors and sharper estimation results than the matrix lasso estimator under the heavy-tailed errors such as Cauchy and log-normal distribution. In particular, our RML outperforms other methods in Cauchy and log-normal cases and guarantee nearly the same performance compared with the best estimator in Gaussian case. This suggests that it is adaptive to a wide range of populations and capable of yielding a better trade-off between robustness and estimation accuracy.

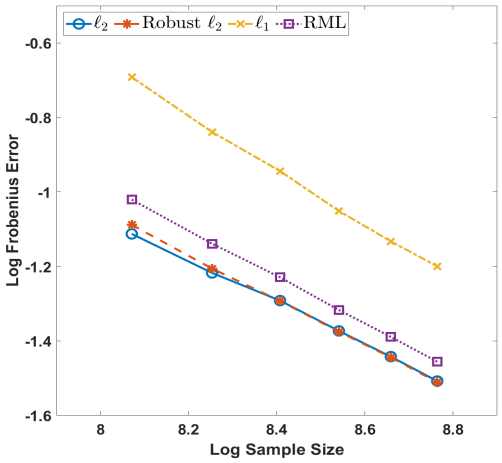
In Figure 2, under the same setting as Figure 1, we compare our method with the state-of-art alternate projection method called Scaled Gradient Descent (SGD) in Tong et al. (2021a) and Scaled Subgradient Methods (SSM) in Tong et al. (2021b). The SGD and

SSM consider the ℓ_2 loss and ℓ_1 loss under matrix factorization framework, respectively. For the two methods, we consider three specifications on rank, $R = 3, 5, 10$ which represent the underestimation, perfect specification and overestimation respectively. We also consider using cross validation to select rank R. It can be seen that the performances of SGD and SSM are sensitive to the choice of pre-specified rank R. Our method is either the best or is close to the best. The cross validation criterion lacks a theoretical guarantee, which is also a manifestation of the usefulness of our method.

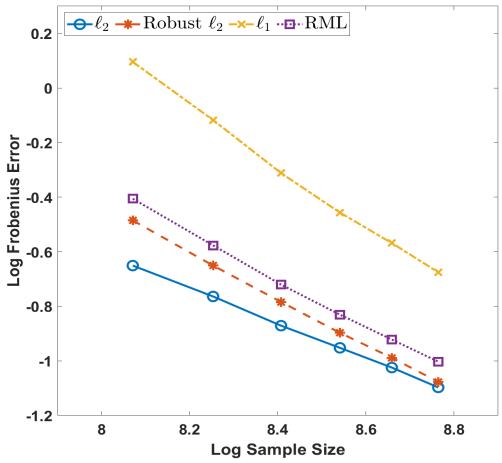
Example 4.2 [Multivariate regression] We consider the multivariate regression model (3). Here we consider $m_1 = m_2 = m = 40$ and $r = 5$, the sample size ranges from 500 to 2000. The ground truth $\mathbf{A}_0$ is generated in the same way as Example 4.1, but this example takes different covariate designs into account. As for the covariates $\mathbf{x}_i$, we take i.i.d. draws from a multivariate normal $\mathcal{N}_m(\mathbf{0}, \mathbf{\Sigma})$. Two choices of $\mathbf{\Sigma}$ are considered: (1)Identity covariance. $\mathbf{\Sigma} = \mathbf{I}_m$, which gives i.i.d. normal components for the random vectors. In this case our Assumption 12 is met with $\kappa_0 = 1$ and $b_3 = 1$. (2) Autoregressive covariance. $\mathbf{\Sigma} = (a_{ij}), a_{ij} = 0.8^{|i-j|}$. This generates the elements of $\mathbf{x}_i$ from an AR(1) model with the coefficient fixed at 0.8. According to Grenander and Szegö (1958), in this case Assumption 12 is satisfied with $\kappa_0 = 1/9$ and $b_3 = 9$. We compare four methods and numerical results are presented in Figure 3. The proposed rank matrix lasso method shows a quite competitive performance within the four candidates and yields sharp accuracy in estimating the ground truth under both designs.

In addition, Figure 4 investigated the effect of heavy-tailed errors. Two settings on the matrix dimension are considered: (1) Lower dimension: $m_1 = m_2 = m = 40$ and $n = 200$; (2) Higher dimension: $m_1 = 40$, $m_2 = 80$, and $n = 60$. The ground truth $\mathbf{A}_0$ is generated in the same way as Example 4.1. The covariates $\mathbf{x}_i$ are drawn from a $\mathcal{N}_{m_2}(\mathbf{0}, \mathbf{I}_{m_2})$ distribution, and we simulate a sequence of independent noise vector $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{m_1}$, for which each component comes from a t distribution with a degree of freedom d. Here d is given by $3k, k = 1, \ldots, 6$. Figure 4 shows the results after averaging over 100 simulation runs. The performance curves of the aforementioned estimators are presented along with the degrees of freedom of the t distributions. Generally as the t distribution approaches the normal, better estimation accuracy can be achieved by all four estimators. When the noise has a relatively heavy tail (small d), ℓ_2 loss yields a poor performance, while ℓ_1 loss and the rank matrix lasso could result in a remarkable improvement. However, for the case of large d, the performances of ℓ_1-based method would be largely compromised. In contrast, our proposed estimator still remains as competitive as ℓ_2-type methods. Thus, the rank matrix lasso achieves a good balance between robustness and estimation accuracy.

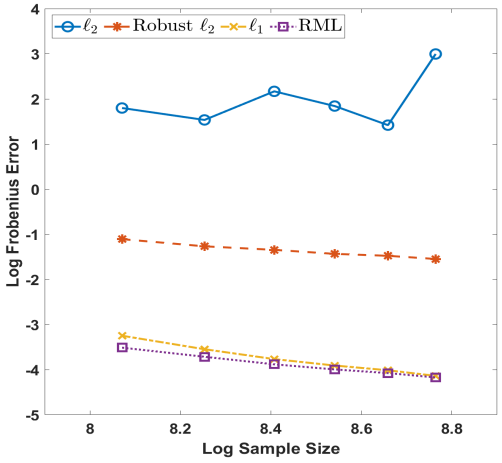
Example 4.3 [Matrix completion] We study the matrix completion model in Example 3. $\mathbf{A}_0$ is generated similarly as before but with an additional step of normalization (divided by $\sqrt{5}$) such that its Forbenius norm equals 1. The masks are i.i.d. samples from all the unit matrices, rescaled by multiplying $\sqrt{m_1 m_2}$ such that the signal-to-noise ratio of the model remains a constant as the dimension grows. Again, we take $m_1 = m_2 = m \in \{40, 80\}$. The random error takes the same form as the matrix regression example in Example 4.1, and the sample size ranges from 3200 to 6400. The results are presented in Figure 5. The numerical results fully demonstrate the satisfactory performances of our rank matrix lasso estimators regardless of the dimension, sample size and random error, especially for log-normal random



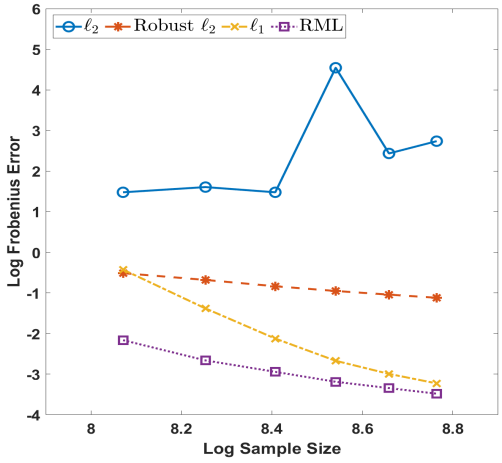
(a) Gaussian, $m = 40$



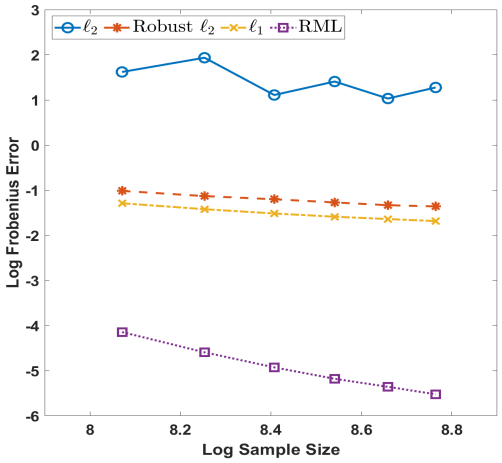
(b) Gaussian, $m = 80$



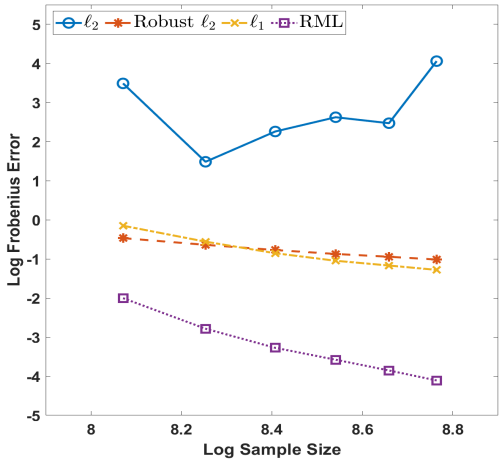
(c) Cauchy, $m = 40$



(d) Cauchy, $m = 80$

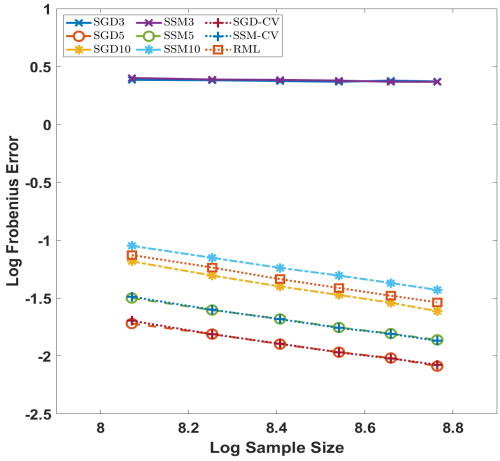


(e) Log-normal, $m = 40$

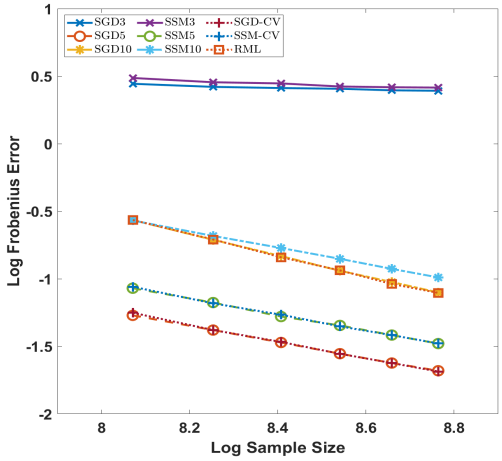


(f) Log-normal, $m = 80$

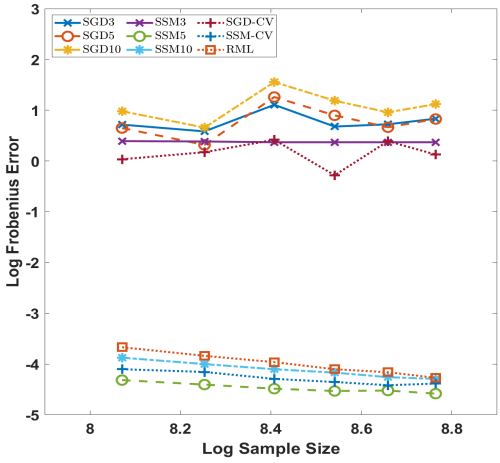
Figure 1: Log Frobenius Errors of different estimators for matrix regression model



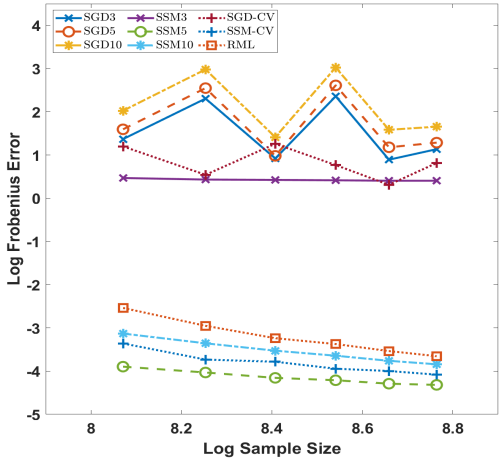
(a) Gaussian, $m = 40$



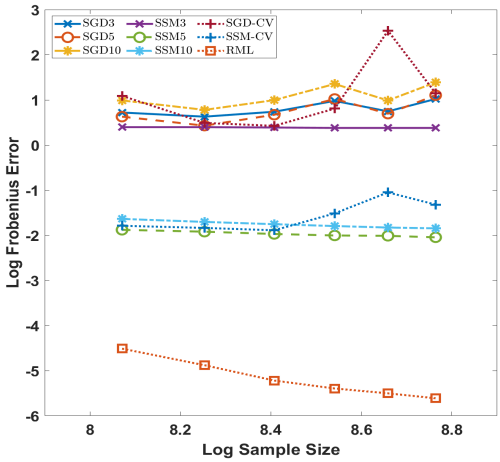
(b) Gaussian, $m = 80$



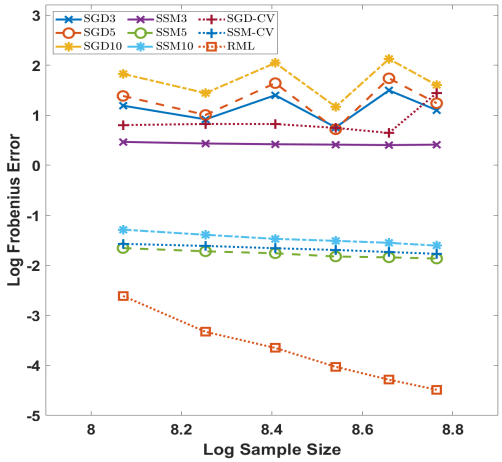
(c) Cauchy, $m = 40$



(d) Cauchy, $m = 80$

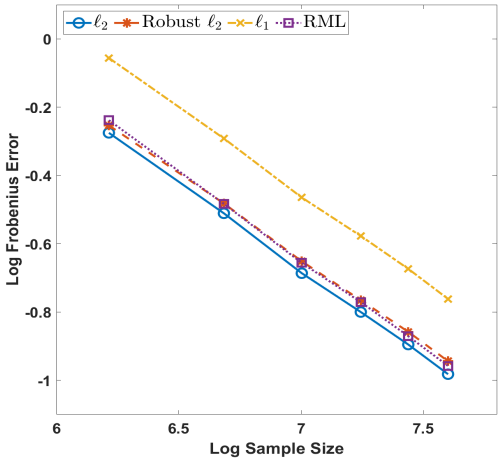


(e) Log-normal, $m = 40$

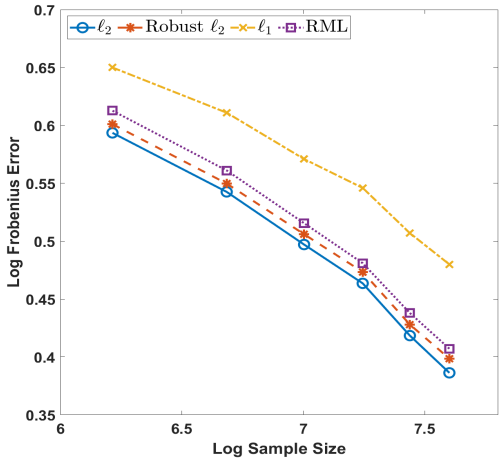


(f) Log-normal, $m = 80$

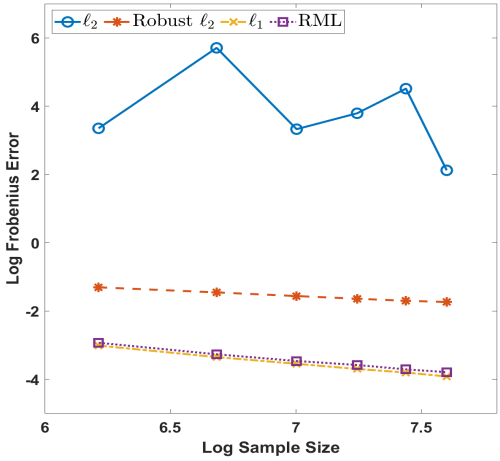
Figure 2: Log Frobenius Errors of different estimators for matrix regression model



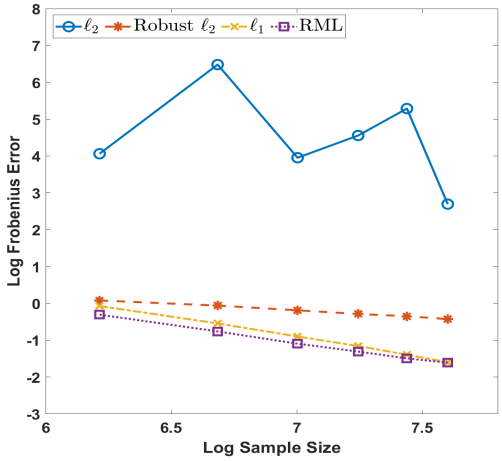
(a) Gaussian noise with identity covariance



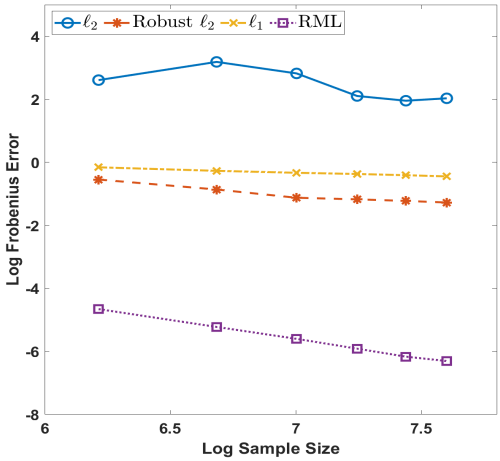
(b) Gaussian noise with AR covariance



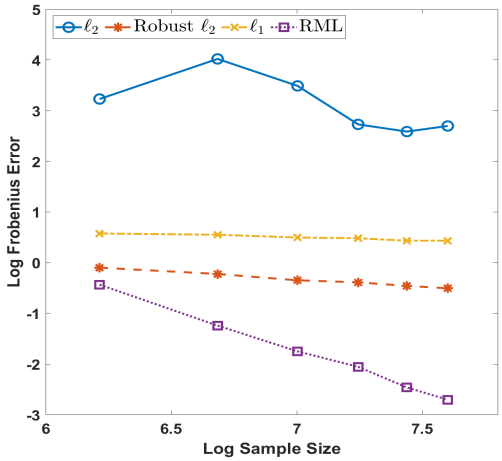
(c) Cauchy noise with identity covariance



(d) Cauchy noise with AR covariance

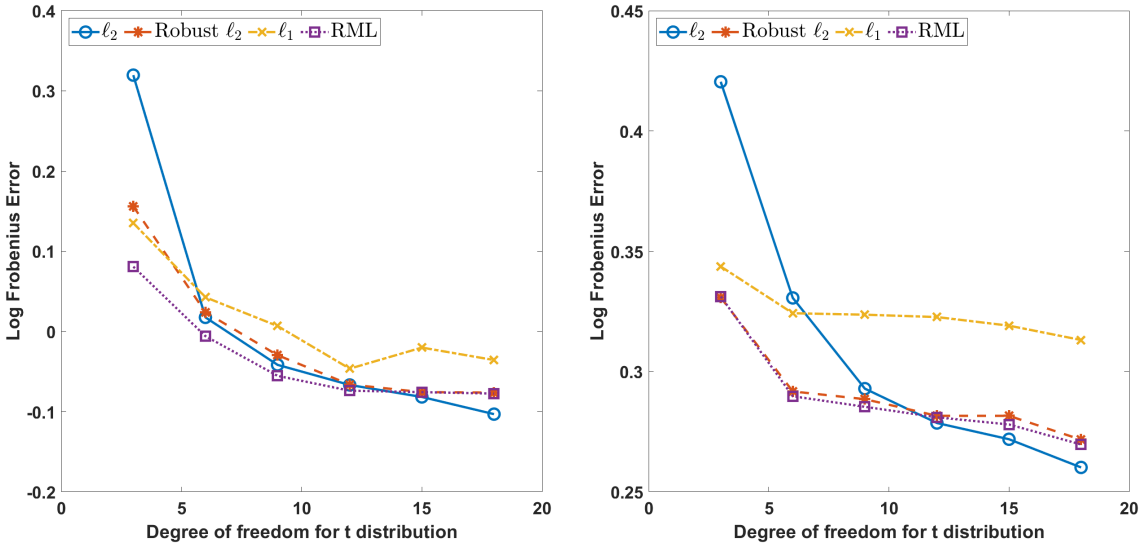


(e) Log-normal noise with identity covariance



(f) Log-normal noise with AR covariance

Figure 3: Log Frobenius Errors of different estimators for multivariate regression model



(a) Case (1): $m_1 = m_2 = m = 40$, $n = 200$ (b) Case (2): $m_1 = 40$, $m_2 = 80$, $n = 60$

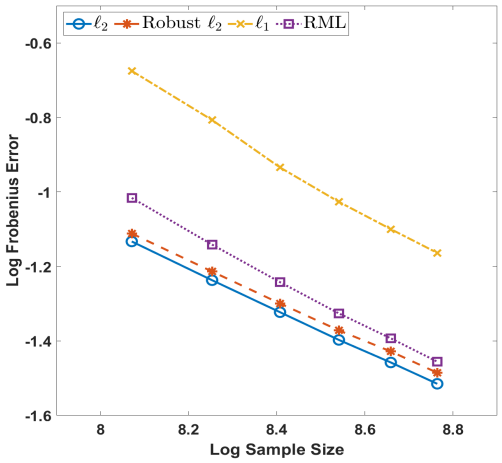
Figure 4: Log Frobenius Errors of different estimators for multivariate regression model under t noises with varying degrees of freedom

errors. The rank matrix lasso can not only be robust to heavy-tailed random errors but also perform similarly as matrix lasso under Gaussian random errors.

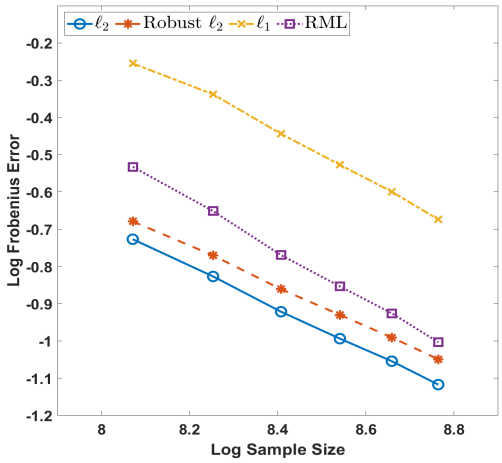
Moreover, the computational merit of our rank matrix lasso is that it is tuning-easy by using a simulated tuning parameter. It overcomes the challenge of tuning parameter selection and substantially saves the computation time. Next, we demonstrate the superiority of pivotal tuning using simulations. We consider the aforementioned three models, (I) matrix regression model with the same settings as Example 4.1 when $m_1 = m_2 = 40$ and $n = 3200$, (II) multivariate regression model with the same settings as Example 4.2 when $\mathbf{\Sigma} = \mathbf{I}_m$, $m_1 = m_2 = 40$ and $n = 500$, and (III) matrix completion with the same settings as Example 4.3 when $m_1 = m_2 = 40$ and $n = 3200$. To save the space, we only consider Fan et al. (2021)'s Robustified matrix lasso (Robust ℓ_2) for comparison, where the robust cross-validation (CV) method is used for the tuning parameter selection. Table 1 summarizes the results including the estimation error $\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2$, both tuning and solving computation time for different estimators. Clearly, pivotal tuning can remarkably reduce the burden of parameter tuning than the CV-based methods without scarifying estimation accuracy.

5. Real Data Analysis

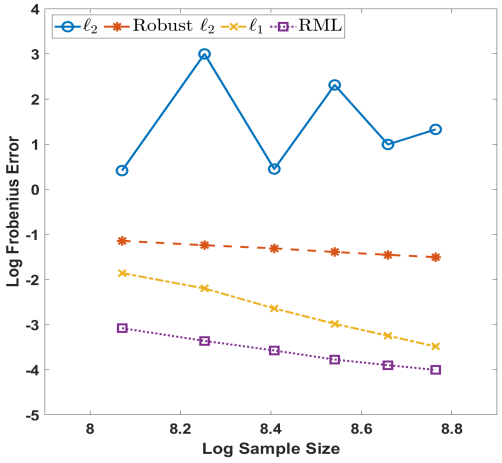
This section is devoted to a numerical study based on the well-known Arabidopsis thaliana data, which monitors the expression levels of a group of genes contributing to the generation of isoprenoids under different experimental conditions. See Wille et al. (2004) and She and



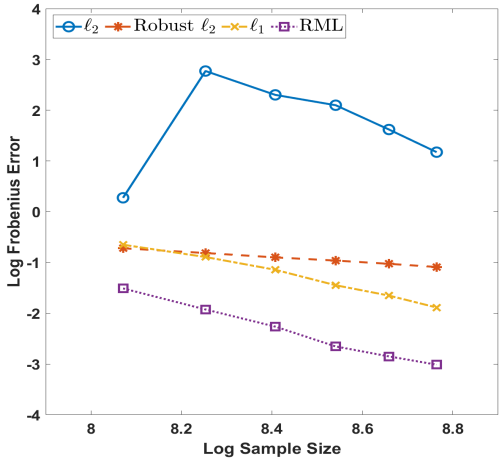
(a) Gaussian, $m = 40$



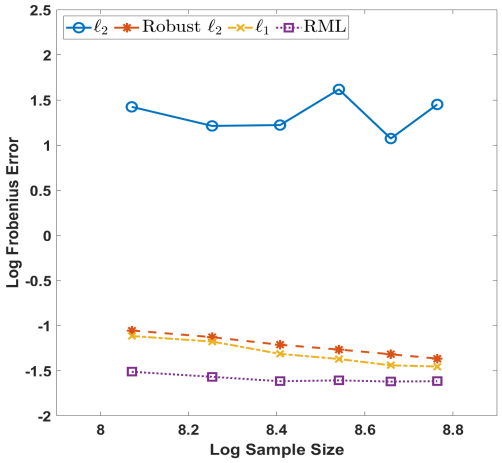
(b) Gaussian, $m = 80$



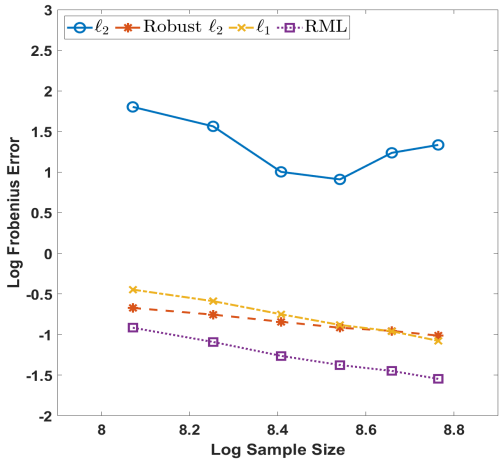
(c) Cauchy, $m = 40$



(d) Cauchy, $m = 80$



(e) Log-normal, $m = 40$



(f) Log-normal, $m = 80$

Figure 5: Log Frobenius Errors of different estimators for matrix completion model

Model	Error	Estimator	$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2$	Rank	Tuning(s)	Solving(s)	Total(s)
(I)	Normal	Robust ℓ_2	0.114(0.007)	5(0.00)	343	0.88	344
		RML	0.130(0.009)	5(0.00)	0.60	1.37	1.97
	Cauchy	Robust ℓ_2	0.105(0.007)	12.11(2.91)	314	0.89	315
		RML	<0.001	5(0.00)	0.61	2.38	2.98
	Log-normal	Robust ℓ_2	0.131(0.012)	14.17(3.26)	389	0.85	390
		RML	<0.001	5(0.00)	0.63	3.80	4.43
(II)	Normal	Robust ℓ_2	0.611(0.035)	5(0.00)	3.88	0.03	3.91
		RML	0.632(0.039)	5(0.00)	0.04	0.06	0.10
	Cauchy	Robust ℓ_2	0.076(0.008)	10.27(2.86)	4.45	0.03	4.48
		RML	0.003(0.000)	5(0.00)	0.04	0.08	0.12
	Log-normal	Robust ℓ_2	0.323(0.074)	21.83(4.64)	4.88	0.03	4.91
		RML	<0.001	5(0.00)	0.04	0.12	0.16
(III)	Normal	Robust ℓ_2	0.107(0.007)	5(0.00)	367	2.94	370
		RML	0.129(0.011)	5(0.00)	0.47	3.19	3.66
	Cauchy	Robust ℓ_2	0.103(0.007)	13.69(2.57)	342	1.83	344
		RML	0.002(0.000)	5(0.00)	0.45	3.22	3.67
	Log-normal	Robust ℓ_2	0.121(0.010)	15.52(3.37)	338	1.81	340
		RML	0.048(0.006)	5.07(0.25)	0.46	3.16	3.62

Table 1: The comparison of estimation accuracy and computation time. (I): matrix regression model; (II): multivariate regression model; (III): matrix completion

Chen (2017) for detailed description. The study contains $n = 118$ GeneChip microarray records, with the expression levels of $m_2 = 39$ genes from two upstream isoprenoid biosynthesis pathways (mevalonate and non-mevalonate) and $m_1 = 62$ genes from downstream pathways (plastoquinone, carotenoid, phytosterol and chlorophyll). We consider a multivariate regression model for the data, using genes from upstream pathways as predictors and the downstream genes as responses.

Here for the sake of comparison, we again consider the four estimators mentioned in Section 4, trained over 80% of the data, $\mathbf{Y}_{train}$, and calculate the prediction accuracy based on the remaining data serving as a test set, $\mathbf{Y}_{test}$. Concretely speaking, the accuracy for the prediction $\mathbf{Y}_{pre}$ is measured by two prediction errors, mean absolute deviation (MAD) and mean square error (MSE), as follows,

$$\text{MAD} = \frac{1}{m_1 n_{test}} \|\mathbf{Y}_{pre} - \mathbf{Y}_{test}\|_{1,1}, \ \text{MSE} = \frac{1}{m_1 n_{test}} \|\mathbf{Y}_{pre} - \mathbf{Y}_{test}\|_F^2.$$

Here $\|\cdot\|_{1,1}$ simply gives the summation of the absolute values of all the entries for a given matrix. For matrix lasso, regularized LAD and our RML we apply the pivotal tuning procedure with $\alpha = 0.2$(for matrix lasso, we simply assume the error follows standard normal distribution), and for Robustified matrix lasso, we determine the tuning parameter using robust cross validation (Fan et al., 2021). We repeat the splitting step 100 times

and report the average prediction error and the standard error. The result is summarized in Table 2. We observe from Table 2 that the matrix lasso shows the highest prediction error. In contrast, the RML has consistently lower prediction errors than other methods. Compared to Robustified matrix lasso, the RML produced estimators with smaller rank and end up with a more parsimonious model, which is beneficial for some follow-up analysis such as principal component analysis or exploratory factor analysis (EFA).

Method	ℓ_2	Robust-ℓ_2	ℓ_1	RML
MSE	0.609(0.079)	0.574(0.075)	0.581(0.077)	0.567(0.072)
MAD	0.560(0.026)	0.543(0.028)	0.539(0.025)	0.531(0.023)
Estimated rank	3.76(0.42)	8.6(0.84)	3.16(0.36)	4(0.20)

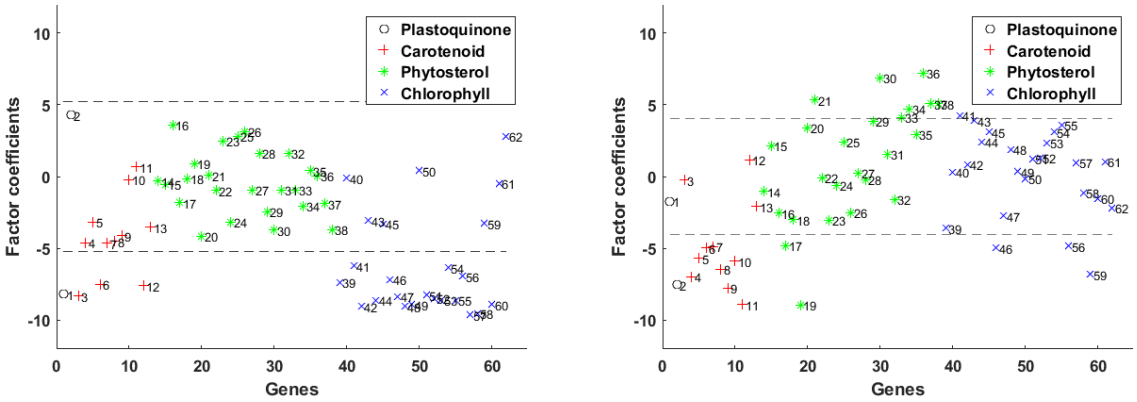
Table 2: Prediction accuracy for the Arabidopsis thaliana generic data

Next, we perform an EFA based on the RML estimators. The analysis follows a similar manner to She and Chen (2017) which conducted a factor analysis following their additive model setup and robust reduced rank regression estimation results. Let the predicted response be $\mathbf{Y}_{pre} = \mathbf{X}\widehat{\mathbf{A}}^\top \in \mathbb{R}^{n\times m_1}$ with singular value decomposition $\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^\top$. Then $\widehat{\mathbf{U}}$ collects five underlying factors, and $\widehat{\mathbf{V}}\widehat{\mathbf{D}}$ records the factor loadings (coefficients) of the 62 genes in the four downstream pathways. We plot the coefficients corresponding to the first two factors in Figure 6. To identify the most significant genes for each factor, we take a same cut-off value as She and Chen (2017), which is given by $\pm m_1^{-1/2}\hat{d}_k$ for the k-th factor. Here $\hat{d}_k$ is the k-th singular value of $\widehat{\mathbf{Y}}_{pre}$, given by the k-th diagonal element of $\widehat{\mathbf{D}}$.

Basically the factor analysis results can unveil some structural information beneath the target genes. We can see the first factor captures some joint characteristics of carotenoid and chlorophyll, and the second factor differentiates the influence of carotenoid and phytosterol, which coincide with the findings in She and Chen (2017). This reveals some group pattern in the downstream pathways that has been verified by many biological studies. For example, Trudel and Ozbun (1970) mentioned that carotenoid and chlorophyll pigments are generally interrelated, since "the two pigment systems are morphologically associated in the cell where they are attached to the same or very similar proteins in the grana or lamellae of the chloroplast", providing evidence of the group structure we summarized from the first factor loadings. It is worthy of further devotion to knit together the line of biological experiments and the insight from statistical analysis to acquire deeper understanding of these natural mechanisms.

6. Conclusion and Discussion

In this article, we study a linear operator model and propose a new rank matrix lasso method for high-dimensional low-rank matrix recovery which can tackle the challenges of tuning parameter selection for regularized estimator. For normal random errors, our estimator behaves very similarly as matrix lasso. It remains robust under heavy-tailed and skewed random errors in the sense that it possesses nearly optimal statistical error rates as other standard estimators under sub-Gaussian errors.



(a) First factor coefficients

(b) Second factor coefficients

Figure 6: Factor loadings of the 62 genes from downstream pathways for the first and second factor

We conclude this article with two remarks. Firstly, it is well known that for the linear regression model, the influence function of rank-based estimators is bounded in the response space but it is unbounded in the covariate space. Hence an outlier in the covariate space can seriously impair a rank estimate. For this aspect of robustness, Fan et al. (2021) proposed several bounded moment conditions on the design matrices and showed their shrinkage principle can achieve the minimax estimation error rate. It deserves to investigate the possibility of similar extensions in our method. Secondly, while the linear operator model has already covered a wide range of problems, it's still limited in some real-life application, especially when additional structures or restrictions are imposed on the model, such as 1-bit matrix completion (Davenport et al., 2014). Hence it is attractive to explore whether this rank based method as well as the pivotal tuning property can be adapted to more general model with other side/structure information.

Acknowledgements

The authors thank the Editor and anonymous referees for their many helpful comments that have resulted in significant improvements in the article. This research was supported by the China National Key R&D Program (Grant Nos. 2022YFA1003703, 2022YFA1003802, 2022YFA1003803), the NNSF of China Grants (Nos. 11925106, 12231011,11931001, 11971247, 71988101) and the National Statistical Science Research Grants of China (2022LD08). All authors equally contributed to this work and are listed in the alphabetical order.

Appendix

The Appendix contains the technical proofs of all Theorems and Corollaries, the associated optimization algorithm and additional simulations results. Hereafter we let $\mathbb{E}_{\mathbb{P}|\mathbf{X}}$ be the conditional expectation conditioning on all observed covariates and $\mathbb{E}_{\mathbf{X}}$ be the expectation taken with respect to covariates.

Appendix A: Proof of Theorem 1

We first give the definition of $\mathbf{\Delta}'$ and $\mathbf{\Delta}''$, which is first proposed by Negahban and Wainwright (2011). Any matrix $\mathbf{A}_0 \in \mathbb{R}^{m_1 \times m_2}$ has a singular value decomposition of the form $\mathbf{A}_0 = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m_1 \times m_1}$ and $\mathbf{V} \in \mathbb{R}^{m_2 \times m_2}$ are orthogonal matrices. For each integer $r \in \{1, 2, \ldots, m\}$, we let $\mathbf{U}^r \in \mathbb{R}^{m_1 \times r}$ and $\mathbf{V}^r \in \mathbb{R}^{m_2 \times r}$ be the sub-matrices of singular vectors associated with the top r singular values of $\mathbf{A}_0$. We then define the following two subspaces of $\mathbb{R}^{m_1 \times m_2}$:

$$\begin{aligned}
\mathcal{A}\left(\mathbf{U}^r, \mathbf{V}^r\right) :&= \left\{\mathbf{\Delta} \in \mathbb{R}^{m_1 \times m_2} \mid \operatorname{row}(\mathbf{\Delta}) \subseteq \mathbf{V}^r \text{ and } \operatorname{col}(\mathbf{\Delta}) \subseteq \mathbf{U}^r\right\} \\
\mathcal{B}\left(\mathbf{U}^r, \mathbf{V}^r\right) :&= \left\{\mathbf{\Delta} \in \mathbb{R}^{m_1 \times m_2} \mid \operatorname{row}(\mathbf{\Delta}) \perp \mathbf{V}^r \text{ and } \operatorname{col}(\mathbf{\Delta}) \perp \mathbf{U}^r\right\},
\end{aligned}$$

where $\operatorname{row}(\mathbf{\Delta}) \subseteq \mathbb{R}^{m_2}$ and $\operatorname{col}(\mathbf{\Delta}) \subseteq \mathbb{R}^{m_1}$ denote the row space and column space, respectively, of the matrix $\mathbf{\Delta}$. Let $\Pi_{\mathcal{B}}$ denote the projection operator onto the subspace $\mathcal{B}$, and define $\mathbf{\Delta}'' = \Pi_{\mathcal{B}}(\mathbf{\Delta})$ and $\mathbf{\Delta}' = \mathbf{\Delta} - \mathbf{\Delta}''$.

Before proving the main theorem, we state a useful lemma.

Lemma 21 *Let $\mathbf{U}^r \in \mathbb{R}^{m_1 \times r}$ and $\mathbf{V}^r \in \mathbb{R}^{m_2 \times r}$ be matrices consisting of the top r left and right singular vectors of $\mathbf{A}_0$, respectively. Then there exists a matrix decomposition $\mathbf{\Delta} = \mathbf{\Delta}' + \mathbf{\Delta}''$ of the error $\mathbf{\Delta} = \widehat{\mathbf{A}} - \mathbf{A}_0$ such that:*

(a) the matrix $\mathbf{\Delta}'$ satisfies $\|\mathbf{\Delta}'\|_F \leq \|\mathbf{\Delta}\|_F$ and the constraint $\operatorname{rank}(\mathbf{\Delta}') \leq 2r$;

(b) if $\lambda \geq 2\|\nabla Q_n(\mathbf{A}_0)\|_{\text{op}}$, then the nuclear norm of $\mathbf{\Delta}''$ is bounded as $\|\mathbf{\Delta}''\|_1 \leq 3\|\mathbf{\Delta}'\|_1 + 4\sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0)$.

The proof of this lemma is the same as the proof of Lemma 1 in Negahban and Wainwright (2011), thus we omit it.

Proof [Proof of Theorem 1] Using the fact that $\widehat{\mathbf{A}}$ is the minimizer of the objective function, we have

$$Q_n(\widehat{\mathbf{A}}) + \lambda\|\widehat{\mathbf{A}}\|_1 \leq Q_n(\mathbf{A}_0) + \lambda\|\mathbf{A}_0\|_1 .$$

From this and the assumption, we obtain the following inequality

$$\begin{aligned}
Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) &\leq -\Big[\{Q_n(\widehat{\mathbf{A}}) - Q(\widehat{\mathbf{A}})\} - \{Q_n(\mathbf{A}_0) - Q(\mathbf{A}_0)\}\Big] + \lambda\|\mathbf{A}_0\|_1 - \lambda\|\widehat{\mathbf{A}}\|_1 \\
&\leq \lambda_\varepsilon\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_1 + \widetilde{\lambda} + \lambda\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_1 \\
&= (\lambda_\varepsilon + \lambda)\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_1 + \widetilde{\lambda}.
\end{aligned}$$

Combining with Lemma 21, we have that

$$Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \leq (\lambda_\varepsilon + \lambda) \left\{ 4\sqrt{2r}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F + 4 \sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0) \right\} + \widetilde{\lambda}$$
$$\leq \max \left\{ 12\sqrt{2r}(\lambda_\varepsilon + \lambda)\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F, 12(\lambda_\varepsilon + \lambda) \sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0), 3\widetilde{\lambda} \right\}.$$

By Lemma 21, we have $\widehat{\mathbf{A}} - \mathbf{A}_0 \in \mathcal{C}$, which implies that $Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \geq \kappa\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2$. Then it follows that

$$\kappa\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2 \leq \max \left\{ 12\sqrt{2r}(\lambda_\varepsilon + \lambda)\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F, 12(\lambda_\varepsilon + \lambda) \sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0), 3\widetilde{\lambda} \right\},$$

which further implies

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F \leq \max \left\{ \frac{24(\lambda + \lambda_\varepsilon)\sqrt{r}}{\kappa}, \left\{ \frac{12(\lambda + \lambda_\varepsilon)\sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0)}{\kappa} \right\}^{1/2}, \sqrt{\frac{3\widetilde{\lambda}}{\kappa}} \right\}.$$

For a threshold $\tau > 0$, we choose $r = \#\{j \in \{1, 2, \ldots, m\} \mid \sigma_j(\mathbf{A}_0) \geq \tau\}$. Then it follows that

$$\sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0) \leq \tau \sum_{j=r+1}^{m} \frac{\sigma_j(\mathbf{A}_0)}{\tau} \leq \tau \sum_{j=r+1}^{m} \left(\frac{\sigma_j(\mathbf{A}_0)}{\tau} \right)^q \leq \tau^{1-q} \sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0)^q \leq \tau^{1-q} R_q$$

On the other hand, $R_q \geq \sum_{j=1}^{r} \sigma_j(\mathbf{A}_0)^q \geq r\tau^q$, so $r \leq R_q \tau^{-q}$. Choose $\tau = (\lambda + \lambda_\varepsilon)/\kappa$ yields that

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F \leq \max \left\{ 24\sqrt{R_q} \left(\frac{\lambda + \lambda_\varepsilon}{\kappa} \right)^{1-q/2}, \sqrt{\frac{3\widetilde{\lambda}}{\kappa}} \right\}.$$

■

Before proceeding to the proof of Theorems 6, 14, and 18, we give an equivalent form for rank-based loss with Wilcoxon score $Q_n(\mathbf{A})$, which is useful in our proof. Denote $\sum_{i \neq j} := \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n}$, we have

$$Q_n(\mathbf{A}) = \frac{2(n+1)}{\sqrt{3}n(n-1)} \sum_{k=1}^{p} \sum_{i=1}^{n} \phi(R(\varepsilon_{ik}(\mathbf{A}))) \cdot \varepsilon_{ik}(\mathbf{A})$$
$$= \{n(n-1)\}^{-1} \sum_{k=1}^{p} \sum_{i \neq j} |\varepsilon_{ik}(\mathbf{A}) - \varepsilon_{jk}(\mathbf{A})|$$
$$= \{n(n-1)\}^{-1} \sum_{k=1}^{p} \sum_{i \neq j} \left| \left\{ y_{ik} - \mathbf{e}_k^\top(p)\, \mathfrak{X}(\mathbf{X}_i; \mathbf{A}) \right\} - \left\{ y_{jk} - \mathbf{e}_k^\top(p)\, \mathfrak{X}(\mathbf{X}_j; \mathbf{A}) \right\} \right|,$$

where $\phi(i) = \varphi(i/(n+1))$ and $\varphi(u) = \sqrt{12}(u - 1/2)$.

Appendix B: Proof of Theorem 6

Combining the following lemma with Theorem 1, we can complete the proof of Theorem 6.

Lemma 22 *Suppose that Assumptions 2-4 hold. Suppose* $\mathbf{A}_0 \in \mathcal{B}_q(R_q) \cap \mathcal{S}$. *Then we have the following conclusions.*

(a) Define for all $M > 0$

$$\mathcal{Z}_M := \sup_{\|\mathbf{A}-\mathbf{A}_0\|_1 \le M} \left|\{Q_n(\mathbf{A}) - Q(\mathbf{A})\} - \{Q_n(\mathbf{A}_0) - Q(\mathbf{A}_0)\}\right|.$$

Then there exists a universal constant $C_1 > 0$, *with* $\lambda'_\varepsilon = 8C_1\kappa_0\sqrt{\frac{m_1+m_2}{n}}$, *such that*

$$\mathcal{Z}_M \le \lambda'_\varepsilon M$$

with probability at least $1 - 2\exp(-4(m_1+m_2))$. *Moreover, let* $\lambda_\varepsilon = 2\lambda'_\varepsilon$ *and* $\widetilde{\lambda} = 2\frac{m_1+m_2}{n}\lambda'_\varepsilon$, *we have, for any* $\mathbf{A} \in \mathcal{S}$

$$\left|\{Q_n(\mathbf{A}) - Q(\mathbf{A})\} - \{Q_n(\mathbf{A}_0) - Q(\mathbf{A}_0)\}\right| \le \lambda_\varepsilon \|\mathbf{A} - \mathbf{A}_0\|_1 + \widetilde{\lambda}$$

with probability at least $1 - 2\exp(-2(m_1+m_2))$.

(b) $\lambda^* > 2\|\nabla Q_n(\mathbf{A}_0)\|_{\text{op}}$ *with probability* $1-\alpha_0$. *If* $m_1+m_2 > \ln(2/\alpha_0)$, *then there exists constant* $C_2 > 0$ *such that* $\lambda^* \le 4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$ *with probability at least* $1 - \exp(-(m_1+m_2))$.

(c) $\mathbf{A}_0$ *is the minimizer of* $Q(\mathbf{A})$. *Take* $\kappa = b_1b_3/2$, *if* $\frac{16\sqrt{2}b_4+16}{b_1b_3b_4^2}\left(\frac{\lambda+\lambda_\varepsilon}{\kappa}\right)^{1-q/2} R_q^{1/2} < 1$, *then we have*

$$Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \ge \frac{b_1b_3}{2}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2.$$

Remark 23 *Part (a) and (b) implies that* $(\lambda^* + \lambda_\varepsilon)^{1-q/2} R_q^{1/2} \lesssim \kappa_0 \left(\frac{m_1+m_2}{n}\right)^{1/2-q/4} R_q^{1/2}$. *A sufficient condition of the assumption in (c) is* $n \gtrsim (m_1+m_2) R_q^{2/(2-q)}$, *where the symbol* $\gtrsim$ *means that the inequality holds up to a multiplicative numerical constant.*

Proof [Proof of Lemma 22] (a) The proof of the first conclusion is based on the Markov inequality, on the symmetrization theorem (Van Der Vaart and Wellner, 1996), on the contraction theorem (Ledoux and Talagrand, 2013) and on the dual norm inequality. Write $h(\varepsilon_i, \varepsilon_j) = |(\varepsilon_i - \varepsilon_j) - \langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0\rangle| - |\varepsilon_i - \varepsilon_j|$,

$$\mathcal{Z}_M = \sup_{\|\mathbf{A}-\mathbf{A}_0\|_1 \le M} \left| \frac{1}{n(n-1)} \sum_{i \ne j} \{h(\varepsilon_i, \varepsilon_j) - \mathbb{E}[h(\varepsilon_i, \varepsilon_j)]\} \right|.$$

For any $\gamma \ge 0$, Markov's inequality implies

$$\mathbb{P}(\mathcal{Z}_M > t) = \mathbb{P}\left(e^{\gamma \mathcal{Z}_M} > e^{\gamma t}\right) \le e^{-\gamma t}\mathbb{E}e^{\gamma \mathcal{Z}_M}.$$

Let $M_n = \lfloor n/2 \rfloor$, the largest integer that is no larger than $n/2$. Let $\sigma_i, i = 1, \ldots, n$ denote a Rademacher sequence independent of $\{\mathbf{X}_i, \varepsilon_i\}_{i=1}^n$. In the following, we denote $\Gamma = \{\mathbf{A} \in \mathbb{R}^{m_1 \times m_2} \mid \|\mathbf{A} - \mathbf{A}_0\|_1 \leq M\}$. By Lemma A.1 of Clémençon et al. (2008) and the convexity of $\exp(x)$, we have

$$\mathbb{E}e^{\gamma \mathcal{Z}_M} \leq \frac{1}{n!} \sum_{\pi} \mathbb{E} \exp\left(\gamma \sup_{\mathbf{A} \in \Gamma} \left| M_n^{-1} \sum_{i=1}^{M_n} \left\{ h\left(\varepsilon_{\pi(i)}, \varepsilon_{\pi(M_n+i)}\right) - \mathbb{E}h\left(\varepsilon_{\pi(i)}, \varepsilon_{\pi(M_n+i)}\right) \right\} \right| \right),$$

where the first summation taking over all permutations π of $\{1, \ldots, n\}$. Applying the symmetrization theorem and the contraction theorem, we obtain

$$\mathbb{E}e^{\gamma \mathcal{Z}_M} \leq \frac{1}{n!} \sum_{\pi} \mathbb{E} \exp\left(4\gamma \sup_{\mathbf{A} \in \Gamma} \left| M_n^{-1} \sum_{i=1}^{M_n} \sigma_i \left\langle \mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)}, \mathbf{A} - \mathbf{A}_0 \right\rangle \right| \right).$$

Then by the dual norm inequality we have

$$\begin{aligned}
\mathbb{E}e^{\gamma \mathcal{Z}_M} &\leq \frac{1}{n!} \sum_{\pi} \mathbb{E} \exp\left(4\gamma \sup_{\mathbf{A} \in \Gamma} \|\mathbf{A} - \mathbf{A}_0\|_1 \left\| M_n^{-1} \sum_{i=1}^{M_n} \sigma_i \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)} \right) \right\|_{\mathrm{op}} \right) \\
&\leq \frac{1}{n!} \sum_{\pi} \mathbb{E} \exp\left(4\gamma M \left\| M_n^{-1} \sum_{i=1}^{M_n} \sigma_i \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)} \right) \right\|_{\mathrm{op}} \right).
\end{aligned}$$

Next we adopt the ε-net argument in the Chapter 4 of Vershynin (2018) to bound the last term. Choose $\varepsilon = 1/4$, by Corollary 4.2.13 of Vershynin (2018), we can find an ε-net $\mathcal{N}$ of the sphere $\mathcal{D}^{m_1 - 1}$ and ε-net $\mathcal{M}$ of the sphere $\mathcal{D}^{m_2 - 1}$ with cardinalities $|\mathcal{N}| \leq 9^{m_1}$ and $|\mathcal{M}| \leq 9^{m_2}$ such that for any $\mathbf{A} \in \mathbb{R}^{m_1 \times m_2}$, the operator norm of $\mathbf{A}$ can be bounded using these nets as follows: $\|\mathbf{A}\|_{\mathrm{op}} \leq 2 \max_{u \in \mathcal{N}, v \in \mathcal{M}} u^\top \mathbf{A} v$. Then we have

$$\begin{aligned}
&\mathbb{E} \exp\left(4\gamma M \left\| M_n^{-1} \sum_{i=1}^{M_n} \sigma_i \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)} \right) \right\|_{\mathrm{op}} \right) \\
&\leq \mathbb{E} \exp\left(8\gamma M \max_{u \in \mathcal{N}, v \in \mathcal{M}} u^\top \left\{ M_n^{-1} \sum_{i=1}^{M_n} \sigma_i \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)} \right) \right\} v \right) \\
&\leq 2 \cdot 9^{m_1+m_2} \max_{u \in \mathcal{N}, v \in \mathcal{M}} \mathbb{E} \exp\left(8\gamma M u^\top \left\{ M_n^{-1} \sum_{i=1}^{M_n} \sigma_i \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)} \right) \right\} v \right) \\
&= 2 \cdot 9^{m_1+m_2} \max_{u \in \mathcal{N}, v \in \mathcal{M}} \mathbb{E} \exp\left(8\gamma M M_n^{-1} \sum_{i=1}^{M_n} u^\top \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)} \right) v \right) \\
&\leq 2 \cdot 9^{m_1+m_2} \exp\left(\frac{C_1^2 \kappa_0^2 \gamma^2 M^2}{n} \right),
\end{aligned}$$

where the second inequality follows from the simple bound

$$\mathbb{E}\left[\max_{1 \leq j \leq p} e^{|z_j|} \right] \leq p \max_{1 \leq j \leq p} \mathbb{E}\left[e^{|z_j|} \right] \leq p \max_{1 \leq j \leq p} \mathbb{E}\left[e^{z_j} + e^{-z_j} \right] \leq 2p \max_{1 \leq j \leq p} \mathbb{E}\left[e^{z_j} \right]$$

holding for symmetric random variables z_j, and the last inequality follows the sub-Gaussian assumption on the covariate. Above discussion implies

$$\begin{aligned}\mathbb{P}\left(\mathcal{Z}_M > t\right) &\leq \inf_{\gamma\geq 0} e^{-\gamma t}\mathbb{E}e^{\gamma \mathcal{Z}_M} \\ &\leq \inf_{\gamma\geq 0} 2\cdot 9^{m_1+m_2}\exp\left(C_1^2\kappa_0^2\gamma^2 M^2/n - \gamma t\right) \\ &= 2\cdot 9^{m_1+m_2}\exp\left(-\frac{nt^2}{4C_1^2\kappa_0^2M^2}\right).\end{aligned}$$

Let $t = 8C_1\kappa_0\sqrt{\frac{m_1+m_2}{n}}M$, then we conclude that $\mathcal{Z}_M \leq 8C_1\kappa_0\sqrt{\frac{m_1+m_2}{n}}M$ with probability at least $1-2\exp\left(-4\left(m_1+m_2\right)\right)$.

For the second conclusion, to obtain a uniform bound for all $\mathbf{A}\in\mathcal{S}$, we apply the first conclusion and peeling device given in Van de Geer (2000) and Elsener and van de Geer (2018). Without loss of generality we assume $m_1 \geq m_2$, then we can subdivide the set $\mathcal{S}$ as follows

$$\mathcal{S} = \left\{\mathbf{A}\in\mathcal{S} : \|\mathbf{A}-\mathbf{A}_0\|_1 \leq \frac{m_1+m_2}{n}\right\}\bigcup\left\{\mathbf{A}\in\mathcal{S} : \frac{m_1+m_2}{n} < \|\mathbf{A}-\mathbf{A}_0\|_1 \leq 2\sqrt{m_1^2m_2}\eta\right\}.$$

On the first set $\mathcal{S}_0 = \left\{\mathbf{A}\in\mathcal{S} : \|\mathbf{A}-\mathbf{A}_0\|_1 \leq \frac{m_1+m_2}{n}\right\}$

$$\begin{aligned}&\mathbb{P}\left(\exists\mathbf{A}\in\mathcal{S}_0 : \left|\left\{Q_n\left(\mathbf{A}\right)-Q\left(\mathbf{A}\right)\right\}-\left\{Q_n\left(\mathbf{A}_0\right)-Q\left(\mathbf{A}_0\right)\right\}\right| > \lambda_\varepsilon\left\|\mathbf{A}-\mathbf{A}_0\right\|_1 + \widetilde{\lambda}\right) \\ &\leq \mathbb{P}\left(\exists\mathbf{A}\in\mathcal{S}_0 : \left|\left\{Q_n\left(\mathbf{A}\right)-Q\left(\mathbf{A}\right)\right\}-\left\{Q_n\left(\mathbf{A}_0\right)-Q\left(\mathbf{A}_0\right)\right\}\right| > \frac{m_1+m_2}{n}\lambda'_\varepsilon\right) \\ &\leq 2\exp\left(-4\left(m_1+m_2\right)\right).\end{aligned}$$

We further subdivide the second set $\left\{\mathbf{A}\in\mathcal{S} : \frac{m_1+m_2}{n} < \|\mathbf{A}-\mathbf{A}_0\|_1 \leq 2\sqrt{m_1^2m_2}\eta\right\}$ into

$$\begin{aligned}&\left\{\mathbf{A}\in\mathcal{S} : \frac{m_1+m_2}{n} < \|\mathbf{A}-\mathbf{A}_0\|_1 \leq 2\sqrt{m_1^2m_2}\eta\right\} \\ &\subset \bigcup_{k=1}^{k_0}\left\{\mathbf{A}\in\mathcal{S} : \frac{m_1+m_2}{n}2^k < \|\mathbf{A}-\mathbf{A}_0\|_1 \leq \frac{m_1+m_2}{n}2^{k+1}\right\} := \bigcup_{k=1}^{k_0}\mathcal{S}_k,\end{aligned}$$

where k_0 is the smallest integer such that $\frac{m_1+m_2}{n}2^{k_0+1} \geq 2\sqrt{m_1^2m_2}\eta$. For each $\mathcal{S}_k$

$$\begin{aligned}&\mathbb{P}\left(\exists\mathbf{A}\in\mathcal{S}_k : \left|\left\{Q_n\left(\mathbf{A}\right)-Q\left(\mathbf{A}\right)\right\}-\left\{Q_n\left(\mathbf{A}_0\right)-Q\left(\mathbf{A}_0\right)\right\}\right| > \lambda_\varepsilon\left\|\mathbf{A}-\mathbf{A}_0\right\|_1 + \widetilde{\lambda}\right) \\ &\leq \mathbb{P}\left(\exists\mathbf{A}\in\mathcal{S}_k : \left|\left\{Q_n\left(\mathbf{A}\right)-Q\left(\mathbf{A}\right)\right\}-\left\{Q_n\left(\mathbf{A}_0\right)-Q\left(\mathbf{A}_0\right)\right\}\right| > \frac{m_1+m_2}{n}2^{k+1}\lambda'_\varepsilon\right) \\ &\leq 2\exp\left(-4\left(m_1+m_2\right)\right).\end{aligned}$$

Note that $k_0 = \log_2\left(\frac{n\sqrt{m_1^2m_2}\eta}{m_1+m_2}\right)$, by the high-dimensional setting $n < m_1m_2$, there exists a constant c such that $(k_0+1)\exp\left(-2\left(m_1+m_2\right)\right) \leq 1$ when $m_1+m_2 > c$. Then by the

union bound

$$\begin{aligned}
&\mathbb{P}\left(\exists \mathbf{A} \in \mathcal{S} : |\{Q_n(\mathbf{A}) - Q(\mathbf{A})\} - \{Q_n(\mathbf{A}_0) - Q(\mathbf{A}_0)\}| > \lambda_\varepsilon \|\mathbf{A} - \mathbf{A}_0\|_1 + \widetilde{\lambda}\right) \\
&\leq \sum_{k=0}^{k_0} \mathbb{P}\left(\exists \mathbf{A} \in \mathcal{S}_k : |\{Q_n(\mathbf{A}) - Q(\mathbf{A})\} - \{Q_n(\mathbf{A}_0) - Q(\mathbf{A}_0)\}| > \lambda_\varepsilon \|\mathbf{A} - \mathbf{A}_0\|_1 + \widetilde{\lambda}\right) \\
&\leq 2(k_0+1)\exp\left(-4(m_1+m_2)\right) \leq 2\exp\left(-2(m_1+m_2)\right),
\end{aligned}$$

which completes the proof of the second conclusion.

(b) The first conclusion follows from the definition of λ^*. Next, we give the proof of the second conclusion. Conditional on $\{\mathbf{X}_i\}_{i=1}^n$. For any $\gamma \geq 0$, Markov's inequality implies

$$\mathbb{P}(S_n > t \mid \mathbf{X}) = \mathbb{P}\left(e^{\gamma S_n} > e^{\gamma t} \mid \mathbf{X}\right) \leq e^{-\gamma t}\mathbb{E}_{\mathbb{P}|\mathbf{X}} e^{\gamma S_n}.$$

Recall that $S_n = \{n(n-1)\}^{-1} \| \sum_{i \neq j} (\mathbf{X}_j - \mathbf{X}_i) \operatorname{sign}(\varepsilon_i - \varepsilon_j) \|_{\mathrm{op}}$, then using the ε-net, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}|\mathbf{X}} e^{\gamma S_n} \\
&\leq 2 \cdot 9^{m_1+m_2} \max_{u \in \mathcal{N}, v \in \mathcal{M}} \mathbb{E}_{\mathbb{P}|\mathbf{X}} \exp\left(2\gamma \{n(n-1)\}^{-1} \sum_{i \neq j} u^\top (\mathbf{X}_j - \mathbf{X}_i) v \operatorname{sign}(\varepsilon_i - \varepsilon_j)\right) \\
&\leq 2 \cdot 9^{m_1+m_2} \max_{u \in \mathcal{N}, v \in \mathcal{M}} \mathbb{E}_{\mathbb{P}|\mathbf{X}} \exp\left(2\gamma \frac{1}{n!} \sum_{\pi} M_n^{-1} \sum_{i=1}^{M_n} u^\top \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)}\right) v \operatorname{sign}\left(\varepsilon_{\pi(i)} - \varepsilon_{\pi(M_n+i)}\right)\right) \\
&\leq \frac{1}{n!} \sum_{\pi} 2 \cdot 9^{m_1+m_2} \max_{u \in \mathcal{N}, v \in \mathcal{M}} \mathbb{E}_{\mathbb{P}|\mathbf{X}} \exp\left(2\gamma M_n^{-1} \sum_{i=1}^{M_n} u^\top \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)}\right) v \operatorname{sign}\left(\varepsilon_{\pi(i)} - \varepsilon_{\pi(M_n+i)}\right)\right) \\
&\leq \frac{1}{n!} \sum_{\pi} 2 \cdot 9^{m_1+m_2} \max_{u \in \mathcal{N}, v \in \mathcal{M}} \exp\left(4\frac{\gamma^2}{M_n^2} \sum_{i=1}^{M_n} \frac{\left\{u^\top \left(\mathbf{X}_{\pi(i)} - \mathbf{X}_{\pi(M_n+i)}\right) v\right\}^2}{2}\right) \\
&\leq \frac{1}{n!} \sum_{\pi} 2 \cdot 9^{m_1+m_2} \exp\left(16\frac{\gamma^2}{n} \max_{u \in \mathcal{N}, v \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \left(u^\top \mathbf{X}_i v\right)^2\right).
\end{aligned}$$

Assume $16 \max_{u \in \mathcal{N}, v \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^n \left(u^\top \mathbf{X}_i v\right)^2 \leq \kappa_0^2 C_2^2$ for some constant C_2, we obtain

$$\begin{aligned}
\mathbb{P}(S_n > t \mid \mathbf{X}) &\leq \inf_{\gamma \geq 0} e^{-\gamma t} \mathbb{E}_{\mathbb{P}|\mathbf{X}} e^{\gamma S_n} \\
&\leq \inf_{\gamma \geq 0} 2 \cdot 9^{m_1+m_2} \exp\left(C_2^2 \kappa_0^2 \gamma^2 / n - \gamma t\right) \\
&= 2 \cdot 9^{m_1+m_2} \exp\left(-\frac{nt^2}{4C_2^2\kappa_0^2}\right).
\end{aligned}$$

Let $t = 4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$, it follows that $S_n \leq 4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$ with probability at least $1 - 2\exp(-(m_1+m_2))$. Then if $m_1 + m_2 > \ln(2/\alpha_0)$, we conclude that $\lambda^* < 4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$.

Finally we only need to show that there exists some constant C_2 such that

$$16 \max_{u\in\mathcal{N},v\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n}\left(u^\top \mathbf{X}_i v\right)^2 \le \kappa_0^2 C_2^2$$

holds with high probability. Under the Assumption 12, it's easy to show that $u^\top \mathbf{X}_i v/\kappa_0$ is sub-Gaussian with $\left\|u^\top \mathbf{X}_i v/\kappa_0\right\|_{\psi_2} \le 1$ and $\mathbb{E}\left(u^\top \mathbf{X}_i v/\kappa_0\right)^2 \le \mu$ for all $u \in \mathcal{N}, v \in \mathcal{M}$, where μ is a constant. By $\| \left(u^\top \mathbf{X}_i v/\kappa_0\right)^2 \|_{\psi_1} = \left(\|u^\top \mathbf{X}_i v/\kappa_0\|_{\psi_2}\right)^2$ we have that $\left(u^\top \mathbf{X}_i v/\kappa_0\right)^2$ is sub-exponential with constant parameter (ν, α) and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left(u^\top \mathbf{X}_i v/\kappa_0\right)^2 - \mathbb{E}\left(u^\top \mathbf{X}_i v/\kappa_0\right)^2\right) > t\right) \le \begin{cases} e^{-\frac{nt^2}{2\nu^2}} & \text{for } 0 \le t \le \frac{\nu^2}{\alpha} \\ e^{-\frac{nt}{2\alpha}} & \text{for } t > \frac{\nu^2}{\alpha} \end{cases}.$$

We take $t = \frac{\nu^2}{\alpha}$, then

$$\begin{aligned}\mathbb{P}\left(\max_{u\in\mathcal{N},v\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n}\left(u^\top \mathbf{X}_i v/\kappa_0\right)^2 > \mu + \frac{\nu^2}{\alpha}\right) &\le \sum_{k=1}^{9^{m_1+m_2}} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(u^\top \mathbf{X}_i v/\kappa_0\right)^2 > \mu + \frac{\nu^2}{\alpha}\right) \\ &\le 9^{m_1+m_2}\exp\left(-\frac{\nu^2}{2\alpha^2}n\right).\end{aligned}$$

Hence there exists constant C_2 and c_2 such that

$$\mathbb{P}\left(16 \max_{u\in\mathcal{N},v\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n}\left(u^\top \mathbf{X}_i v\right)^2 \le \kappa_0^2 C_2^2\right) \ge 1 - 9^{m_1+m_2}\exp\left(-c_2 n\right).$$

Note that $16\max_{u\in\mathcal{N},v\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n}\left(u^\top \mathbf{X}_i v\right)^2 \le \kappa_0^2 C_2^2$ with probability at least $1-9^{m_1+m_2}\exp\left(-c_2 n\right)$. If $n > 4\left(m_1+m_2\right)/c_2$, we obtain $\lambda^* < 4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$ with probability at least $1 - \exp\left(-\left(m_1+m_2\right)\right)$.

(c) We denote $\mathrm{Cov}\left(\mathrm{vec}\left(\mathbf{X}_1\right)\right)$ by $\mathbf{J}$. First we show that, for any $\mathbf{A}$

$$Q\left(\mathbf{A}\right) - Q\left(\mathbf{A}_0\right) \ge \frac{b_1}{2}\|\mathbf{J}^{\frac{1}{2}}\,\mathrm{vec}\left(\mathbf{A}-\mathbf{A}_0\right)\|_2^2 \wedge \frac{b_1 b_4}{2}\|\mathbf{J}^{\frac{1}{2}}\,\mathrm{vec}\left(\mathbf{A}-\mathbf{A}_0\right)\|_2$$

which implies that $\mathbf{A}_0$ is the minimizer of $Q\left(\mathbf{A}\right)$. We define the maximal radius over which the criterion function can be minorated by a quadratic function

$$r_{\mathbf{A}_0} = \sup_r\left\{r : Q\left(\mathbf{A}\right) - Q\left(\mathbf{A}_0\right) \ge \frac{b_1}{2}\|\mathbf{J}^{1/2}\,\mathrm{vec}\left(\mathbf{A}-\mathbf{A}_0\right)\|_2^2 \text{ for all } \|\mathbf{J}^{1/2}\,\mathrm{vec}\left(\mathbf{A}-\mathbf{A}_0\right)\|_2 \le r\right\}.$$

We claim that $r_{\mathbf{A}_0} \ge b_4$. By Knight's identity (Koenker, 2005),

$$\begin{aligned}Q_n\left(\mathbf{A}\right) - Q_n\left(\mathbf{A}_0\right) = &\left\{n\left(n-1\right)\right\}^{-1}\sum_{i\ne j}\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A}-\mathbf{A}_0\rangle\left\{\mathrm{I}\left(\zeta_{ij} < 0\right) - 1/2\right\} \\ &+ \left\{n\left(n-1\right)\right\}^{-1}\sum_{i\ne j}\int_0^{\langle \mathbf{X}_i-\mathbf{X}_j, \mathbf{A}-\mathbf{A}_0\rangle}\left\{\mathrm{I}\left(\zeta_{ij}\le s\right) - \mathrm{I}\left(\zeta_{ij}\le 0\right)\right\}ds.\end{aligned}$$

Denote the distribution function of ζ_{ij} by $F^*(\cdot)$. By the independence of $\mathbf{X}$ and ε, we have

$$Q(\mathbf{A}) - Q(\mathbf{A}_0) = \mathbb{E}_{\mathbf{X}} \int_0^{\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \{F^*(s) - F^*(0)\} \, ds.$$

By the Taylor expansion and Assumption 2, for some ξ_{ij} between 0 and $\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle$, we have

$$\begin{aligned} Q(\mathbf{A}) - Q(\mathbf{A}_0) &= \mathbb{E}_{\mathbf{X}} \int_0^{\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \left\{ f^*(0)\, s + \frac{\partial f^*(\xi_{ij})}{\partial t} \frac{s^2}{2} \right\} ds \\ &\geq \frac{b_1}{2} \mathbb{E}_{\mathbf{X}} \left[\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle^2 \right] - \frac{b_2}{6} \mathbb{E}_{\mathbf{X}} \left[|\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle|^3 \right]. \end{aligned}$$

Note that for all $\mathbf{A} - \mathbf{A}_0$, if

$$\|\mathbf{J}^{1/2} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2 \leq b_4 \leq \frac{3b_1}{2\sqrt{2}b_2} \inf_{\mathbf{A} - \mathbf{A}_0 \neq 0} \frac{\left(\mathbb{E} \langle \mathbf{X}_1 - \mathbf{X}_2, \mathbf{A} - \mathbf{A}_0 \rangle^2 \right)^{3/2}}{\mathbb{E} \left| \langle \mathbf{X}_1 - \mathbf{X}_2, \mathbf{A} - \mathbf{A}_0 \rangle \right|^3},$$

it follows that

$$\frac{b_1}{4} \mathbb{E}_{\mathbf{X}} \langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle^2 \geq \frac{b_2}{6} \mathbb{E}_{\mathbf{X}} \left| \langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle \right|^3.$$

Hence for any $\|\mathbf{J}^{1/2} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2 \leq b_4$, we have

$$\begin{aligned} Q(\mathbf{A}) - Q(\mathbf{A}_0) &\geq \frac{b_1}{4} \mathbb{E}_{\mathbf{X}} \left(\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle^2 \right) \\ &= \frac{b_1}{2} \|\mathbf{J}^{\frac{1}{2}} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2^2. \end{aligned}$$

This implies $r_{\mathbf{A}_0} \geq b_4$.

By construction of $r_{\mathbf{A}_0}$ and the convexity of Q, for any $\mathbf{A}$ we have

$$\begin{aligned} & Q(\mathbf{A}) - Q(\mathbf{A}_0) \\ \geq\ & \frac{b_1}{2} \|\mathbf{J}^{1/2} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2^2 \wedge \left\{ \frac{\|\mathbf{J}^{1/2} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2}{r_{\mathbf{A}_0}} \cdot \inf_{\|\mathbf{J}^{1/2} \operatorname{vec}(\widetilde{\mathbf{A}} - \mathbf{A}_0)\|_2 = r_{\mathbf{A}_0}} Q(\widetilde{\mathbf{A}}) - Q(\mathbf{A}_0) \right\} \\ \geq\ & \frac{b_1}{2} \|\mathbf{J}^{1/2} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2^2 \wedge \left\{ \frac{\|\mathbf{J}^{1/2} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2}{r_{\mathbf{A}_0}} \frac{b_1}{2} r_{\mathbf{A}_0}^2 \right\} \\ \geq\ & \frac{b_1}{2} \|\mathbf{J}^{1/2} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2^2 \wedge \left\{ \frac{b_1 b_4}{2} \|\mathbf{J}^{1/2} \operatorname{vec}(\mathbf{A} - \mathbf{A}_0)\|_2 \right\}, \end{aligned}$$

which implies $\mathbf{A}_0$ is the minimizer of $Q(\mathbf{A})$. By Assumption 3, we further have

$$Q(\mathbf{A}) - Q(\mathbf{A}_0) \geq \frac{b_1 b_3}{2} \|\mathbf{A} - \mathbf{A}_0\|_F^2 \wedge \frac{b_1 b_3 b_4}{2} \|\mathbf{A} - \mathbf{A}_0\|_F$$

for all $\mathbf{A} - \mathbf{A}_0 \in \mathcal{C}$. Using the fact that $\widehat{\mathbf{A}}$ is the minimizer of the objective function:

$$Q_n(\widehat{\mathbf{A}}) + \lambda \|\widehat{\mathbf{A}}\|_1 \ \leq Q_n(\mathbf{A}_0) + \lambda \|\mathbf{A}_0\|_1 \ .$$

Similar to the proof of Theorem 1, we obtain

$$Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \leq (\lambda_\varepsilon + \lambda) \|\widehat{\mathbf{A}} - \mathbf{A}_0\|_1 + \widetilde{\lambda}.$$

By Lemma 21, we have $\widehat{\mathbf{A}} - \mathbf{A}_0 \in \mathcal{C}$, then

$$\begin{aligned}
\frac{b_1 b_3}{2}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2 \wedge \frac{b_1 b_3 b_4}{2}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F &\leq (\lambda_\varepsilon + \lambda) \|\widehat{\mathbf{A}} - \mathbf{A}_0\|_1 + \widetilde{\lambda} \\
&\leq (\lambda_\varepsilon + \lambda) \left\{ 4\sqrt{2r}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F + 4 \sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0) \right\} + \widetilde{\lambda} \\
&\leq (\lambda_\varepsilon + \lambda) \max \left\{ 8\sqrt{2r}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F, 8 \sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0) \right\} + \widetilde{\lambda}.
\end{aligned}$$

If $\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F \leq b_4$, combining the fact that $\sqrt{\widetilde{\lambda}}$ is dominated by λ_ε when $n \gtrsim (m_1 + m_2)$, it follows that $Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \geq \frac{b_1 b_3}{2}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2$ and

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F \leq \max \left\{ \frac{32(\lambda + \lambda_\varepsilon)\sqrt{r}}{b_1 b_3}, \left\{ \frac{16(\lambda + \lambda_\varepsilon) \sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0)}{b_1 b_3} \right\}^{1/2} \right\}.$$

If $\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F \geq b_4$, then

$$\frac{b_1 b_3 b_4}{2}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F \leq 2(\lambda_\varepsilon + \lambda) \left\{ 4\sqrt{2r}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F + 4 \sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0) \right\}$$

which implies

$$1 \leq \frac{4(\lambda_\varepsilon + \lambda)}{b_1 b_3 b_4} \left\{ 4\sqrt{2r} + 4\frac{\sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0)}{\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F} \right\} \leq \frac{4(\lambda_\varepsilon + \lambda)}{b_1 b_3 b_4} \left\{ 4\sqrt{2r} + 4\frac{\sum_{j=r+1}^{m} \sigma_j(\mathbf{A}_0)}{b_4} \right\}.$$

Similar to the proof of Theorem 1, we choose $r = \#\{j \in \{1, 2, \ldots, p\} \mid \sigma_j(\mathbf{A}_0) \geq \tau\}$, then we have $\sum_{j=r+1}^{p} \sigma_j(\mathbf{A}_0) \leq \tau^{1-q} R_q$ and $r \leq R_q \tau^{-q}$. Taking $\kappa = b_1 b_3/2$ and $\tau = (\lambda + \lambda_\varepsilon)/\kappa$, this gives

$$1 \leq \left\{ \sqrt{2} + \frac{1}{b_4} \right\} \frac{16}{b_1 b_3 b_4} \left(\frac{\lambda + \lambda_\varepsilon}{\kappa} \right)^{1-q/2} R_q^{1/2} < 1$$

which is a contradiction. Hence if $\frac{16\sqrt{2}b_4 + 16}{b_1 b_3 b_4^2} \left(\frac{\lambda + \lambda_\varepsilon}{\kappa}\right)^{1-q/2} R_q^{1/2} < 1$, we have

$$Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \geq \frac{b_1 b_3}{2}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2.$$

■

Appendix C: Proof of Corollary 10

In the proof of Theorem 6, we use ε-net argument to bound $\|M_n^{-1}\sum_{i=1}^{M_n}\sigma_i\left(\mathbf{X}_{\pi(i)}-\mathbf{X}_{\pi(M_n+i)}\right)\|_{\text{op}}$ and $\|\sum_{i\neq j}\left(\mathbf{X}_i-\mathbf{X}_j\right)\text{sign}\left(\varepsilon_i-\varepsilon_j\right)\|_{\text{op}}$. But in the high dimensional linear regression model, due to the diagonal structure of $\mathbf{X}$, we bound them in the following way to get a faster rate:

$$\left\|M_n^{-1}\sum_{i=1}^{M_n}\sigma_i\left(\mathbf{X}_{\pi(i)}-\mathbf{X}_{\pi(M_n+i)}\right)\right\|_{\text{op}}\leq\max_{1\leq j\leq d}\left|M_n^{-1}\sum_{i=1}^{M_n}\sigma_i\left(\mathbf{X}_{\pi(i)jj}-\mathbf{X}_{\pi(M_n+i)jj}\right)\right|,$$

$$\left\|\sum_{i\neq j}\left(\mathbf{X}_i-\mathbf{X}_j\right)\text{sign}\left(\varepsilon_i-\varepsilon_j\right)\right\|_{\text{op}}\leq\max_{1\leq k\leq d}\left|\sum_{i\neq j}\left(\mathbf{X}_{ikk}-\mathbf{X}_{jkk}\right)\text{sign}\left(\varepsilon_i-\varepsilon_j\right)\right|,$$

where $\mathbf{X}_{\pi(i)jj}$ and $\mathbf{X}_{ikk}$ denote the j-th and k-th diagonal element of $\mathbf{X}_{\pi(i)}$ and $\mathbf{X}_i$ respectively. Then with the two inequalities, similar to the proof of Theorem 6, we can get the conclusion in Corollary 10.

Appendix D: Proof of Theorem 14

Combining the following lemma with Theorem 1, we can complete the proof of Theorem 14.

Lemma 24 *Suppose that Assumptions 11-13 hold. Suppose* $\mathbf{A}_0\in\mathcal{B}_q\left(R_q\right)\cap\mathcal{S}$. *Then we have the following conclusions.*

(a) Define for all $M>0$

$$\mathcal{Z}_M:=\sup_{\|\mathbf{A}-\mathbf{A}_0\|_1\leq M}\left|\left\{Q_n\left(\mathbf{A}\right)-Q\left(\mathbf{A}\right)\right\}-\left\{Q_n\left(\mathbf{A}_0\right)-Q\left(\mathbf{A}_0\right)\right\}\right|.$$

Then there exists a universal constant $C_1>0$, *with* $\lambda_\varepsilon'=8C_1\kappa_0\sqrt{\frac{m_1+m_2}{n}}$, *such that*

$$\mathcal{Z}_M\leq\lambda_\varepsilon' M$$

with probability at least $1-2\exp\left(-4\left(m_1+m_2\right)\right)$. *Moreover, let* $\lambda_\varepsilon=2\lambda_\varepsilon'$ *and* $\widetilde{\lambda}=2\frac{m_1+m_2}{n}\lambda_\varepsilon'$, *we have, for any* $\mathbf{A}\in\mathcal{S}$

$$\left|\left\{Q_n\left(\mathbf{A}\right)-Q\left(\mathbf{A}\right)\right\}-\left\{Q_n\left(\mathbf{A}_0\right)-Q\left(\mathbf{A}_0\right)\right\}\right|\leq\lambda_\varepsilon\left\|\mathbf{A}-\mathbf{A}_0\right\|_1+\widetilde{\lambda}$$

with probability at least $1-2\exp\left(-2\left(m_1+m_2\right)\right)$.

(b) $\lambda^*>2\left\|\nabla Q_n\left(\mathbf{A}_0\right)\right\|_{\text{op}}$ *with probability* $1-\alpha_0$. *If* $m_1+m_2>\ln\left(2/\alpha_0\right)$, *then there exists constant* $C_2>0$ *such that* $\lambda^*\leq 4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$ *with probability at least* $1-\exp\left(-\left(m_1+m_2\right)\right)$.

(c) $\mathbf{A}_0$ *is the minimizer of* $Q\left(\mathbf{A}\right)$. *Take* $\kappa=b_1b_3/2$, *if* $\frac{16\sqrt{2}b_4+16}{b_1b_3b_4^2}\left(\frac{\lambda+\lambda_\varepsilon}{\kappa}\right)^{1-q/2}R_q^{1/2}<1$, *then we have*

$$Q(\widehat{\mathbf{A}})-Q\left(\mathbf{A}_0\right)\geq\frac{b_1b_3}{2}\|\widehat{\mathbf{A}}-\mathbf{A}_0\|_F^2.$$

Proof [Proof of Lemma 24] (a) We only prove the first conclusion since the proof of the second conclusion follows similarly to that of Lemma 22(a). Write $h\left(\varepsilon_{ik}, \varepsilon_{jk}\right) = |\left(\varepsilon_{ik} - \varepsilon_{jk}\right) - \langle \mathbf{e}_k \mathbf{x}_i^\top - \mathbf{e}_k \mathbf{x}_j^\top, \mathbf{A} - \mathbf{A}_0 \rangle| - |\varepsilon_{ik} - \varepsilon_{jk}|$,

$$\mathcal{Z}_M = \sup_{\|\mathbf{A}-\mathbf{A}_0\|_1 \le M} \left| \frac{1}{n\left(n-1\right)} \sum_{i \neq j} \sum_{k=1}^{m_1} \left\{ h\left(\varepsilon_{ik}, \varepsilon_{jk}\right) - \mathbb{E}\left[h\left(\varepsilon_{ik}, \varepsilon_{jk}\right)\right] \right\} \right|.$$

For any $\gamma \ge 0$, Markov's inequality implies

$$\mathbb{P}\left(\mathcal{Z}_M > t\right) = \mathbb{P}\left(e^{\gamma \mathcal{Z}_M} > e^{\gamma t}\right) \le e^{-\gamma t} \mathbb{E} e^{\gamma \mathcal{Z}_M}.$$

Let $M_n = \lfloor n/2 \rfloor$, the largest integer that is no larger than $n/2$. Let $\{\sigma_{ik}\}, i = 1, \ldots, n, k = 1, \ldots, m_1$ denote a Rademacher sequence independent of $\{\mathbf{x}_i, \boldsymbol{\varepsilon_i}\}, i = 1, \ldots, n$. Similar to the proof of Lemma 22 we have

$$\mathbb{E} e^{\gamma \mathcal{Z}_M} \le \frac{1}{n!} \sum_{\pi} \mathbb{E} \exp\left(4\gamma M \left\| \frac{1}{M_n} \sum_{i=1}^{M_n} \sum_{k=1}^{m_1} \sigma_{ik} \left(\mathbf{x}_{\pi(i)} \mathbf{e}_k^\top(m_1) - \mathbf{x}_{\pi(M_n+i)} \mathbf{e}_k^\top(m_1) \right) \right\|_{\mathrm{op}} \right).$$

To bound this term, using the ε-net discretization method in Lemma 22 and the sub-Gaussion assumption on covariates, similarly we obtain

$$\mathbb{E} e^{\gamma \mathcal{Z}_M} \le 2 \cdot 9^{m_1+m_2} \exp\left(\frac{64\gamma^2 M^2 \kappa_0^2 C_1^2}{M_n} \right)$$

for some constant C_1. Thus we have

$$\begin{aligned} \mathbb{P}\left(\mathcal{Z}_M > t\right) &\le \inf_{\gamma \ge 0} e^{-\gamma t} \mathbb{E} e^{\gamma \mathcal{Z}_M} \\ &\le \inf_{\gamma \ge 0} 2 \cdot 9^{m_1+m_2} \exp\left(C_1^2 \kappa_0^2 \gamma^2 M^2 / n - \gamma t \right) \\ &= 2 \cdot 9^{m_1+m_2} \exp\left(-\frac{n t^2}{4 C_1^2 \kappa_0^2 M^2} \right). \end{aligned}$$

Now we take $t = 8C_1\kappa_0 \sqrt{\frac{m_1+m_2}{n}} M$, this implies $\mathcal{Z}_M \le 8C_1\kappa_0 \sqrt{\frac{m_1+m_2}{n}} M$ with probability at least $1 - 2\exp\left(-4\left(m_1+m_2\right)\right)$.

(b) The first conclusion follows from the definition of λ^*. Next we give the proof of the second conclusion. Conditional on $\{\mathbf{X}_i\}_{i=1}^n$. For any $\gamma \ge 0$, Markov's inequality implies

$$\mathbb{P}\left(S_n > t \mid \mathbf{X}\right) = \mathbb{P}\left(e^{\gamma S_n} > e^{\gamma t} \mid \mathbf{X}\right) \le e^{-\gamma t} \mathbb{E}_{\mathbb{P}|\mathbf{X}} e^{\gamma S_n}.$$

By the ε-net discretization method and the sub-Gaussian assumption on covariates, we have

$$\mathbb{E}_{\mathbb{P}|\mathbf{X}} e^{\gamma S_n} \le 2 \cdot 9^{m_1+m_2} \exp\left(\frac{16\gamma^2}{n} \max_{u \in \mathcal{N}, v \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \left(u^\top \mathbf{x}_i \right)^2 \right).$$

Assume $16\max_{u\in\mathcal{N}}\frac{1}{n}\sum_{i=1}^{n}\left(u^{\top}\mathbf{x}_i\right)^2\leq\kappa_0^2C_2^2$ for some constant C_2, we have

$$\begin{aligned}\mathbb{P}\left(S_n>t\mid\mathbf{X}\right)&\leq\inf_{\gamma\geq0}e^{-\gamma t}\mathbb{E}_{\mathbb{P}|\mathbf{X}}e^{\gamma S_n}\\&\leq\inf_{\gamma\geq0}2\cdot9^{m_1+m_2}\exp\left(C_2^2\kappa_0^2\gamma^2/n-\gamma t\right)\\&=2\cdot9^{m_1+m_2}\exp\left(-\frac{nt^2}{4C_2^2\kappa_0^2}\right).\end{aligned}$$

Now we take $t=4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$, this implies $S_n\leq4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$ with probability at least $1-2\exp\left(-\left(m_1+m_2\right)\right)$. Then if $m_1+m_2>\ln\left(2/\alpha_0\right)$, we have $\lambda^*<4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$.

Similar to the discussion of Lemma 22(b), we have that there exists constant C_2 and c_2 such that $16\max_{u\in\mathcal{N}}\frac{1}{n}\sum_{i=1}^{n}\left(u^{\top}\mathbf{x}_i\right)^2\leq\kappa_0^2C_2^2$ with probability at least $1-9^{m_1+m_2}\exp\left(-c_2n\right)$. If $n>4\left(m_1+m_2\right)/c_2$, then we have $\lambda^*<4C_2\kappa_0\sqrt{\frac{m_1+m_2}{n}}$ holds with probability at least $1-\exp\left(-\left(m_1+m_2\right)\right)$.

(c) We denote $\mathrm{Cov}\left(\mathbf{x}_1\right)$ by $\mathbf{J}$. First we show that

$$Q\left(\mathbf{A}\right)-Q\left(\mathbf{A}_0\right)\geq\frac{b_1}{2}\|\mathbf{J}^{\frac{1}{2}}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F^2\wedge\frac{b_1b_4}{2}\|\mathbf{J}^{\frac{1}{2}}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F$$

which implies that $\mathbf{A}_0$ is the minimizer of $Q\left(\mathbf{A}\right)$. We define the maximal radius over which the criterion function can be minorated by a quadratic function

$$r_{\mathbf{A}_0}=\sup_r\left\{r:Q\left(\mathbf{A}\right)-Q\left(\mathbf{A}_0\right)\geq\frac{b_1}{2}\|\mathbf{J}^{1/2}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F^2\text{ for all }\|\mathbf{J}^{1/2}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F\leq r\right\}.$$

We claim that $r_{\mathbf{A}_0}\geq b_4$. Note that the empirical loss of multivariate regression is a summation of m_1 empirical loss of linear regression, we have

$$Q\left(\mathbf{A}\right)-Q\left(\mathbf{A}_0\right)\geq b_1\|\mathbf{J}^{1/2}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F^2-\frac{b_2}{6}\sum_{k=1}^{m_1}\mathbb{E}_{\mathbf{X}}\left\{\left|\left(\mathbf{x}_1-\mathbf{x}_2\right)^{\top}\left(\mathbf{A}_k-\mathbf{A}_{0k}\right)\right|^3\right\}.$$

Note that for all $\mathbf{A}-\mathbf{A}_0$, if

$$\|\mathbf{J}^{1/2}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F\leq b_4\leq\frac{3b_1}{2\sqrt{2}b_2}\inf_{\mathbf{A}-\mathbf{A}_0\neq0}\frac{\left(\sum_{k=1}^{m_1}\mathbb{E}\left|\left(\mathbf{x}_i-\mathbf{x}_j\right)^{\top}\left(\mathbf{A}_k-\mathbf{A}_{0k}\right)\right|^2\right)^{3/2}}{\sum_{k=1}^{m_1}\mathbb{E}\left|\left(\mathbf{x}_i-\mathbf{x}_j\right)^{\top}\left(\mathbf{A}_k-\mathbf{A}_{0k}\right)\right|^3},$$

it follows that $\frac{b_1}{2}\|\mathbf{J}^{1/2}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F^2\geq\frac{b_2}{6}\sum_{k=1}^{m_1}\mathbb{E}_{\mathbf{X}}\left\{\left|\left(\mathbf{x}_1-\mathbf{x}_2\right)^{\top}\left(\mathbf{A}_k-\mathbf{A}_{0k}\right)\right|^3\right\}$. Then we obtain

$$Q\left(\mathbf{A}\right)-Q\left(\mathbf{A}_0\right)\geq\frac{b_1}{2}\|\mathbf{J}^{1/2}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F^2\text{ for all }\|\mathbf{J}^{1/2}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F\leq b_4.$$

This implies $r_{\mathbf{A}_0}\geq b_4$. Similar to the proof of Lemma 22(c), we have

$$Q\left(\mathbf{A}\right)-Q\left(\mathbf{A}_0\right)\geq\frac{b_1}{2}\|\mathbf{J}^{\frac{1}{2}}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F^2\wedge\frac{b_1b_4}{2}\|\mathbf{J}^{\frac{1}{2}}\left(\mathbf{A}-\mathbf{A}_0\right)^{\top}\|_F$$

for any $\mathbf{A}$. This implies that $\mathbf{A}_0$ is the minimizer of $Q(\mathbf{A})$, and

$$Q(\mathbf{A}) - Q(\mathbf{A}_0) \geq \frac{b_1 b_3}{2}\|\mathbf{A}-\mathbf{A}_0\|_F^2 \wedge \frac{b_1 b_3 b_4}{2}\|\mathbf{A}-\mathbf{A}_0\|_F$$

for any $\mathbf{A}-\mathbf{A}_0 \in \mathcal{C}$. Similar to the proof of Lemma 22(c), we can prove that

$$Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \geq \frac{b_1 b_3}{2}\|\widehat{\mathbf{A}}-\mathbf{A}_0\|_F^2.$$

■

Appendix E: Proof of Theorem 18

Hereafter we use $\widetilde{\mathbf{X}}$ and $\widetilde{\varepsilon}$ to denote $a\mathbf{X}$ and $a\varepsilon$ respectively, and let $\widetilde{\zeta}_{ij} = a_i\varepsilon_i - a_j\varepsilon_j$. With the following lemma and Theorem 1, we can get Theorem 18.

Lemma 25 *Suppose that Assumptions 15-17 hold. Assume $\mathbf{A}_0 \in \mathcal{B}_q(R_q) \cap \mathcal{S}$. Then we have following conclusions:*

(a) Define for all $M > 0$

$$\mathcal{Z}_M := \sup_{\|\mathbf{A}-\mathbf{A}_0\|_1 \leq M} \left|\{Q_n(\mathbf{A}) - Q(\mathbf{A})\} - \{Q_n(\mathbf{A}_0) - Q(\mathbf{A}_0)\}\right|.$$

Then we have for a constant $C_0 > 0$, with $\lambda'_\varepsilon = 8\sqrt{2}C_0\sqrt{\frac{L\log(m_1+m_2)}{nm_2}}$ and $\widetilde{\lambda}' = 8\eta\sqrt{\frac{\log(m_1+m_2)}{n}}$, such that

$$\mathcal{Z}_M \leq \lambda'_\varepsilon M + \widetilde{\lambda}'$$

with probability at least $1-\exp(-4\log(m_1+m_2))$. Moreover, let $\lambda_\varepsilon = 2\lambda'_\varepsilon$ and $\widetilde{\lambda} = 2\widetilde{\lambda}'$, we have, for any $\mathbf{A} \in \mathcal{S}$

$$\left|\{Q_n(\mathbf{A}) - Q(\mathbf{A})\} - \{Q_n(\mathbf{A}_0) - Q(\mathbf{A}_0)\}\right| \leq \lambda_\varepsilon\|\mathbf{A}-\mathbf{A}_0\|_1 + \widetilde{\lambda}$$

with probability at least $1-\exp(-2\log(m_1+m_2))$.

(b) $\lambda^* > 2\|\nabla Q_n(\mathbf{A}_0)\|_{\text{op}}$ with probability $1-\alpha_0$ if ε_i's are symmetric. If $m_1+m_2 \geq (3/\alpha_0)^{1/3}$, then $\lambda^* \leq \sqrt{\frac{5L\log(m_1+m_2)}{nm_2}}$ with probability at least $1-\exp(-2\log(m_1+m_2))$.

(c) $\mathbf{A}_0$ is the minimizer of $Q(\mathbf{A})$. For all $\mathbf{A}$, $Q(\mathbf{A}) - Q(\mathbf{A}_0) \geq \frac{1}{c_1^2\mu m_1 m_2}\|\mathbf{A}-\mathbf{A}_0\|_F^2$.

Proof [Proof of Lemma 25] (a) We only prove the first conclusion since the proof of the second conclusion follows similarly to that of Lemma 22(a). Due to the special structure of $\widetilde{\mathbf{X}}$, we adopt the bounded difference inequality to control the empirical process. Write $h(\widetilde{\varepsilon}_i, \widetilde{\varepsilon}_j) = |(\widetilde{\varepsilon}_i - \widetilde{\varepsilon}_j) - \langle \widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j, \mathbf{A}-\mathbf{A}_0\rangle| - |\widetilde{\varepsilon}_i - \widetilde{\varepsilon}_j|$,

$$\mathcal{Z}_M = \sup_{\|\mathbf{A}-\mathbf{A}_0\|_1 \leq M} \left|\frac{1}{n(n-1)}\sum_{i\neq j}\{h(\widetilde{\varepsilon}_i, \widetilde{\varepsilon}_j) - \mathbb{E}[h(\widetilde{\varepsilon}_i, \widetilde{\varepsilon}_j)]\}\right|.$$

Note that $|h(\widetilde{\varepsilon}_i, \widetilde{\varepsilon}_j)| \le |\langle \widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j, \mathbf{A} - \mathbf{A}_0 \rangle| \le 4\eta$. Hence if we perturb one observation of the data set, the value of $\mathcal{Z}_M$ changes at $\frac{4\eta}{n}$. By the bounded difference inequality, $\forall t > 0$

$$P\left(\mathcal{Z}_M - \mathbb{E}\left\{\mathcal{Z}_M\right\} > t\right) \le \exp\left(-\frac{nt^2}{16\eta^2}\right).$$

Let $M_n = \lfloor n/2 \rfloor$, the largest integer that is no larger than $n/2$. Let $\sigma_i, i = 1, \ldots, n$ denote a Rademacher sequence independent of $\{\widetilde{\mathbf{X}}_i, \widetilde{\varepsilon}_i\}_{i=1}^n$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Z}_M\right] &= \mathbb{E}\left[\sup_{\|\mathbf{A}\|_\infty \le \eta : \|\mathbf{A}-\mathbf{A}_0\|_1 \le M} \frac{1}{n!} \left| \sum_{\pi} M_n^{-1} \sum_{i=1}^{M_n} \left\{ h\left(\widetilde{\varepsilon}_{\pi(i)}, \widetilde{\varepsilon}_{\pi(M_n+i)}\right) - \mathbb{E} h\left(\widetilde{\varepsilon}_{\pi(i)}, \widetilde{\varepsilon}_{\pi(M_n+i)}\right) \right\} \right| \right] \\
&\le 4\mathbb{E}\left[\sup_{\|\mathbf{A}\|_\infty \le \eta : \|\mathbf{A}-\mathbf{A}_0\|_1 \le M} \left\| M_n^{-1} \sum_{i=1}^{M_n} \sigma_i \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)}\right)^\top \right\|_{\mathrm{op}} \|\mathbf{A} - \mathbf{A}_0\|_1 \right] \\
&\le 4M\mathbb{E}\left\| M_n^{-1} \sum_{i=1}^{M_n} \sigma_i \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)}\right)^\top \right\|_{\mathrm{op}} \\
&\le 8\sqrt{2} M C_0 \sqrt{\frac{L \log\left(m_1 + m_2\right)}{n m_2}} \quad \text{by triangle inequality and Lemma 29.}
\end{aligned}$$

Now we take $t = 8\eta\sqrt{\frac{\log(m_1+m_2)}{n}}$, this gives

$$\mathcal{Z}_M \le 8\sqrt{2} M C_0 \sqrt{\frac{L \log\left(m_1 + m_2\right)}{n m_2}} + 8\eta \sqrt{\frac{\log\left(m_1 + m_2\right)}{n}}$$

with probability at least $1 - \exp\left(-4\log\left(m_1 + m_2\right)\right)$.

(b) First we show that the symmetrized masks $a_i \mathbf{X}_i$ and noise $a_i \varepsilon_i$ are independent. Without loss of generality, we can calculate

$$\begin{aligned}
&\mathbb{P}\left(a_i \mathbf{X}_i = \mathbf{e}_1(m_1)\mathbf{e}_1(m_2)^\top, a_i\varepsilon_i = \nu\right) \\
&= \mathbb{P}\left(a_i = 1, \mathbf{X}_i = \mathbf{e}_1(m_1)\mathbf{e}_1(m_2)^\top, \varepsilon_i = \nu\right) \\
&= \mathbb{P}\left(a_i = 1\right) \mathbb{P}\left(\mathbf{X}_i = \mathbf{e}_1(m_1)\mathbf{e}_1(m_2)^\top\right) \mathbb{P}\left(\varepsilon_i = \nu\right) \\
&= \mathbb{P}\left(a_i = 1\right) \mathbb{P}\left(\mathbf{X}_i = \mathbf{e}_1(m_1)\mathbf{e}_1(m_2)^\top\right) \mathbb{P}\left(a_i\varepsilon_i = \nu\right) \\
&= \mathbb{P}\left(a_i \mathbf{X}_i = \mathbf{e}_1(m_1)\mathbf{e}_1(m_2)^\top\right) \mathbb{P}\left(a_i\varepsilon_i = \nu\right).
\end{aligned}$$

Then the first conclusion follows from the fact that $\{a_i \mathbf{X}_i\}_{i=1}^n$ is independent of $\{a_i\varepsilon_i\}_{i=1}^n$ and $\|\nabla Q_n(\mathbf{A}_0)\|_{\mathrm{op}}$ has the same distribution as S_n when ε_i's are symmetric.

Next we give the proof of the second conclusion. Conditional on $\{\widetilde{\mathbf{X}}_i\}_{i=1}^n$. For any $\gamma \ge 0$, Markov's inequality implies

$$\mathbb{P}\left(S_n > t \mid \widetilde{\mathbf{X}}\right) = \mathbb{P}\left(e^{\gamma S_n} > e^{\gamma t} \mid \widetilde{\mathbf{X}}\right) \le e^{-\gamma t} \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} e^{\gamma S_n}.$$

Recall that $S_n = \| \{n(n-1)\}^{-1} \sum_{i \neq j} \left(\widetilde{\mathbf{X}}_j - \widetilde{\mathbf{X}}_i \right) \operatorname{sign}(\varepsilon_i - \varepsilon_j) \|_{\mathrm{op}}$, by the definition of operator norm, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} e^{\gamma S_n} \\
&= \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} \exp\left(\gamma \sup_{\|u\|_2 \leq 1, \|v\|_2 \leq 1} \{n(n-1)\}^{-1} \sum_{i \neq j} u^\top \left(\widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j \right) v \operatorname{sign}(\varepsilon_i - \varepsilon_j) \right) \\
&= \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} \exp\left(\gamma \sup_{\|u\|_2 \leq 1, \|v\|_2 \leq 1} \frac{1}{n!} \sum_{\pi} M_n^{-1} \sum_{i=1}^{M_n} u^\top \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right) v \operatorname{sign}\left(\varepsilon_{\pi(i)} - \varepsilon_{\pi(M_n+i)} \right) \right) \\
&\leq \frac{1}{n!} \sum_{\pi} \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} \exp\left(\gamma \sup_{\|u\|_2 \leq 1, \|v\|_2 \leq 1} M_n^{-1} \sum_{i=1}^{M_n} u^\top \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right) v \operatorname{sign}\left(\varepsilon_{\pi(i)} - \varepsilon_{\pi(M_n+i)} \right) \right) \\
&= \frac{1}{n!} \sum_{\pi} \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} \exp\left(\gamma \left\| M_n^{-1} \sum_{i=1}^{M_n} \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right) \operatorname{sign}\left(\varepsilon_{\pi(i)} - \varepsilon_{\pi(M_n+i)} \right) \right\|_{\mathrm{op}} \right).
\end{aligned}
$$

We denote $Z_\pi = \| M_n^{-1} \sum_{i=1}^{M_n} \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right) \operatorname{sign}\left(\varepsilon_{\pi(i)} - \varepsilon_{\pi(M_n+i)} \right) \|_{\mathrm{op}}$ for the simplicity of notation, then we have

$$
\mathbb{P}\left(S_n > t \mid \widetilde{\mathbf{X}} \right) \leq \inf_{\gamma \geq 0} e^{-\gamma t} \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} e^{\gamma S_n} \leq \inf_{\gamma \geq 0} e^{-\gamma t} \frac{1}{n!} \sum_{\pi} \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} e^{\gamma Z_\pi}.
$$

We do not use the ε-net argument to bound the operator norm term, instead we apply the Matrix Bernstein inequality (Theorem 6.1.1 in Tropp (2015)) to take advantage of the singleton design under the matrix completion setting. We state the Matrix Bernstein inequality in Lemma 26 for the sake of completeness. Now we calculate the quantity needed in Matrix Bernstein inequality. For all permutation π, we have

$$
\begin{aligned}
&\left\| M_n^{-1} \sum_{i=1}^{M_n} \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right) \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right)^\top \right\|_{\mathrm{op}} \\
&= \sup_{\|u\|_2 \leq 1, \|v\|_2 \leq 1} M_n^{-1} \sum_{i=1}^{M_n} u^\top \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right) \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right)^\top v
\end{aligned}
$$

By the special structure of $\widetilde{\mathbf{X}}$, we have $\pm u^\top \widetilde{\mathbf{X}}_a \widetilde{\mathbf{X}}_b^\top v \leq \frac{1}{2}(u^\top \widetilde{\mathbf{X}}_a \widetilde{\mathbf{X}}_a^\top u + v^\top \widetilde{\mathbf{X}}_b \widetilde{\mathbf{X}}_b^\top v)$, then for any $\|u\|_2 \leq 1, \|v\|_2 \leq 1$,

$$
\begin{aligned}
&M_n^{-1} \sum_{i=1}^{M_n} u^\top \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right) \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)} \right)^\top v \\
&\leq (2M_n)^{-1} \sum_{i=1}^{n} \left\{ 2u^\top \widetilde{\mathbf{X}}_i \widetilde{\mathbf{X}}_i^\top v + u^\top \widetilde{\mathbf{X}}_i \widetilde{\mathbf{X}}_i^\top u + v^\top \widetilde{\mathbf{X}}_i \widetilde{\mathbf{X}}_i^\top v \right\} \\
&= 4n^{-1} \left(\frac{u+v}{2} \right)^\top \sum_{i=1}^{n} \widetilde{\mathbf{X}}_i \widetilde{\mathbf{X}}_i^\top \left(\frac{u+v}{2} \right).
\end{aligned}
$$

Note that $\left\|\frac{u+v}{2}\right\|_2 \le 1$, we obtain

$$\left\| M_n^{-1} \sum_{i=1}^{M_n} \left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)}\right)\left(\widetilde{\mathbf{X}}_{\pi(i)} - \widetilde{\mathbf{X}}_{\pi(M_n+i)}\right)^\top \right\|_{\mathrm{op}} \le 4 \left\| n^{-1} \sum_{i=1}^{n} \widetilde{\mathbf{X}}_i \widetilde{\mathbf{X}}_i^\top \right\|_{\mathrm{op}}.$$

Assume that $4\|n^{-1}\sum_{i=1}^n \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top\|_{\mathrm{op}} \le \frac{8L}{m_2}$, then we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} e^{\gamma Z_\pi} \\
&= \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} \int_0^Z \gamma e^{\gamma y} dy + 1 \\
&= \int_0^{+\infty} \mathbb{P}\left(Z_\pi > y \mid \widetilde{\mathbf{X}}\right) \gamma e^{\gamma y} dy + 1 \quad \text{by Fubini's theorem} \\
&\le \gamma \int_0^{+\infty} \exp\left(-\frac{2n^2y^2}{\frac{8nL}{m_2} + ny}\right) dy + 1 \quad \text{by Lemma 26 Matrix Bernstein inequality} \\
&\le \gamma \left\{ \int_0^{+\infty} \exp\left(-\frac{nm_2y^2}{8L} + \gamma y\right) dy + \int_0^{+\infty} \exp\left(-ny + \gamma y\right) dy \right\} + 1 \\
&\le \gamma \sqrt{\frac{8\pi L}{nm_2}} \exp\left(\frac{8\gamma^2 L}{4nm_2}\right) + \int_0^{+\infty} \exp\left(-ny + \gamma y\right) dy + 1.
\end{aligned}$$

Take $t = \sqrt{\frac{40L\log(m_1+m_2)}{nm_2}}$ and $\gamma = nm_2t/8L$, if $n \gtrsim m_2 \log(m_1+m_2)$, we obtain

$$\begin{aligned}
&\mathbb{P}\left(S_n > t \mid \widetilde{\mathbf{X}}\right) \\
&\le \inf_{\gamma \ge 0} e^{-\gamma t} \mathbb{E}_{\mathbb{P}|\widetilde{\mathbf{X}}} e^{\gamma Z_\pi} \\
&\le e^{-5\log(m_1+m_2)} \left\{ \frac{nm_2}{8L} \sqrt{\frac{40L\log(m_1+m_2)}{nm_2}} \sqrt{\frac{8\pi L}{nm_2}} \exp\left(\frac{5}{4}\log(m_1+m_2)\right) + \frac{1}{n-\gamma} + 1 \right\} \\
&\le e^{-5\log(m_1+m_2)} \left\{ \sqrt{5\pi \log(m_1+m_2)} \exp\left(\frac{5}{4}\log(m_1+m_2)\right) + 2 \right\} \\
&\le 3\exp\left(-3\log(m_1+m_2)\right),
\end{aligned}$$

thus if $m_1 + m_2 \ge (3/\alpha_0)^{1/3}$, we conclude that $\lambda^* \le \sqrt{\frac{40L\log(m_1+m_2)}{nm_2}}$.

Finally we show that $4\|n^{-1}\sum_{i=1}^n \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top\|_{\mathrm{op}} \le \frac{8L}{m_2}$ with high probability. It's easy to see that $\|\mathbb{E}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top\|_{\mathrm{op}} \le \frac{L}{m_2}$, then by Matrix Bernstein inequality and $n \gtrsim m_2\log(m_1+m_2)$, we have

$$\mathbb{P}\left(\left\| n^{-1} \sum_{i=1}^n \widetilde{\mathbf{X}}_i \widetilde{\mathbf{X}}_i^\top \right\|_{\mathrm{op}} > \frac{2L}{m_2} \right) \le (m_1+m_2) \exp\left(\frac{-\frac{L^2}{m_2^2}}{\frac{L}{nm_2} + \frac{1}{3n}\frac{L}{m_2}} \right) \le \exp\left(-2\log(m_1+m_2)\right).$$

Therefore, we conclude that $\lambda^* \le \sqrt{\frac{5L\log(m_1+m_2)}{nm_2}}$ with probability at least $1 - \exp\left(-2\log(m_1+m_2)\right)$.

(c) Similar to the proof of Lemma 22, we have

$$\begin{aligned} Q(\mathbf{A}) - Q(\mathbf{A}_0) =& \mathbb{E}\langle \widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j, \mathbf{A} - \mathbf{A}_0 \rangle \left\{ \mathrm{I}\left(\widetilde{\zeta}_{ij} < 0\right) - 1/2 \right\} \\ &+ \mathbb{E} \int_0^{\langle \widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \left\{ \mathrm{I}\left(\widetilde{\zeta}_{ij} \le s\right) - \mathrm{I}\left(\widetilde{\zeta}_{ij} \le 0\right) \right\} ds. \end{aligned}$$

For the first part,

$$\begin{aligned} & \mathbb{E}\langle \widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j, \mathbf{A} - \mathbf{A}_0 \rangle \left\{ \mathrm{I}\left(\widetilde{\zeta}_{ij} < 0\right) - 1/2 \right\} \\ =& \frac{1}{2}\mathbb{E}\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle \left\{ \mathrm{I}\left(\zeta_{ij}^- < 0\right) - 1/2 \right\} + \frac{1}{2}\mathbb{E}\langle \mathbf{X}_i + \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle \left\{ \mathrm{I}\left(\zeta_{ij}^+ < 0\right) - 1/2 \right\} \\ =& 0. \end{aligned}$$

For the second part,

$$\begin{aligned} & \mathbb{E} \int_0^{\langle \widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \left\{ \mathrm{I}\left(\widetilde{\zeta}_{ij} \le s\right) - \mathrm{I}\left(\widetilde{\zeta}_{ij} \le 0\right) \right\} ds \\ =& \frac{1}{2}\mathbb{E} \int_0^{\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \left\{ \mathrm{I}\left(\zeta_{ij}^- \le s\right) - \mathrm{I}\left(\zeta_{ij}^- \le 0\right) \right\} ds \\ &+ \frac{1}{2}\mathbb{E} \int_0^{\langle \mathbf{X}_i + \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \left\{ \mathrm{I}\left(\zeta_{ij}^+ \le s\right) - \mathrm{I}\left(\zeta_{ij}^+ \le 0\right) \right\} ds. \end{aligned}$$

Denote the distribution function of ζ_{ij}^- by $F^*(\cdot)$. By the independence of $\mathbf{X}$ and ε,

$$\begin{aligned} & \frac{1}{2}\mathbb{E} \int_0^{\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \left\{ \mathrm{I}\left(\zeta_{ij}^- \le s\right) - \mathrm{I}\left(\zeta_{ij}^- \le 0\right) \right\} ds \\ =& \frac{1}{2}\mathbb{E}_{\mathbf{X}_i, \mathbf{X}_j} \int_0^{\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \left\{ F^*(s) - F^*(0) \right\} ds \\ =& \frac{1}{2}\mathbb{E}_{\mathbf{X}_i, \mathbf{X}_j} \int_0^{\langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle} f^-(\xi_{ij})\, s ds \\ \ge& \frac{1}{4c_1^2}\mathbb{E}_{\mathbf{X}_i, \mathbf{X}_j} \langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle^2. \end{aligned}$$

For the same reason, we have

$$\frac{1}{2}\mathbb{E} \int_0^{\langle \mathbf{X}_i + \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle} \left\{ \mathrm{I}\left(\zeta_{ij}^+ \le s\right) - \mathrm{I}\left(\zeta_{ij}^+ \le 0\right) \right\} ds \ge \frac{1}{4c_1^2}\mathbb{E}_{\mathbf{X}_i, \mathbf{X}_j} \langle \mathbf{X}_i + \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle^2.$$

From the above discussion, we conclude that

$$\begin{aligned} Q(\mathbf{A}) - Q(\mathbf{A}_0) &\ge \frac{1}{4c_1^2}\mathbb{E}_{\mathbf{X}_i, \mathbf{X}_j} \langle \mathbf{X}_i - \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle^2 + \frac{1}{4c_1^2}\mathbb{E}_{\mathbf{X}_i, \mathbf{X}_j} \langle \mathbf{X}_i + \mathbf{X}_j, \mathbf{A} - \mathbf{A}_0 \rangle^2. \\ &= \frac{1}{c_1^2}\mathbb{E}_{\mathbf{X}_i} \langle \mathbf{X}_i, \mathbf{A} - \mathbf{A}_0 \rangle^2 \\ &\ge \frac{1}{c_1^2 \mu m_1 m_2} \left\| \mathbf{A} - \mathbf{A}_0 \right\|_F^2, \end{aligned}$$

which also implies that $\mathbf{A}_0$ is the minimizer of $Q(\mathbf{A})$. ■

Appendix F: Collection of Lemmas

In this part, we give some lemmas used in the previous proofs.

Lemma 26 (Matrix Bernstein, Theorem 6.1.1 in Tropp (2015)) *Consider a finite sequence $\{\mathbf{Z}_i\}_{i=1}^n$ of independent, random matrices with common dimension $m_1 \times m_2$. Assume that $\mathbb{E}\mathbf{Z}_i = \mathbf{0}$ and $\|\mathbf{Z}_i\|_{\text{op}} \leq H$ for each index k, introduce the random matrix*

$$\mathbf{S} = \sum_k \mathbf{Z}_i.$$

Let $v(\mathbf{S})$ be the matrix variance statistic of the sum:

$$\begin{aligned} v(\mathbf{S}) &= \max\left\{\left\|\mathbb{E}\left(\mathbf{S}\mathbf{S}^\top\right)\right\|_{\text{op}}, \left\|\mathbb{E}\left(\mathbf{S}^\top\mathbf{S}\right)\right\|_{\text{op}}\right\} \\ &= \max\left\{\left\|\sum_k \mathbb{E}\left(\mathbf{Z}_i\mathbf{Z}_i^\top\right)\right\|_{\text{op}}, \left\|\sum_k \mathbb{E}\left(\mathbf{Z}_i^\top\mathbf{Z}_i\right)\right\|_{\text{op}}\right\}. \end{aligned}$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{\|\mathbf{S}\|_{\text{op}} \geq t\right\} \leq (m_1 + m_2)\exp\left(\frac{-t^2/2}{v(\mathbf{S}) + Ht/3}\right).$$

Matrix Bernstein requires the constant bound on $\|\mathbf{Z}_i\|_{\text{op}}$, it's possible to replace this requirement by the bound on the weaker ψ_p-norms of $\|\mathbf{Z}_i\|_{\text{op}}$, we have following result which are very useful in matrix completion problem. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be independent random matrices with dimensions $m_1 \times m_2$. Define

$$\sigma_{\mathbf{Z}}^2 = \max\left\{\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\mathbf{Z}_i\mathbf{Z}_i^\top\right)\right\|_{\text{op}}, \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\mathbf{Z}_i^\top\mathbf{Z}_i\right)\right\|_{\text{op}}\right\}.$$

Lemma 27 (Proposition 2 in Koltchinskii et al. (2011)) *Let $\mathbf{Z}, \mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be i.i.d. random matrices with dimensions $p \times q$ that satisfy $\mathbb{E}(\mathbf{Z}) = 0$. Suppose that $\| \|\mathbf{Z}\|_{\text{op}} \|_{\psi_p} < \infty$ for some $p \geq 1$. Then there exists a constant $C > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$*

$$\begin{aligned} \left\|\frac{\mathbf{Z}_1 + \cdots + \mathbf{Z}_n}{n}\right\|_{\text{op}} \leq C\max\Bigg\{&\sigma_{\mathbf{Z}}\sqrt{\frac{t + \log(m_1 + m_2)}{n}}, \\ &\| \|\mathbf{Z}\|_{\text{op}} \|_{\psi_p}\left(\log\frac{\| \|\mathbf{Z}\|_{\text{op}} \|_{\psi_p}}{\sigma_{\mathbf{Z}}}\right)^{1/p}\frac{t + \log(m_1 + m_2)}{n}\Bigg\}. \end{aligned}$$

Lemma 28 *Let $\{\mathbf{X}_i\}_{i=1}^n$ be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 15 and 16. Assume that $(\sigma_i)_{i=1}^n$ are i.i.d. Rademacher random variables independent of $\{\mathbf{X}_i\}_{i=1}^n$. Then, there exists an absolute constant $C^* > 0$ that depends only on K and such that, for all $t > 0$ with probability at least $1 - e^{-t}$ we have*

$$\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i\mathbf{X}_i\right\|_{\text{op}} \leq C^*\max\left\{\sqrt{\frac{L(t + \log(m_1 + m_2))}{nm_2}}, \sqrt{\log m_2}\frac{(t + \log(m_1 + m_2))}{n}\right\}.$$

Proof We apply Lemma 27 to $\mathbf{Z}_i = \sigma_i \mathbf{X}_i$. We first estimate $\sigma_{\mathbf{Z}}$ and $\| \|\mathbf{Z}\|_{\text{op}} \|_{\psi_2}$. Note that $\mathbf{Z}_i$ is a zero-mean random matrix which satisfies $\|\mathbf{Z}_i\|_{\text{op}} \leq |\sigma_i|$, hence $\| \|\mathbf{Z}_i\|_{\text{op}} \|_{\psi_2} \leq \|\sigma_i\|_{\psi_2}$, together with the fact that the ψ_2 norm of a Rademacher random variable σ_i is equal to $\|\sigma_i\|_{\psi_2} = \sqrt{1/\log 2}$, we have $\| \|\mathbf{Z}\|_{\text{op}} \|_{\psi_2} \leq \sqrt{1/\log 2}$. Then we compute $\mathbb{E}\left(\mathbf{Z}_i \mathbf{Z}_i^\top\right) = R$ and $\mathbb{E}\left(\mathbf{Z}_i^\top \mathbf{Z}_i\right) = C$, where C and R is the diagonal matrix with C_k and R_j on the diagonal respectively. This and the fact that the $\mathbf{Z}_i$ are i.i.d. imply that $\sigma_{\mathbf{Z}}^2 = \max_{i,j}(C_i, R_j) \leq L/m_2$. Note that $\max_{i,j}(C_i, R_j) \geq 1/q$ together with the concavity of the logarithm imply that $\sqrt{\log(\| \|\mathbf{Z}\|_{\text{op}} \|_{\psi_p}/\sigma_{\mathbf{Z}})} \leq \sqrt{\log m_2}$. Hence the statement of Lemma 28 follows. ■

We choose the parameter t to be $t = \log(m_1 + m_2)$. Next we use Lemma 28 to bound the expectation of the largest singular value of the sum of masks.

Lemma 29 *Let $\{\mathbf{X}_i\}_{i=1}^n$ be i.i.d. with distribution Π on $\mathbf{X}$ which satisfies Assumptions 15 and 16. Assume that $\{\sigma_i\}_{i=1}^n$ are i.i.d. Rademacher random variables independent of $\{\mathbf{X}_i\}_{i=1}^n$. Then, for $n > 2m_2 \log^2(m_1 + m_2)/L$, there exists an absolute constant $C_0 > 0$ such that*

$$\mathbb{E}\left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \mathbf{X}_i \right\|_{\text{op}} \leq C_0 \sqrt{\frac{L \log(m_1 + m_2)}{n m_2}}.$$

Proof This lemma is obtained by integrating the tail probability of Lemma 28. ■

Appendix G: Algorithms and Complexity Analysis

G1: Proximal Gradient Algorithm for the Rank Matrix Lasso

In this subsection, we propose an accelerated proximal gradient (APG) algorithm to solve the minimization (4). Specifically, to minimize a penalized loss function, i.e.,

$$\min_{\mathbf{A}} \ \{Q_n(\mathbf{A}) + \lambda \|\mathbf{A}\|_1\}, \tag{G.1}$$

we employ the quadratic function

$$Q_{\text{Major}}(\mathbf{A}; \mathbf{A}^{(l)}) = Q_n(\mathbf{A}^{(l)}) + \left\langle \nabla Q_n(\mathbf{A}^{(l)}), \mathbf{A} - \mathbf{A}^{(l)} \right\rangle + (L/2)\|\mathbf{A} - \mathbf{A}^{(l)}\|_F^2 \tag{G.2}$$

to locally "majorize" $Q_n(\mathbf{A})$ for the t-th iteration (Fan et al., 2018), where L is a constant such that in a neighborhood of $\mathbf{A}^{(l)}$ we have $Q_n(\mathbf{A}) \leq Q_{\text{Major}}(\mathbf{A}; \mathbf{A}^{(l)})$ and the equality can be attained at $\mathbf{A}^{(l)}$. Then, we solve

$$\min_{\mathbf{A}} L_{\text{Major}}(\mathbf{A}; \mathbf{A}^{(l)}) = \min_{\mathbf{A}} \left\{ Q_{\text{Major}}(\mathbf{A}; \mathbf{A}^{(l)}) + \lambda \|\mathbf{A}\|_1 \right\} \tag{G.3}$$

and set the solution as $\mathbf{A}^{(l+1)}$, which gives

$$Q_n(A^{(l+1)}) + \lambda \|A^{(l+1)}\|_1 \leq L_{\text{Major}}(A^{(l+1)}; A^{(l)}) \leq L_{\text{Major}}(A^{(l)}; A^{(l)}) = Q_n(A^{(l)}) + \lambda \|A^{(l)}\|_1.$$

While typically (G.1) does not have a closed-form solution, the minimizer in (G.3) can be expressed using the singular value soft-thresholding operator (see e.g. Toh and Yun

(2010)): for any given $\mathbf{M} \in \mathbb{R}^{m_1 \times m_2}$, $\text{Soft}(\mathbf{M}; \tau) = \mathbf{U}_{\mathbf{M}}^{\top} \text{diag}\{(\sigma_i(\mathbf{M}) - \tau)_+\} \mathbf{V}_{\mathbf{M}}$. Here $\mathbf{M} = \mathbf{U}_{\mathbf{M}}^{\top} \text{diag}\{(\sigma_i(\mathbf{M}))\} \mathbf{V}_{\mathbf{M}}$ is the singular value decomposition of $\mathbf{M}$. Detailed description and explicit mathematical expressions are provided in Algorithm 1. In our simulations, the `tol` we picked is 10^{-4} for all the trials, and the maximal iteration $T = 100$. We take $L^{(0)} = 10^{-4} L_{\max}$, and set $L_{\max} = 3 \times 10^p$ which is empirically found good enough for convergence in most scenarios.

Algorithm 1: Accelerated proximal gradient algorithm for the rank matrix lasso

Input: Observed data $(\mathbf{y}_i, \mathbf{X}_i)$, for $i = 1, \cdots, n$; tuning parameter $\lambda^\star$ by (8); floor curvature $L^{(0)}$; ceiling curvature L_{max}; updating rate η; $t^{(0)} = t^{(1)} = 1$; convergence tolerance `tol` and maximal iteration T; initial estimate $\mathbf{A}^{(0)} = \mathbf{A}^{(1)} = \mathbf{0}$.

Output: Estimator $\widehat{\mathbf{A}}$.

Set $\mathbf{B}^{(l)} = \mathbf{A}^{(l)} + \frac{t^{(l-1)}-1}{t^{(l)}}(\mathbf{A}^{(l)} - \mathbf{A}^{(l-1)})$.

Calculate the sub-gradient $\mathbf{G}^{(l)} = \nabla Q_n(\mathbf{A})|_{\mathbf{A}=\mathbf{B}^{(l)}}$.

Set $L = \min\{\eta L^{(l-1)}, L_{max}\}$.

while $L < L_{max}$ **do**

  Compute $\mathbf{S} = \text{Soft}\left(\mathbf{B}^{(l)} - L^{-1}\mathbf{G}^{(l)}; L^{-1}\lambda^\star\right)$;

  Calculate $Q_n(\mathbf{S})$ and $Q_{\text{Major}}(\mathbf{S}; \mathbf{B}^{(l)})$;

  if $Q_n(\mathbf{S}) \le Q_{\text{Major}}(\mathbf{S}; \mathbf{B}^{(l)})$ **then**

    Set $\mathbf{A}^{(l+1)} = \mathbf{S}$, $L^{(l)} = L$

    break

  else

    Set $L = \min\{L/\eta, L_{max}\}$.

Compute $t^{(l+1)} = \frac{1+\sqrt{1+4(t^{(l)})^2}}{2}$.

Repeat above steps until the stop criterion is meet: $\|\mathbf{A}^{(l+1)} - \mathbf{A}^{(l)}\|_F / \|\mathbf{A}^{(l)}\|_F \le$ `tol` or the maximal number of iteration T is hit. Set $\widehat{\mathbf{A}} = \mathbf{A}^{(l+1)}$.

Remark 30 *The suggested APG is kind of a "first-order" algorithm, while some second-order algorithms, such as quasi-Newton algorithms, are available for tackling nuclear norm penalization problems. In Appendix H2, we implement a state-of-the-art quasi-Newton algorithm proposed by Becker et al. (2019) for solving the rank-based optimization and make a comparison with Algorithm 1. An interesting trade-off is that, while the second-order method has certain advantage on the total steps for convergence, it generally requires more time in each step since first-order methods have to be implied to solve the subproblems. Our empirical results reveal that the proposed APG algorithm is at least comparable with the quasi-Newton algorithm in terms of average computation time.*

Remark 31 *While "smooth + nonsmooth" optimization has been extensively studied (Nesterov, 2013; Lee et al., 2014; Chambolle and Dossal, 2015), the global convergence rates of the APG method for "nonsmooth + nonsmooth" optimization has not been fully understood (Bian and Wu, 2021). Current theoretical progress on this issue typically involves*

some modification on the procedure. For example, Yu et al. (2010) uses a "tightest pseudo quadratic fit" in the proximal step, which gives global convergence in objective function value. However, this would render the subproblem hard to solve, especially for the nuclear norm penalization. Another popular direction is to locally approximate the the loss function by smoothing methods (Nesterov, 2005; Zhang and Chen, 2009; Bian and Chen, 2020; Bian and Wu, 2021), but a careful tuning procedure is generally required. It is still not clear whether these solutions with certain guarantee of global convergence can be extended to our objective function and tuning scheme, but this certainly warrants future research.

G2: Complexity Analysis

We present an analysis of the time complexity of the entire algorithm including both parts of parameter selection and optimizations. We take the trace regression as an example for careful investigation, but other models such as multivariate regression can be studied analogously. Our standpoint is the primitive form of the algorithm without taking special structural blessing (such as sparse matrices or factor matrices) into consideration. For trace regression models, the complexity of Algorithm 1 can be decomposed into the following steps:

Pivotal tuning. Computing the summation of the gradient is of order $O(nm_1m_2)$. Uses SVD for obtaining the operator norm of $\nabla Q_n(\mathbf{A})$ requires $O(m_1m_2\min\{m_1, m_2\})$. It is usually redundant to perform a full SVD. There are other algorithms that are less expensive, like power iteration or Lanczos bidiagonalization (Baglama and Reichel, 2005; Larsen, 2004). Hence, the total complexity with B rounds of simulations is about $O\{B(nm_1m_2 + m_1^2m_2)\}$ (assume $m_1 \leq m_2$). In most problems, n is typically required to be larger than m_1 and m_2 to achieve successful recovery. Therefore, the complexity for Pivotal tuning is basically $O(Bnm_1m_2)$.

APG optimization The following steps contribute most significantly to the computation costs: (1) Subgradient calculation in Step 2, which by our previous argument, takes $O(nm_1m_2)$ operations for the problems of interest. (2) SVD in Step 5. In medium scale problems full SVD can be directly applied, which takes $O(m_1m_2\min\{m_1, m_2\})$ operations. Again, utilizing other algorithms for structured problems could overcome this barrier and accelerate the program significantly. (3) Calculation of the objective value and the majorized value in Step 6. The former involves a sorting of the residuals in $O(n\log n)$ operations and summation of n linear products. The latter demands a calculation of $Q_n(\mathbf{B}^{(l)})$ as well as two inner products for the first-order and quadratic approximation terms (see (G.2)). Thus the dominant part takes up to $O(nm_1m_2)$ operations.

From the above analysis we can conclude that APG proceeds with the complexity of $O(nm_1m_2)$ each iteration, which aligns well with the Pivotal tuning step. Our simulation provides evidence that the time for parameter selection and for optimization remains in a similar magnitude among various models and settings. Empirically, the pivotal tuning usually takes less time than the APG program. These results justify the superiority of pivotal tuning compared with some conventional methods like cross-validation in matrix settings.

In Appendix H1, we also discuss how to make our algorithm scalable to large-scale structured matrix recovery problems and verify its effectiveness via a simulation study.

Appendix H: Additional Simulations Results

H1: Adaptivity to Large-Scale Problems: Large-Scale Matrix Completion

Real-life application typically involves computation of large matrices, especially for matrix completion problem. By taking advantage of the structure of the problem(like sparsity in matrix completion scenarios) as well as state-of-art algorithms for handling operator norm and SVD computation(Larsen, 1998; Baglama and Reichel, 2005; Larsen, 2004), our program can be adapted to large-scale structures and demonstrate the computational efficiency gain of pivotal tuning. For example, large-scale matrix completion problems are endowed with a perfect sparse structure. Then instead of the $O(nm_1m_2)$ costs we only need $O(n)$ operations for calculating the sub-gradients and objective values(with $O(n \log n)$ rounds of sorting), which is undoubtedly a striking improvement over the direct summation. Furthermore, instead of performing full SVD each time, a more appealing trick is to gradually increase and explore the rank as the program progresses(Toh and Yun, 2010). With this strategy we typically only need to focus on several largest eigen-pairs in each iteraction, for which many fast algorithms exist. In our large-scale implementation we use the PROPACK package by Larsen (2004) to achieve this convenience, rendering our algorithm perfectly scalable to large matrix computation problems(m_1, m_2 as high as 5×10^4). As a concluding remark, some other improvements are also possible in large-scale extensions. For example, it is appealing to take advantage of robust distributed programs and parallelization algorithms. See for example Chen et al. (2020).

Here we report some numerical results on large-scale matrix completion problems. For operator norm and SVD calculation we make use of the `lansvd` function in PROPACK. We consider $m = 5 \times 10^3, 1 \times 10^4, 5 \times 10^4$ and $r = 10$. The ground truth $\mathbf{A}_0$ and the noise are generated similarly to our small-scale matrix completion setting. Each time we sample $N = 6 \cdot df$ observations, where the degree of freedom of a matrix $df = r(m_1 + m_2 - r) = r(2m - r)$. Table S.3 summarised the results. The results demonstrate both the computational efficiency and the statistical accuracy of the proposed algorithm.

Noise	Dimension	$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2$	Rank	Tuning(s)	Solving(s)	Iteration
Gaussian	5×10^3	0.180(8.7e-4)	10.0(0)	4.77	13.8	39.9
	1×10^4	0.180(6.9e-4)	10.0(0)	9.85	26.8	36.4
	5×10^4	0.179(3.4e-4)	10.0(0)	57.4	225	32.0
Cauchy	5×10^3	0.004(6.6e-5)	10.0(0)	4.74	34.5	84.0
	1×10^4	0.004(7.5e-5)	10.0(0)	9.86	70.5	84.4
	5×10^4	0.004(2.0e-5)	10.0(0)	58.7	723	87.4

Table S.3: Simulation for large-scale matrix completion

H2: Comparison of First-Order and Second-Order Algorithms

Our main trials are based on a first-order algorithm. The population version of the proposed loss function has higher order curvature which is coarsely represented by a constant diagonal matrix in first-order methods. We now frame a quasi-newton algorithm for minimizing the

rank-based objective, attempting to borrow information from second-order geometry of the formulation.

Algorithm 2: Proximal quasi-Newton algorithm for the rank matrix lasso

Input: Observed data $(\mathbf{y}_i, \mathbf{X}_i)$, for $i = 1, \cdots, n$; pivotal tuning parameter α; updating rate η; convergence tolerance `tol` and maximal iteration T; initial estimate $\mathbf{A}^{(0)}$; tuning parameter L.

Output: Estimator $\widehat{\mathbf{A}}$.

`/* Pivotal tuning */`

1 Calculate $\lambda^\star$ using the pivotal tuning scheme with confidence level α.

`/* Start optimization */`

2 **Repeat**

3 Choose $\mathbb{B}^{(l)}$ to be a Hessian approximation at $\mathbf{A}^{(l)}$.

4 Solve the subproblem for a searching direction using a first-order forward-backward splitting algorithm:

$$\Delta\mathbf{A}^{(l)} \leftarrow \arg\min_{\mathbf{D}} Q_n(\mathbf{A}^{(l)}) + \left\langle \nabla Q_n(\mathbf{A}^{(l)}), \mathbf{D} \right\rangle + (\mathbf{D}^V)^\top \mathbb{B}^{(l)} \mathbf{D}^V + \lambda^\star \|\mathbf{A}^{(l)} + \mathbf{D}\|_1.$$

5 Select $t^{(l)}$ with a backtracking line search;

6 Update: $\mathbf{A}^{(l+1)} = \mathbf{A}^{(l)} + t^{(l)} \Delta\mathbf{A}^{(l)}$.

7 **until:** $\|\mathbf{A}^{(l+1)} - \mathbf{A}^{(l)}\|_F / \|\mathbf{A}^{(l)}\|_F \leq$ `tol` or the maximal number of iteration T is hit.

8 Set $\widehat{\mathbf{A}} = \mathbf{A}^{(l+1)}$.

We consider several details regarding the implementation of the algorithm:

- Steps to update the Hessian approximation matrix(Step 3 in Algorithm 2). Since we are targeting problems involving large matrices, it is a necessity to avoid saving or loading dense Hessian matrices. To this end, we pursue using limited memories Hessian approximation schemes. Two simple yet admirable options in the literature are SR1 update(Becker and Fadili, 2012; Becker et al., 2019) and BFGS update(Lee et al., 2014), which provide rank-1 and rank-2 updates respectively. Concretely speaking, let

$$\mathbf{dS}^{(l)} = \{\mathbf{A}^{(l)} - \mathbf{A}^{(l-1)}\}^V, \ \mathbf{dG}^{(l)} = \{\nabla Q_n(\mathbf{A}^{(l)}) - \nabla Q_n(\mathbf{A}^{(l-1)})\}^V.$$

  The 1-step SR1 update gives:

$$\mathbb{B}^{(l)} = \mathbb{B}_0^{(l)} + \mathbf{u}^{(l)} \mathbf{u}^{(l)\top},$$

  where

$$\mathbf{u}^{(l)} = \frac{\mathbf{v}^{(l)} \mathbf{v}^{(l)\top}}{\mathbf{v}^{(l)\top} \mathbf{dS}^{(l)}}, \ \mathbf{v}^{(l)} = \mathbf{dG}^{(l)} - \mathbb{B}^{(l)} \mathbf{dS}^{(l)}.$$

  The 1-step BFGS gives:

$$\mathbb{B}^{(l)} = \mathbb{B}_0^{(l)} + \mathbf{u_1}^{(l)} \mathbf{u_1}^{(l)\top} - \mathbf{u_2}^{(l)} \mathbf{u_2}^{(l)\top},$$

where

$$\mathbf{u}_1^{(l)} = \frac{\mathbf{dG}^{(l)}}{\sqrt{\mathbf{dG}^{(l)\top}\mathbf{dS}^{(l)}}},\ \mathbf{u}_2^{(l)} = \frac{\mathbb{B}^{(l)}\mathbf{dS}^{(l)}}{\sqrt{\mathbf{dS}^{(l)\top}\mathbb{B}_0^{(l)}\mathbf{dS}^{(l)}}}.$$

If we simply choose $\mathbb{B}^{(l)}$ to be a constant diagonal matrix, we recover our first-order proximal gradient algorithm.

Note that an arbitrary step of updates can be implemented for these two schemes, which generally do not cost much more space due to the compact representation formula available(Nocedal and Wright, 2006). However, LSR1 generally with more steps generally performs less effectively than LBFGS due to the indefiniteness issue.

- Steps to update the stepsize. Many powerful schemes are suggested for updating the stepsize by the literature. We adopt a line search procedure with backtracking(Boyd et al., 2004). See Section 3 of Lee et al. (2014) for more details.

- Solver for the subproblem in Step 4. Step 4 breaks down to a "quadratic + nuclear norm penalty " type problem, which can be efficiently solved by first-order solvers. In our implementation, we adopt our Algorithm 1 which, in the case of quadratic smooth part, performs in the same spirit as the NNLS method by Toh and Yun (2010).

Below we run some synthetic simulation based on the multivariate regression model($m = 80, N = 6320$, and the ground truth is generated in the same way as the simulation in our main context) to compare the computational efficiency and statistical accuracy of first-order and second-order algorithms. Table S.4 presents the comparison of first-order and second-order implementations(with 1-step limited memory SR update and 1-step limited memory BFGS update, termed as **LSR1** and **LBFGS** respectively). In Table S.5 we compare BFGS approximation with different steps of storage. The "Error" column in the tables is measured by $\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_F^2$.

Method	**Criterion**	**Noise**		
		Gaussian	**Cauchy**	**Lognormal**
First order	**Error**	9.48e-2(3.97e-3)	3.21e-4(1.23e-5)	1.28e-6(1.59e-7)
	Rank	5.00(0)	5.00(0)	5.00(0)
	Iterations	32.0	35.3	56.6
	Time(s)	2.22	2.51	3.88
LSR1	**Error**	9.48e-2(3.97e-3)	3.21e-4(1.29e-5)	1.48e-6(3.30e-7)
	Rank	5.00(0)	5.00(0)	5.00(0)
	Iterations	28.1	31.8	51.7
	Time(s)	4.81	4.77	7.48
LBFGS	**Error**	9.48e-2(3.98e-3)	3.20e-4(1.36e-5)	1.49e-6(3.38e-7)
	Rank	5.00(0)	5.00(0)	5.00(0)
	Iterations	24.5	29.2	35.3
	Time(s)	4.19	4.33	4.95

Table S.4: Comparison of first order and second order implementation of RML

Method	Criterion	Noise		
		Gaussian	Cauchy	Lognormal
1-step LBFGS	**Error**	9.52e-2(4.11e-3)	3.19e-4(1.35e-5)	1.45e-6(3.04e-7)
	Rank	5.00(0)	5.00(0)	5.00(0)
	Iterations	24.5	29.4	35.9
	Time(s)	4.15	4.34	5.06
5-step LBFGS	**Error**	9.52e-2(4.12e-3)	3.19e-4(1.26e-5)	1.41e-6(2.34e-7)
	Rank	5.00(0)	5.00(0)	5.00(0)
	Iterations	25.3	30.8	41.1
	Time(s)	4.26	4.59	6.22
10-step LBFGS	**Error**	9.52e-2(4.12e-3)	3.19e-4(1.25e-5)	1.34e-6(1.84e-7)
	Rank	5.00(0)	5.00(0)	5.00(0)
	Iterations	25.2	31.1	42.1
	Time(s)	4.25	4.65	6.51

Table S.5: Comparison of BFGS with different steps of approximation for multivariate regression

Here is a brief summary for the tables:

- **Accuracy of solutions.** As we can see from the results, under the specified tolerance, all methods give quite accurate solutions. There is no clear superiority among the tested solvers. For second-order methods, both rank-1 update(LSR1) and rank-2 update(LBFGS) are able to recover the true estimand in a decent manner.

- **Number of iterations.** In terms of total steps of iterations, generally the rank-1 and rank-2 update(LBFGS) can reduce the steps that the algorithm requires to converge under a variety of noisy perturbations. The rank-2 update(LBFGS) appears more powerful than rank-1 update(LSR1) in accelerating the convergence.

- **Time per iteration and total running time.** Although the second-order schemes witness a reduction in the steps needed for convergence, they pay the cost of additional time consumption per iteration step. This is mostly due to the fact that one needs to seek the optimal solution for a sub-problem(see Algorithm 2 step 4), which again relies on first-order solvers, thus leading to more rounds of singular value decomposition. It is notable that a reduction in the number of iteration steps does lend a large improvement in sparse vector-variate problems, since solving L_1 penalization breaks down to subproblems that involves much less computation burden than matrix settings(Lee et al., 2014; Becker and Fadili, 2012). Therefore, the blessing of iteration steps and the curse of per step convergence time become a negligible trade-off in nuclear norm penalization problems.

- **Sensitivity to the choice of initial curvature.** First order methods can be regarded as second order methods with an extremely simple Hessian approximation

scheme(constant diagonal). However, it is in general trickier to specify an initial curvature parameter. A small specification might render the iteration divergent, while an overly large number would shrink the length of steps the algorithm can take. However, second-order programs can cleverly circumvent this issue since they are endowed with the capability of adapting themselves gradually to the appropriate scale of the Hessian structure. Therefore, while more time is required in our matrix-variate optimization settings, we do further free ourselves from the dilemma of curvature tuning and obtain more numerical stability of the solutions.

- **Space of memory for the second-order implementation.** The second-order limited memory implementation like BFGS can proceed with different space of memory to store the Hessian information within each step. While the statistical accuracy appears less sensitive to the Hessian memory space, more complex Hessian structure could add more numerical stability in certain cases(see the Cauchy and Lognormal column of Table S.5).

H3: Influence of the Choice of α_0 on the Estimation Error

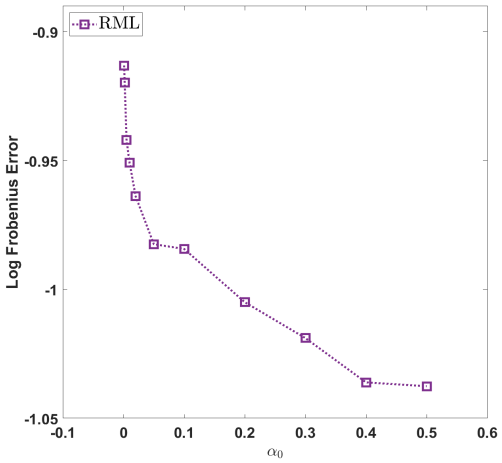
In our theory, the number $1-\alpha_0$ can be regarded as the "confidence level" in the sense that our non-asymptotic bounds on the estimation error will be controlled at the optimal rate with probability close to $1-\alpha_0$. To see how the size of α_0 influence the estimation error, we conduct some simulations in Figure S.7 under the same setting as Example 4.1 with $m_1 = m_2 = 40$. We show the estimation error of our estimator. In the three pictures on the left, we consider three errors, fixed sample size $n = 1600$, and varying α_0 belonging to $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For different random errors, when α_0 is not very small, the estimation error is not so sensitive to the choice of α_0. But when α_0 approaches zero, the estimation error will increase rapidly, because the selected λ will be relatively large. In the three pictures on the right, we consider three errors, three α_0's and varying sample sizes. The α_0 affects estimation error much less than sample size. In order for the optimal estimation rate to hold with high probability and good numerical performance, we suggest taking $1-\alpha_0 \in [0.8, 0.9]$.

H4: Vertical Outliers and Leverage Points

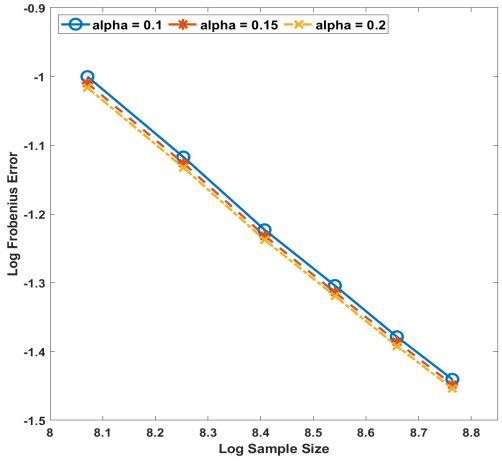
In robust statistics, vertical outliers and leverage points are also worthy of attention. In this section, we compare the performance of different methods in these two cases.

For the vertical outliers, we consider the same setting as Example 4.1 with $m_1 = m_2 = 40, n = 3200$. The random errors are sampled from $\mathcal{N}(0, 0.25) + a \cdot \mathcal{N}(10, 1)$, where a follows the binomial distribution $bino(1, \pi)$. The π represents the average proportion of vertical outliers. We show four methods' performance at different vertical outliers proportions in the left of Figure S.8. Our method is still robust to vertical outliers.

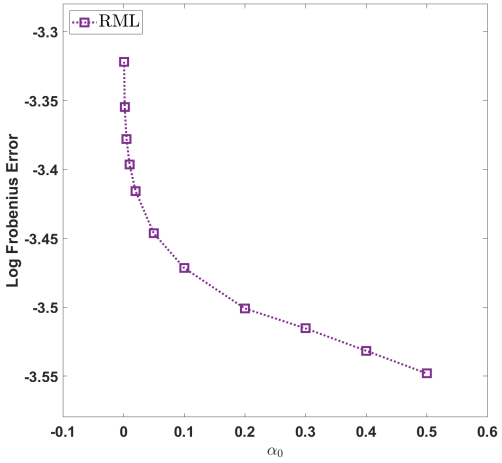
Our method is not applicable to situations where leverage points exist. In our method, the selection of regularization parameter λ is related to the covariates. The existence of leverage points may cause λ to be very large, affecting the estimation error. We consider the settings in Example 4.1: $m_1 = m_2 = 40, n = 3200$, and the error is Gaussian $\mathcal{N}(0, 0.25)$.



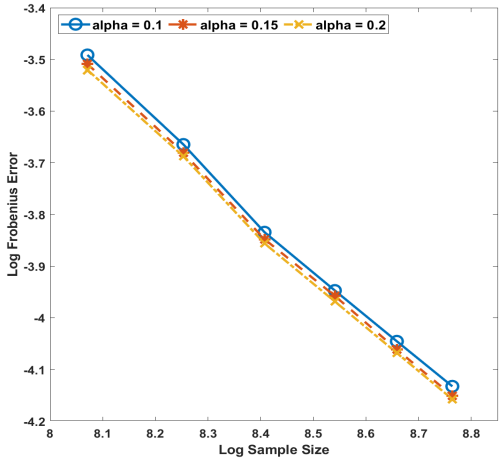
(a) Gaussian



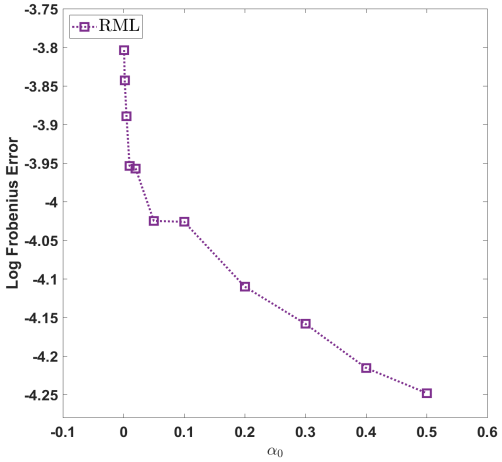
(b) Gaussian



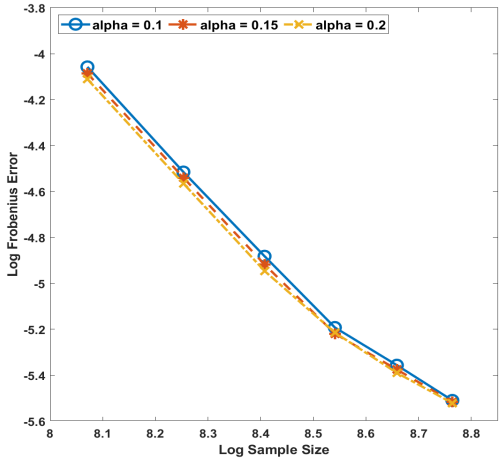
(c) Cauchy



(d) Cauchy



(e) Lognormal



(f) Lognormal

Figure S.7: The influence of the choice of α_0 on estimation error of RML

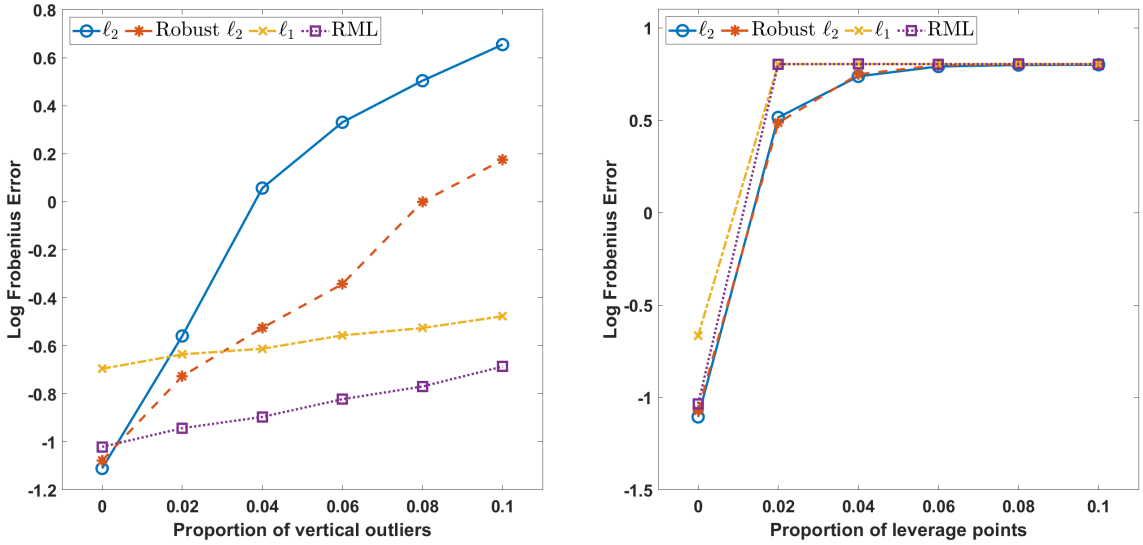


Figure S.8: The influence of vertical outliers and leverage points on estimation error

Consider that $\mathbf{X}$ is generated in the following way,

$$\mathbf{X} = \mathbf{Z}_1 + a \cdot \mathbf{Z}_2,$$

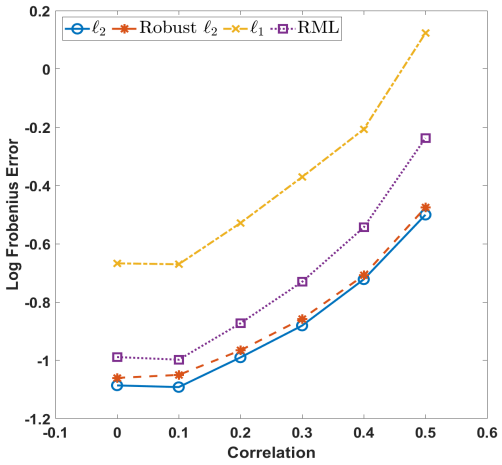
where $\mathbf{Z}_1$ is composed of $\mathcal{N}(0,1)$ entries, $\mathbf{Z}_2$ is composed of $\mathcal{N}(10,1)$ entries, and a follows the binomial distribution $bino(1,\pi)$. The π represents the average proportion of leverage points. We show four methods' performance at different leverage point proportions in the right of Figure S.8. As the proportion of leverage points increases, the performance of all methods collapses rapidly.

H5: Additional Simulations for Examples 4.1

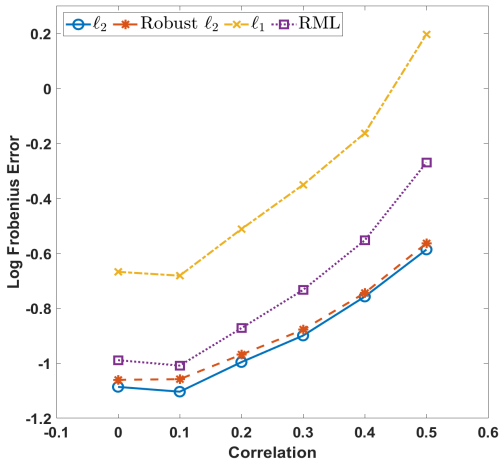
In this section, we present the additional simulation results for Examples 4.1 in the main text. We compare the performance of the proposed method with its competitors when the features come from highly correlated designs with different correlation strengths. In Example 4.1, we consider two correlation structures under sample size $n = 3200$: (i) The autoregressive covariance structure: $\mathrm{Cov}(\mathrm{vec}(\mathbf{X})) = (a_{ij})$, $a_{ij} = a^{|i-j|}$. (ii) The blockwise diagonal Toeplitz covariance matrix, of which each block along the diagonal is set to be

$$\begin{bmatrix} 1 & \frac{(p-2)a}{p-1} & \frac{(p-3)a}{p-1} & \cdots & \frac{a}{p-1} & 0 \\ \frac{(p-2)a}{p-1} & 1 & \frac{(p-2)a}{p-1} & \cdots & \frac{2a}{p-1} & \frac{a}{p-1} \\ \vdots & & \ddots & & & \vdots \\ 0 & \frac{a}{p-1} & \frac{2a}{p-1} & \cdots & \frac{(p-2)a}{p-1} & 1 \end{bmatrix},$$

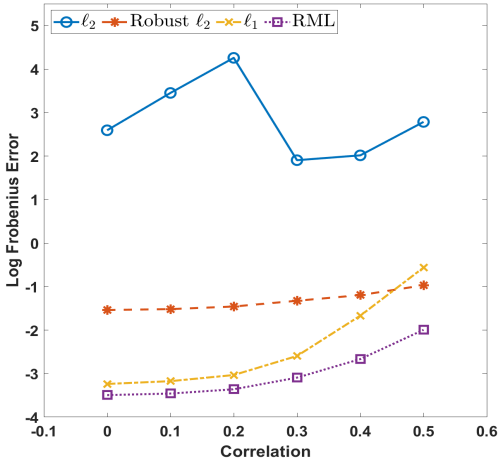
where $p = m_1 m_2/200$. Throughout, we refer $a \in (0,1)$ as the correlation strength. We show the performance of different methods under varying correlation strengths in Figure S.9. As the correlation increases, the estimation errors of all methods rise, and the proposed rank matrix lasso method still outperforms all competitors.



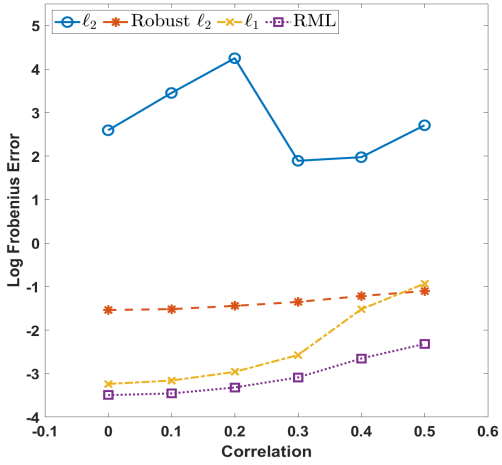
(a) Gaussian AR



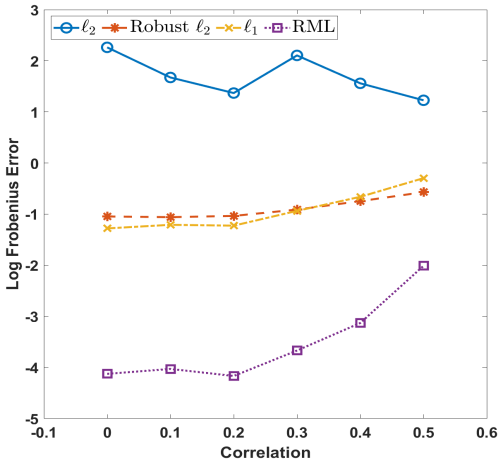
(b) Gaussian Toep



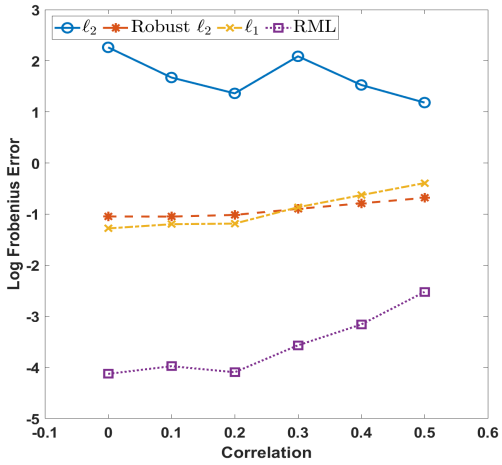
(c) Cauchy AR



(d) Cauchy Toep



(e) Lognormal AR



(f) Lognormal Toep

Figure S.9: Log Frobenius Errors for matrix regression under different correlations

References

James Baglama and Lothar Reichel. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, 27(1):19–42, 2005.

Stephen Becker and Jalal M Fadili. A quasi-newton proximal splitting method. In *Neural Information Processing Systems (NIPS) 2012*, volume 25, pages 2618–2626, 2012.

Stephen Becker, Jalal Fadili, and Peter Ochs. On quasi-newton forward-backward splitting: Proximal calculus and convergence. *SIAM Journal on Optimization*, 29(4):2445–2481, 2019.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Wei Bian and Xiaojun Chen. A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. *SIAM Journal on Numerical Analysis*, 58(1): 858–883, 2020.

Wei Bian and Fan Wu. Accelerated forward-backward method with fast convergence rate for nonsmooth convex optimization beyond differentiability. *arXiv preprint arXiv:2110.01454*, 2021.

Xin Bing, Marten H Wegkamp, et al. Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *The Annals of Statistics*, 47(6):3157–3184, 2019.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40 (5):2359–2388, 2012.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Antonin Chambolle and Ch Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization theory and Applications*, 166 (3):968–982, 2015.

Xi Chen, Weidong Liu, Xiaojun Mao, and Zhuoyi Yang. Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, 21 (182):1–43, 2020.

Stéphan Clémençon, Gábor Lugosi, Nicolas Vayatis, et al. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

James B Davis and Joseph W McKean. Rank-based methods for multivariate linear models. *Journal of the American Statistical Association*, 88(421):245–251, 1993.

Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *The Annals of Statistics*, 46(6B):3481–3509, 2018.

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

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2):814–841, 2018.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *The Annals of Statistics*, 49(3):1239–1266, 2021.

Maryam Fazel, E Candes, Benjamin Recht, and P Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2008.

Mohammad Golbabaee and Pierre Vandergheynst. Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2741–2744, 2012.

Ulf Grenander and Gabor Szegö. *Toeplitz forms and their applications*. Berkeley and Los Angeles: University of California Press, 1958.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Yuwen Gu and Hui Zou. Sparse composite quantile regression in ultrahigh dimensions with tuning parameter calibration. *IEEE Transactions on Information Theory*, 66(11): 7132–7154, 2020.

TP Hettmansperger and JW McKean. *Robust Nonparametric Statistical Methods*. London: Arnold., 1998.

Wei Hu, Weining Shen, Hua Zhou, and Dehan Kong. Matrix linear discriminant analysis. *Technometrics*, 62(2):196–205, 2020.

Wei Hu, Tianyu Pan, Dehan Kong, and Weining Shen. Nonparametric matrix response regression with application to brain imaging data analysis. *Biometrics*, 77(4):1227–1240, 2021.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Roger Koenker. *Quantile regression*. New York: Cambridge University Press, 2005.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Dehan Kong, Baiguo An, Jingwen Zhang, and Hongtu Zhu. L2rm: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association*, 115(529):403–424, 2020.

Rasmus Munk Larsen. Lanczos bidiagonalization with partial reorthogonalization. *DAIMI Report Series*, 27(537), 1998.

Rasmus Munk Larsen. Propack-software for large and sparse svd calculations, 2004.

Michael Law, Ya'acov Ritov, Ruixiang Zhang, and Ziwei Zhu. Rank-constrained least-squares: Prediction and inference. *arXiv preprint arXiv:2111.14287*, 2021.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Berlin: Springer Science & Business Media, 2013.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Luong Trung Nguyen, Junhan Kim, Sangtae Kim, and Byonghyo Shim. Localization of iot networks via low-rank matrix completion. *IEEE Transactions on Communications*, 67 (8):5833–5847, 2019.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

Michael I Parzen, Lee-Jen Wei, and Zhiliang Ying. A resampling method based on pivotal estimating functions. *Biometrika*, 81(2):341–350, 1994.

Andy Ramlatchan, Mengyun Yang, Quan Liu, Min Li, Jianxin Wang, and Yaohang Li. A survey of matrix completion methods for recommendation systems. *Big Data Mining and Analytics*, 1(4):308–323, 2018.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8:45–114, 2014.

Yiyuan She and Kun Chen. Robust reduced-rank regression. *Biometrika*, 104(3):633–647, 2017.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Kean Ming Tan, Qiang Sun, and Daniela Witten. Sparse reduced rank huber regression in high dimensions. *Journal of the American Statistical Association*, pages 1–11, 2022.

Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021a.

Tian Tong, Cong Ma, and Yuejie Chi. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409, 2021b.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

MJ Trudel and JL Ozbun. Relationship between chlorophylls and carotenoids of ripening tomato fruit as influenced by potassium nutrition. *Journal of Experimental Botany*, 21 (4):881–886, 1970.

Sara A Van de Geer. *Applications of empirical process theory*, volume 91. Cambridge University Press Cambridge, 2000.

Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. New York: Springer, 1996.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge: Cambridge university press, 2018.

Lan Wang and Runze Li. Weighted wilcoxon-type smoothly clipped absolute deviation method. *Biometrics*, 65(2):564–571, 2009.

Lan Wang, Bo Peng, Jelena Bradic, Runze Li, and Yunan Wu. A tuning-free robust and efficient approach to high-dimensional regression. *Journal of the American Statistical Association*, 115:1–44, 2020.

Anja Wille, Philip Zimmermann, Eva Vranová, Andreas Fürholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelić, Peter von Rohr, Lothar Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology*, 5(11):1–13, 2004.

Dengdeng Yu, Linbo Wang, Dehan Kong, and Hongtu Zhu. Mapping the genetic-imaging-clinical pathway with applications to alzheimer's disease. *Journal of the American Statistical Association*, pages 1–13, 2022.

Jin Yu, SVN Vishwanathan, Simon Günter, and Nicol N Schraudolph. A quasi-newton approach to nonsmooth convex optimization problems in machine learning. *Journal of Machine Learning Research*, 11:1145–1200, 2010.

Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346, 2007.

Chao Zhang and Xiaojun Chen. Smoothing projected gradient method and its application to stochastic linear complementarity problems. *SIAM Journal on Optimization*, 20(2): 627–649, 2009.

Yichi Zhang, Weining Shen, and Dehan Kong. Covariance estimation for matrix-valued data. *Journal of the American Statistical Association*, pages 1–12, 2022.

Jianwei Zheng, Mengjie Qin, HongChuan Yu, and Wanliang Wang. An efficient truncated nuclear norm constrained matrix completion for image inpainting. In *Proceedings of Computer Graphics International 2018*, pages 97–106. Association for Computing Machinery, 2018.

Le Zhou, Boxiang Wang, and Hui Zou. Sparse convoluted rank regression in high dimensions. *Journal of the American Statistical Association*, just-accepted:1–27, 2023.