Rank Matrix LASSO: a tuning-free robust scheme for high-dimensional low-rank matrix estimation

Xiaolong Cui[1], Lei Shi[1], Wei Zhong[2] and Changliang Zou[1]

[1]*School of Statistics and Data Science, Nankai University, China*

[2]*Department of Statistics and Data Science, Xiamen University, China*

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 outlying observations and heavy-tailed distributions. This paper introduces a robust procedure by incorporating a Wilcoxon-type rank 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. Fourth, we develop an accelerated proximal gradient descent algorithm that is able to solve the rank loss minimization problem and yield accurate results with low computation expenses. 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

$$\mathbf{y} = \mathfrak{X}\left(\mathbf{X}; \mathbf{A}_0\right) + \boldsymbol{\varepsilon}, \tag{1.1}$$

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 $p = 1$, $\mathfrak{X}\left(\mathbf{X}; \mathbf{A}\right) = \operatorname{tr}\left(\mathbf{A}^\top \mathbf{X}\right)$ and $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$ is a matrix of explanatory variables, the linear operator model (1.1) 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

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A} \in \mathcal{S}} \left\{ Q_n\left(\mathbf{A}\right) + \lambda P\left(\mathbf{A}\right) \right\}, \tag{1.2}$$

where $\mathcal{S}$ is a convex set in $\mathbb{R}^{m_1 \times m_2}$, $Q_n\left(\mathbf{A}\right)$ is an empirical loss function, λ is a tuning parameter and $P\left(\mathbf{A}\right)$ 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) establish a nearly optimal in-sample prediction risk bound for the rank-constrained least-squares estimator under no assumptions on the design matrix. For compressed sensing, 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), She (2017), Bing et al. (2019) 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 include Jain et al. (2013), Sun and Luo (2016), Ma et al. (2018), Fan et al. (2021a) and Tong et al. (2021a,b), which are based on matrix factorization framework and stand beyond the consideration of the penalized estimation framework of the current paper. Despite significant advances in theoretical developments, at least two challenges about robustness and tuning parameter selection still remain.

The first challenge is how to deal with heavy-tailed or skewed random errors for low-rank matrix recovery methods in high dimensions. Heavy-tailed error contamination often deteriorate the robustness of the existing quadratic loss based methods which are sensitive to outliers. Fan et al. (2021b) introduced a shrinkage principle by truncating or shrinking appropriately the response variables and the designs to handle heavy-tailed random errors and covariates. 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 (CV) method is inevitably required to select the regularization parameter λ and the truncation level together, which is time-consuming and lacks theoretical guarantee. Similar problems are also found in the methods based on Huber loss (Fan et al., 2017; Elsener and van de Geer, 2018; Tan et al., 2018; Sun et al., 2020). Elsener and van de Geer (2018), Ma and Fattahi (2021) and Tong et al. (2021b) investigated the least absolute deviation (LAD) loss for low-rank matrix recovery problem to alleviate the influence of heavy-tailed random errors. Although being robust, the LAD estimator typically yields a less favorable result under normal errors. She and Chen (2017) considers a robust reduced rank regression framework which could bridge the worlds of penalization and robust M-estimation. However, their methodology starts from an additive outlier characterization, which is essentially different from our straightforward noise quantification.

The second issue is how to select the tuning parameter λ for the regularization methods (or similarly a pre-specified rank in matrix factorization framework) in a computationally efficient way with theoretical guarantee. The existing literature either only give the rate of the parameter, or only give the parameter value empirically for a specific setting without theoretical guarantee which is not applicable in practice, see, for example, Elsener and van de Geer (2018), Chen et al. (2020) and Fan et al. (2021b). Commonly used cross-validation (CV) or information criteria techniques are computationally inefficient. Negahban and Wainwright (2011) provided some insight into the choice of regularization parameter and suggested that λ is proportional to the operator norm of a random matrix defined by the gradient of loss function evaluated at $\mathbf{A}_0$ which depends on both the random error and the design. For example, their choice of λ for low-rank multivariate regression is $10\sigma m_1^{-1}\sqrt{\lambda_{\max}(\mathbf{\Sigma})}\sqrt{n^{-1}(m_1+m_2)}$, where σ is the standard deviation of random error, $\mathbf{\Sigma}$ is the covariance matrix of covariates and $\lambda_{\max}(\mathbf{\Sigma})$ denotes the largest eigenvalue of $\mathbf{\Sigma}$. Unfortunately, the quantities σ and $\mathbf{\Sigma}$ are usually unknown in practice, and the appropriateness of the constant "10" may vary across the error distributions. To circumvent this difficulty, Klopp (2014) proposed the square-root matrix LASSO to alleviate the influence of σ, which adopts the idea of square-root LASSO in the context of linear regression (Belloni et al., 2011), but the choice of λ still relies on the error distributions. She and Tran (2019) proposed a structural CV for jointly row sparse and rank deficient multivariate regression model. They showed that the prediction risk of their estimator achieves the minimax optimal rate, but an optimal estimation guarantee is absent for that CV-based method.

More critically, these two challenges are usually intertwined. Solutions specifically designed for only one aspect of the two challenges could leave the other aspect more unsatisfactory. In a recent work, Wang et al. (2020) proposed a tuning-free robust and efficient approach to deal with these two challenges specifically for high-dimensional linear regression. They developed a novel rank LASSO based on a penalized Wilcoxon-type loss. They showed that the new rank LASSO estimator overcomes the tuning parameter selection difficulty and is robust with substantial efficiency gain for heavy-tailed random errors. However, this approach can not be directly applied to high dimensional low-rank matrix recovery problems and the extension is nontrivial from either a theoretical or a computational perspective.

In this paper, we propose a tuning-free robust procedure, termed as Rank Matrix LASSO, in a unified high-dimensional low-rank matrix estimation framework. It is able to simultaneously tackle the two challenges and 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.

- From the methodology aspect, a new tuning-free robust method incorporates a Wilcoxon-type rank loss function with the nuclear-norm penalty for the low-rank matrix estimation model (1.1). It relaxes the distributional conditions on random errors from sub-exponential or sub-Gaussian to more general distributions. It is not only robust with substantial efficiency gain for heavy-tailed error distributions and outliers, like Rank LASSO in Wang et al. (2020), but also more applicable for many popular high-dimensional low-rank matrix recovery problems including matrix regression, multivariate linear regression and matrix completion, among others.

- 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-free. Moreover, we extend the development of optimization with the nuclear-norm penalty (Nesterov, 2013; Ji and Ye, 2009; Toh and Yun, 2010) to introduce a proximal gradient descent algorithm which can effectively cope with the minimization in the Rank Matrix LASSO and yield accurate results with low computation expenses.

- 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 random errors, the proposed robust estimator is shown to achieve the same rates as those derived by Negahban and Wainwright (2011), Negahban and Wainwright (2012), Klopp (2014). Technical arguments for extending the existing trace 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 with an easy-to-implement algorithm. 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. Section 5 concludes the paper. Technical proofs and additional simulations results are provided in the Supplementary Material.

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, $\|\mathbf{A}\|_{\mathrm{op}}$ 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 $\|\mathbf{x}\|_{\psi_p} := \sup_{\mathbf{v} \in \mathcal{D}^{d-1}} \left\|\mathbf{v}^\top \mathbf{x}\right\|_{\psi_p}$, where $\mathcal{D}^{d-1}$ is the d-dimensional unit sphere. Let $\mathbf{e}_k(m)$ be the k-th m-dimensional unit vector and $\sum_{i \neq j} := \sum_{i=1}^n \sum_{j=1, j \neq i}^n$.

2 Methodology

2.1 Model

Consider n independent observations collected from the linear operator model (1.1)

$$\mathbf{y}_i = \mathfrak{X}(\mathbf{X}_i; \mathbf{A}_0) + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{2.1}$$

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

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

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 2.1 (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*

$$y_i = \langle \mathbf{X}_i, \mathbf{A}_0 \rangle + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.2}$$

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,* $\mathfrak{X}\left(\mathbf{X}_i; \mathbf{A}_0\right) = \langle \mathbf{X}_i, \mathbf{A}_0 \rangle = \operatorname{tr}\left(\mathbf{X}_i^\top \mathbf{A}_0\right)$ *and* $p = 1$*. Moreover, model* (2.2) *includes the high-dimensional linear regression model considered by Wang et al. (2020) 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 2.2 (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*

$$\mathbf{y}_i = \mathbf{A}_0 \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{2.3}$$

where $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{m_1}$*. We can write the multivariate regression as an instance of the linear operator model* (2.1) *with* $\mathbf{X}_i = \mathbf{x}_i \in \mathcal{L} = \mathbb{R}^{m_2}$ *and* $\mathfrak{X}\left(\mathbf{x}_i; \mathbf{A}_0\right) = \mathbf{A}_0 \mathbf{x}_i$*.*

Example 2.3 (Matrix completion)**.** *The matrix completion problem can be formulated into the trace inner products model (2.2), where the matrices* $\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}$ *are so-called masks. They are assumed to lie in*

$$\mathcal{X} = \left\{\mathbf{e}_k\left(m_1\right) \mathbf{e}_l^\top\left(m_2\right) : 1 \le k \le m_1, 1 \le l \le m_2\right\}. \tag{2.4}$$

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 (2.1), we consider a new estimator of $\mathbf{A}_0$ by minimizing the following penalized loss function

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A}\in\mathcal{S}} \left\{ Q_n\left(\mathbf{A}\right) + \lambda \left\|\mathbf{A}\right\|_1 \right\}, \tag{2.5}$$

where the loss function is defined as

$$\begin{aligned} Q_n\left(\mathbf{A}\right) &= \left\{n\left(n-1\right)\right\}^{-1} \sum_{i\neq j} \left\|\boldsymbol{\varepsilon}_i\left(\mathbf{A}\right) - \boldsymbol{\varepsilon}_j\left(\mathbf{A}\right)\right\|_1 \\ &= \left\{n\left(n-1\right)\right\}^{-1} \sum_{i\neq j}\sum_{k=1}^{p} \left|\varepsilon_{ik}\left(\mathbf{A}\right) - \varepsilon_{jk}\left(\mathbf{A}\right)\right| \\ &= \left\{n\left(n-1\right)\right\}^{-1} \sum_{i\neq j}\sum_{k=1}^{p} \left|\left\{y_{ik} - \mathbf{e}_k^\top\left(p\right)\mathfrak{X}\left(\mathbf{X}_i;\mathbf{A}\right)\right\} - \left\{y_{jk} - \mathbf{e}_k^\top\left(p\right)\mathfrak{X}\left(\mathbf{X}_j;\mathbf{A}\right)\right\}\right|, \end{aligned} \tag{2.6}$$

$\boldsymbol{\varepsilon}_i\left(\mathbf{A}\right) = \mathbf{y}_i - \mathfrak{X}\left(\mathbf{X}_i;\mathbf{A}\right)$, $\varepsilon_{ik}\left(\mathbf{A}\right)$ and y_{ik} are the k-th elements of $\boldsymbol{\varepsilon}_i\left(\mathbf{A}\right)$ and $\mathbf{y}_i$, respectively, λ denotes the tuning parameter and $\mathcal{S}$ is a convex constraint set which is determined based on the concrete settings. 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 unique minimizer of $Q\left(\mathbf{A}\right)$ under some weak conditions.

The loss function in (2.6) is an extension of the univariate rank based method which originates from classical nonparametric statistics (Hettmansperger and McKean, 1998). It has been shown that minimizing this loss function is equivalent to minimizing Jaeckel's Wilcoxon-type dispersion function (Jaeckel, 1972),

$$\sqrt{12}\sum_{k=1}^{p}\sum_{i=1}^{n}\left[\frac{R\left(\varepsilon_{ik}\left(\mathbf{A}\right)\right)}{n+1} - \frac{1}{2}\right]\cdot\varepsilon_{ik}\left(\mathbf{A}\right)$$

where $\varepsilon_{ik}\left(\mathbf{A}\right) = y_{ik} - \mathbf{e}_k^\top\left(p\right)\mathfrak{X}\left(\mathbf{X}_i;\mathbf{A}\right)$ and $R\left(\varepsilon_{ik}\left(\mathbf{A}\right)\right)$ denotes the rank of $\varepsilon_{ik}\left(\mathbf{A}\right)$ among $\varepsilon_{1k}\left(\mathbf{A}\right),\ldots,\varepsilon_{nk}\left(\mathbf{A}\right)$. 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 $\frac{R(\varepsilon_{ik}(\mathbf{A}))}{n+1} - \frac{1}{2}$ to weight $\varepsilon_{ik}\left(\mathbf{A}\right)$, while ℓ_2 loss and ℓ_1 loss use $\varepsilon_{ik}\left(\mathbf{A}\right)$ and sign($\varepsilon_{ik}\left(\mathbf{A}\right)$) as weights respectively. In light of this observation, the weight $\frac{R(\varepsilon_{ik}(\mathbf{A}))}{n+1} - \frac{1}{2}$ can be regarded as a balance between weight $\varepsilon_{ik}\left(\mathbf{A}\right)$ and weight sign($\varepsilon_{ik}\left(\mathbf{A}\right)$). Intuitively, outliers will not have as much impact on $\frac{R(\varepsilon_{ik}(\mathbf{A}))}{n+1} - \frac{1}{2}$ 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 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 rank loss function for high-dimensional linear regression. A natural question is whether the new estimator $\widehat{\mathbf{A}}$ in (2.5) 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 Wilcoxon-type rank loss function with the nuclear-norm penalty in (2.5) 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 λ

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-free 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.

By the definition of $Q_n(\mathbf{A})$, we have

$$Q_n(\mathbf{A}) = \{n(n-1)\}^{-1} \sum_{k=1}^{p} \sum_{i \neq j} \left| y_{ik} - y_{jk} - \mathbf{e}_k^\top (p) \mathfrak{X}(\mathbf{X}_i - \mathbf{X}_j; \mathbf{A}) \right|$$

and accordingly the gradient of $Q_n(\mathbf{A})$ evaluated at $\mathbf{A}_0$

$$\nabla Q_n(\mathbf{A}_0) = -\left\{n(n-1)\right\}^{-1} \sum_{k=1}^{p} \sum_{i \neq j} (\mathbf{H}_{ik} - \mathbf{H}_{jk}) \operatorname{sign}(\varepsilon_{ik} - \varepsilon_{jk}),$$

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

$$\nabla Q_n(\mathbf{A}_0) = -2\left\{n(n-1)\right\}^{-1} \sum_{k=1}^{p} \sum_{i=1}^{n} \mathbf{H}_{ik} \sum_{j=1, j\neq i}^{n} \operatorname{sign}(\varepsilon_{ik} - \varepsilon_{jk}).$$

Let $\xi_{ik} = \sum_{j=1,j\neq i}^{n} \operatorname{sign}(\varepsilon_{ik} - \varepsilon_{jk})$ for $i = 1, \ldots, n, k = 1, \ldots, p$. It is important to observe that ξ_{ik} is closely related to the rank of ε_{ik} among $\{\varepsilon_{1k}, \ldots, \varepsilon_{nk}\}$. Let $\operatorname{rank}(\varepsilon_{ik}) := R_{ik}$, then $\{R_{1k}, R_{2k}, \ldots, R_{nk}\}$ follows the uniform distribution on the permutations of the integers $\{1, 2, \ldots, n\}$. Write

$$\xi_{ik} = \sum_{j=1, j\neq i}^{n} \operatorname{sign}(\varepsilon_{ik} - \varepsilon_{jk}) = 2R_{ik} - (n+1).$$

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

$$\lambda \geq 2\left\|\nabla Q_n(\mathbf{A}_0)\right\|_{\mathrm{op}}, \tag{2.7}$$

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 (2.7) 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

$$\lambda^* = 2G_{S_n}^{-1}(1 - \alpha_0), \tag{2.8}$$

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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

2.4 Algorithms

In this subsection, we propose an accelerated proximal gradient (APG) algorithm to solve the minimization (2.5). Specifically, to minimize a penalized loss function, i.e.,

$$\min_{\mathbf{A}} \ \{Q_n(\mathbf{A}) + \lambda\|\mathbf{A}\|_1\}, \tag{2.9}$$

we employ the quadratic function

$$Q_{\text{Major}}(\mathbf{A};\mathbf{A}^{(l)}) = Q_n(\mathbf{A}^{(l)}) + \left\langle \nabla Q_n(\mathbf{A}^{(l)}),\, \mathbf{A}-\mathbf{A}^{(l)}\right\rangle + (L/2)\left\|\mathbf{A}-\mathbf{A}^{(l)}\right\|_F^2 \tag{2.10}$$

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

$$\min_{\mathbf{A}} L_{\text{Major}}(\mathbf{A};\mathbf{A}^{(l)}) = \min_{\mathbf{A}} \ \left\{Q_{\text{Major}}(\mathbf{A};\mathbf{A}^{(l)}) + \lambda\|\mathbf{A}\|_1\right\} \tag{2.11}$$

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

$$Q_n(\mathbf{A}^{(l+1)}) \leq Q_{\text{Major}}(\mathbf{A}^{(l+1)};\mathbf{A}^{(l)}) \leq Q_{\text{Major}}(\mathbf{A}^{(l)};\mathbf{A}^{(l)}) = Q_n(\mathbf{A}^{(l)}).$$

While typically (2.9) does not have a closed-form solution, the minimizer in (2.11) 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.

To tackle the computational issue due to the U-statistic structure, we rearrange the loss in (2.5) into a weighted summation. Recall the equivalent expression for the gradient given in Section 2.3:

$$\nabla Q_n(\mathbf{A}) = -2\{n(n-1)\}^{-1}\sum_{k=1}^{p}\sum_{i=1}^{n}\mathbf{H}_{ik}\xi_{ik}(\mathbf{A}), \tag{2.12}$$

where $\xi_{ik}(\mathbf{A}) = \sum_{j=1,j\neq i}^{n}\text{sign}(\varepsilon_{ik}(\mathbf{A}) - \varepsilon_{jk}(\mathbf{A}))$ for $i = 1,\ldots,n, k = 1,\ldots,p$. Similar calculation yields a counterpart for the loss function:

$$Q_n(\mathbf{A}) = 2\{n(n-1)\}^{-1}\sum_{k=1}^{p}\sum_{i=1}^{n}\xi_{ik}(\mathbf{A})\left\{y_{ik} - \mathbf{e}_k^{\top}(p)\mathfrak{X}(\mathbf{X}_i;\mathbf{A})\right\}. \tag{2.13}$$

Using the weighted version, we only need to deal with $O(n)$ terms in the summation, plus the cost $O\{n\log(n)\}$ for sorting the residuals.

Remark 2.1. *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 G1 of the Supplementary Material, 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.*

```
Algorithm 1: Proximal gradient algorithm for the Rank Matrix LASSO
  Input: Observed data $(y_i, \mathbf{X}_i)$, for $i = 1, \cdots, n$; tuning parameter $\lambda^\star$ by (2.8); floor
         curvature $L_0$; ceiling curvature $L_{max}$; updating rate $\eta$; $t^{(0)} = t^{(-1)} = 1$;
         convergence tolerance tol and maximal iteration $T$; initial estimate $\mathbf{A}^{(0)}$.
  Output: Estimator $\widehat{\mathbf{A}}$.
  Set $\mathbf{B}^{(l)} = \mathbf{A}^{(l)} + \frac{t^{(l-1)}-1}{t^{(l)}}(\mathbf{A}^{(l)} - \mathbf{A}^{(l-1)})$.
  Set $L = L_0$. Calculate the sub-gradient $\mathbf{G}^{(l)} = \nabla Q_n(\mathbf{A})|_{\mathbf{A}=\mathbf{A}^{(l)}}$ using (2.12).
  Set $L = \min\{\eta L, L_{max}\}$.
  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 $Q_{\text{Major}}(\mathbf{S}; \mathbf{B}^{(l)})$ using (2.10) (2.12) and (2.13);
  if $Q_n(\mathbf{S}) \le Q_{\text{Major}}(\mathbf{S}; \mathbf{B}^{(l)})$ then
      Set $\mathbf{A}^{(l+1)} = \mathbf{S}$.
  else
      Set $\mathbf{A}^{(l+1)} = \mathbf{A}^{(l)}$.
      Set $L = L/\eta$.
  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 2.2. *While "smooth + nonsmooth" optimization has been extensively studied (Nes-*

terov, 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.

2.5 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 (2.12) 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 4. 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 (2.10)). 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 G3 of the Supplementary Material, we also discuss how to make our algorithm scalable to large-scale structured matrix recovery problems and verify its effectiveness via a simulation study.

3 Non-asymptotic properties

We first present Theorem 3.1 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 constraint set which is determined based on the concrete settings.

Theorem 3.1. *Suppose $\mathbf{A}_0 \in \mathcal{B}_q(R_q) \cap \mathcal{S}$ and that the regularization parameter λ is chosen such that $\lambda \geq 2\|\nabla Q_n(\mathbf{A}_0)\|_{\mathrm{op}}$. 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

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

If $Q(\widehat{\mathbf{A}}) - Q(\mathbf{A}_0) \geq \kappa\left\|\widehat{\mathbf{A}} - \mathbf{A}_0\right\|_F^2$ for some positive constant κ, then we have

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

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 $2\|\nabla Q_n(\mathbf{A}_0)\|_{\mathrm{op}}$. 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.

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).

3.1 Matrix regression

For the matrix regression model (2.2), $Q_n(\mathbf{A})$ is

$$Q_n(\mathbf{A}) = \{n(n-1)\}^{-1}\sum_{i \neq j}|(y_i - \langle \mathbf{X}_i, \mathbf{A}\rangle) - (y_j - \langle \mathbf{X}_j, \mathbf{A}\rangle)|, \tag{3.1}$$

and we take $\mathcal{S}$ to be $\mathbb{R}^{m_1 \times m_2}$. Denote $\xi_i = \sum_{j=1, j\neq i}^{n} \operatorname{sign}(\varepsilon_i - \varepsilon_j) = 2R_i - (n+1)\,, i = 1, \ldots, n$, where R_i is the rank of ε_i among $\{\varepsilon_1, \ldots, \varepsilon_n\}$. Direct calculation yields

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

We impose the following conditions on observation matrices and random errors.

Assumption 3.1.1. *The random errors ε_i's are i.i.d. with density function $f(\cdot)$. Let $\zeta_{ij} = \varepsilon_i - \varepsilon_j, 1 \leq i \neq j \leq n$. Denote $f^*(\cdot)$ as the probability density function of ζ_{ij}. There exists a positive constant b_1 such that $f^*(0) \geq b_1$ and $|\frac{\partial f^*(t)}{\partial t}| \leq b_2$ for all t.*

Assumption 3.1.2. *The $\{\operatorname{vec}(\mathbf{X}_i)\}_{i=1}^{n}$ are i.i.d. sub-Gaussian vectors with $\left\|\operatorname{vec}(\mathbf{X}_i)\right\|_{\psi_2} \leq \kappa_0 < \infty$ and $\lambda_{\min}(\operatorname{Cov}(\operatorname{vec}(\mathbf{X}_i))) \geq b_3 > 0$.*

Assumption 3.1.3. *Let $\boldsymbol{\Delta} \in \mathbb{R}^{m_1 \times m_2}$. There exists a positive constant b_4 such that*

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

Remark 3.1. *Existing works on low-rank matrix estimation usually impose sub-Gaussian distribution (Negahban and Wainwright, 2011, 2012) or bounded moment condition (Fan et al., 2021b) on random errors which excludes many heavy-tailed distributions and skewed distributions such as Cauchy distribution, log-normal distribution and χ^2 distribution. Assumption 3.1.1 relaxes such requirement to a large degree. Assumption 3.1.2 is standard to studying the error bound for matrix LASSO estimators. Alternatively, we can consider fixed designs and the results of the paper also hold under mild conditions. Assumption 3.1.3 controls the quality of minoration of the loss function by a quadratic function. It also appears in Belloni and Chernozhukov (2011) as part of the "sparse identifiability and nonlinearity" condition. Indeed, if the vectorized covariates* $\operatorname{vec}(\mathbf{X})$ *have a log-concave density, which includes many interesting distributions such as multivariate normal distributions and uniform distribution, then $b_4 \geq 3b_1/\left(2\sqrt{2}b_2 K\right)$ for a universal constant K. This follows from the fact that $\mathbb{E}|\langle \mathbf{X}, \boldsymbol{\Delta}\rangle|^3 \leq K(\mathbb{E}|\langle \mathbf{X}, \boldsymbol{\Delta}\rangle|^2)^{3/2}$ holds for log-concave* $\operatorname{vec}(\mathbf{X})$ *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 unique minimizer of population version loss function $Q(\mathbf{A})$. See Lemma **??** in the Supplementary material. The following theorem gives the estimation error rate of $\widehat{\mathbf{A}}$.

Theorem 3.2. *Suppose that $\mathbf{A}_0 \in \mathcal{B}_q(R_q)$ and Assumptions 3.1.1-3.1.3 hold. The regularization parameter λ is chosen as λ^*. Then there exists universal constant $c > 0$ and $C > 0$ such that when $\frac{\sqrt{2}b_4+1}{b_1 b_3 b_4^2}\kappa_0\sqrt{\frac{m_1+m_2}{n}}R_q^{1/(2-q)} < c$, the estimator $\widehat{\mathbf{A}}$ satisfies*

$$\left\|\widehat{\mathbf{A}} - \mathbf{A}_0\right\|_F^2 \leq C\left(\frac{\kappa_0}{b_1 b_3}\right)^{2-q} R_q\left(\frac{m_1+m_2}{n}\right)^{1-q/2} \tag{3.2}$$

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

Remark 3.2. *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). 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*

$$\left\|\widehat{\mathbf{A}} - \mathbf{A}_0\right\|_F^2 \leq C\left(\frac{\kappa_0\sigma}{b_3}\right)^{2-q} R_q\left(\frac{m_1+m_2}{n}\right)^{1-q/2}, \tag{3.3}$$

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

Remark 3.3. *Comparing our result to the result in Fan et al. (2021b). Our Assumption 3.1.1 cannot be directly compared with the moment condition in Fan et al. (2021b) 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. (2021b) under the Cauchy distribution. Fan et al. (2021b) do not assume the existence of error's density function and include independent but not identical distributions. In Theorem 3.2, as b_1 approaches infinity, our error bound approaches zero. In contrast, Fan et al. (2021b) works with the* *moment condition* $\forall i = 1, \ldots, n, \mathbb{E}\left|y_i\right|^{2k} \leq M < \infty$ *for some* $k > 1$*, and their estimation error bound is*

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

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

Remark 3.4. *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 G2 of the Supplementary Material, 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-free" 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.*

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 3.1. *Suppose that Assumptions 3.1.1-3.1.3 hold and* $\sum_{i=1}^m \left|\theta_{0i}\right|^q \leq R_q$*, 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 universal constant* $c > 0$ *and* $C > 0$ *such that when* $\frac{\sqrt{2}b_4+1}{b_1 b_3 b_4^2}\kappa_0\sqrt{\frac{\log m}{n}}R_q^{1/(2-q)} < c$*, the rank estimator, denoted as* $\widehat{\boldsymbol{\theta}}$*, satisfies*

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_2^2 \leq C \left(\frac{\kappa_0}{b_1 b_3}\right)^{2-q} R_q \left(\frac{\log m}{n}\right)^{1-q/2} \tag{3.4}$$

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 (3.2) in Theorem 3.2. 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 (2.3), $Q_n(\mathbf{A})$ becomes

$$\begin{aligned} Q_n\left(\mathbf{A}\right) &= \frac{1}{n\left(n-1\right)} \sum_{i \neq j} \sum_{k=1}^{m_1} \left| \left\{ y_{ik} - \mathbf{e}_k^\top \left(m_1\right) \mathfrak{X}\left(\mathbf{X}_i; \mathbf{A}\right) \right\} - \left\{ y_{jk} - \mathbf{e}_k^\top \left(m_1\right) \mathfrak{X}\left(\mathbf{X}_j; \mathbf{A}\right) \right\} \right| \\ &= \frac{1}{n\left(n-1\right)} \sum_{i \neq j} \sum_{k=1}^{m_1} \left| \left(y_{ik} - \mathbf{x}_i^\top \mathbf{A}_k \right) - \left(y_{jk} - \mathbf{x}_j^\top \mathbf{A}_k \right) \right|, \end{aligned} \tag{3.5}$$

where $\mathbf{A}_k^\top$ is the k-th row of $\mathbf{A}$. Here we take $\mathcal{S}$ to be $\mathbb{R}^{m_1 \times m_2}$. To specify a concrete form of S_n, we define two matrices $\mathbf{X} = \left(\mathbf{x}_1, \dots, \mathbf{x}_n\right)^\top \in \mathbb{R}^{n \times m_2}$ and $\boldsymbol{\xi} = \left(\boldsymbol{\xi}_1, , \dots, \boldsymbol{\xi}_{m_1}\right) \in \mathbb{R}^{n \times m_1}$, where the j-th element of $\boldsymbol{\xi}_k$ is $\sum_{i=1, i \neq j}^{n} \operatorname{sign}\left(\varepsilon_{jk} - \varepsilon_{ik}\right) = 2R_{jk} - (n+1)$, R_{jk} is the rank of ε_{jk} among $\{\varepsilon_{1k}, \dots, \varepsilon_{nk}\}$. Direct calculation yields

$$S_n = \left\| \nabla Q_n\left(\mathbf{A}_0\right) \right\|_{\mathrm{op}} = \frac{2}{n\left(n-1\right)} \left\| \boldsymbol{\xi}^\top \mathbf{X} \right\|_{\mathrm{op}}.$$

We need the following conditions.

Assumption 3.2.1. *The random errors $\boldsymbol{\varepsilon}_i$'s are i.i.d. with marginal density function $f_k\left(\cdot\right)$ for ε_{ik}. Let $\zeta_{ijk} = \varepsilon_{ik} - \varepsilon_{jk}, 1 \leq i \neq j \leq n, 1 \leq k \leq m_1$. Let $f_k^*\left(\cdot\right)$ denote the probability density function of ζ_{ijk}. There exists a positive constant b_1 such that $f_k^*\left(0\right) \geq b_1$ and $|\frac{\partial f_k^*(t)}{\partial t}| \leq b_2$ for all t, uniformly in k.*

Assumption 3.2.2. *The covariates $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. sub-Gaussian vectors with $\left\|\mathbf{x}_i\right\|_{\psi_2} \leq \kappa_0 < \infty$ and $\lambda_{\min}\left(\mathrm{Cov}\left(\mathbf{x}_i\right)\right) \geq b_3 > 0$.*

Assumption 3.2.3. *Let $\boldsymbol{\Delta} \in \mathbb{R}^{m_1 \times m_2}$. There exists a positive constant b_4 such that*

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

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

Theorem 3.3. *Suppose $\mathbf{A}_0 \in \mathcal{B}_q\left(R_q\right)$ and Assumptions 3.2.1-3.2.3 hold. The regularization parameter λ is chosen such that $\lambda = \lambda^*$. Then there exists universal constant $c > 0$ and $C > 0$ such that when $\frac{\sqrt{2}b_4+1}{b_1 b_3 b_4^2} \kappa_0 \sqrt{\frac{m_1+m_2}{n}} R_q^{1/(2-q)} < c$, the estimator $\widehat{\mathbf{A}}$ satisfies*

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

with probability at least $1-\alpha_0-3\exp\left\{-\left(m_1+m_2\right)\right\}.$

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 2.3), the trace regression formulation (2.2) 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: $\left\langle \mathbf{X}_i\,,\, c\mathbb{1}\right\rangle = c$, for any $c\in\mathbb{R}$, where $\mathbb{1}=\mathbf{1}\mathbf{1}^\top$. Consequently, we can always reformulate (2.2) by offsetting the ground truth and introducing an intercept term

$$y_i = c + \left\langle \mathbf{X}_i\,,\, \mathbf{A}_0 - c\mathbb{1}\right\rangle + \varepsilon_i.$$

Hence, the failure for the estimation of c will cause the problem of identifiability between $\mathbf{A}_0$ and $\mathbf{A}_0 - c\mathbb{1}$.

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

$$a_i y_i = \left\langle a_i\mathbf{X}_i\,,\, \mathbf{A}_0\right\rangle + a_i\varepsilon_i, \tag{3.6}$$

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 $\left\langle a_i\mathbf{X}_i\,,\, \mathbf{C}\right\rangle = c$. 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 (3.6). 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

$$Q_n\left(\mathbf{A}\right) = \left\{n\left(n-1\right)\right\}^{-1}\sum_{i\neq j}\left|\left(a_i y_i - \langle a_i\mathbf{X}_i, \mathbf{A}\rangle\right) - \left(a_j y_j - \langle a_j\mathbf{X}_j, \mathbf{A}\rangle\right)\right|. \tag{3.7}$$

We show that in Appendix E of the Supplementary material, (3.6) can guarantee that $\mathbf{A}_0$ is the unique minimizer of $Q(\mathbf{A})$ under appropriate condition.

For theoretical analysis, assume that $\|\mathbf{A}_0\|_\infty \leq \eta$, which is reasonable since a bound on the maximum of the elements is often known in application. Accordingly, $\mathcal{S} = \{\mathbf{A} : \|\mathbf{A}\|_\infty \leq \eta\}$. 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 $\pi_{jk} = \mathbb{P}\left(\mathbf{X} = \mathbf{e}_j\left(m_1\right)\mathbf{e}_k^\top\left(m_2\right)\right)$ 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 $\max_{i,j}\left(C_i, R_j\right) \geq 1/\min\left(m_1, m_2\right) = 1/m_2$, 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 3.3.1. *There exists a positive constant $L \geq 1$ such that* $\max_{i,j}\left(C_i, R_j\right) \leq L/m_2$.

Assumption 3.3.2. *There exists a positive constant $\mu \geq 1$ such that $\pi_{jk} \geq \left(\mu m_1 m_2\right)^{-1}$.*

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

Assumption 3.3.3. *The random errors ε_i's are i.i.d with density function $f(\cdot)$. Let $\zeta_{ij}^- = \varepsilon_i - \varepsilon_j$, $\zeta_{ij}^+ = \varepsilon_i + \varepsilon_j, 1 \leq i \neq j \leq n$. 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 $f^-(t) \geq \frac{1}{2c_1^2}$ and $f^+(t) \geq \frac{1}{2c_1^2}$ for all $|t| \leq 4\eta$.*

Write

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

where $\xi_i = 2R_i - \left(n+1\right), i = 1, \ldots, n$ 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

$$\lambda^* = 2G_{\mathbf{S}_n}^{-1}\left(1-\alpha_0\right),$$

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

Theorem 3.4. *Suppose that Assumptions 3.3.1-3.3.3 hold. Consider the regularization parameter $\lambda = \lambda^*$. Then there exist a numerical constant C such that*

$$\left\|\widehat{\mathbf{A}} - \mathbf{A}_0\right\|_F^2 \leq C \max\left\{R_q \left(c_1^2 \mu m_1 m_2 \sqrt{\frac{L \log(m_1+m_2)}{n m_2}}\right)^{2-q}, \ c_1^2 \eta \mu m_1 m_2 \sqrt{\frac{\log(m_1+m_2)}{n}}\right\} \tag{3.8}$$

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

Remark 3.5. *When the random errors ε_i's are symmetric, $a\mathbf{X}$ is independent of $a\varepsilon$ (see Lemma* **??** *in the Supplementary material), which implies that $\left\|\nabla Q_n(\mathbf{A}_0)\right\|_{\mathrm{op}}$ has the same distribution as $\mathbf{S}_n$ 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 $\left\|\nabla Q_n(\mathbf{A}_0)\right\|_{\mathrm{op}}$ 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 3.6. *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*

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

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 (3.8)*, 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 Negahban and Wainwright (2012).*

4 Simulation

In this section we investigate the performance of our proposed Rank Matrix LASSO estimator through numerical studies based on both synthetic data and a real data. The results presented in this section are evaluated through 100 Monte Carlo replications. 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. The Matlab codes that implement the proposed scheme are available in the Supplementary Material.

4.1 Matrix regression

We firstly study the matrix regression model (2.2) and compare different estimators under various noises. We mainly 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. 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.

In Figure 1, for each setting, we compare four nuclear norm penalization estimators: Matrix LASSO (Negahban and Wainwright, 2011), Robustified Matrix LASSO (Fan et al., 2021b), 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 (2.8). In practice, (2.8) 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 (2.8) 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. (2021b). In the figures, 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 $\left\|\widehat{\mathbf{A}}-\mathbf{A}_0\right\|_F$ 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.

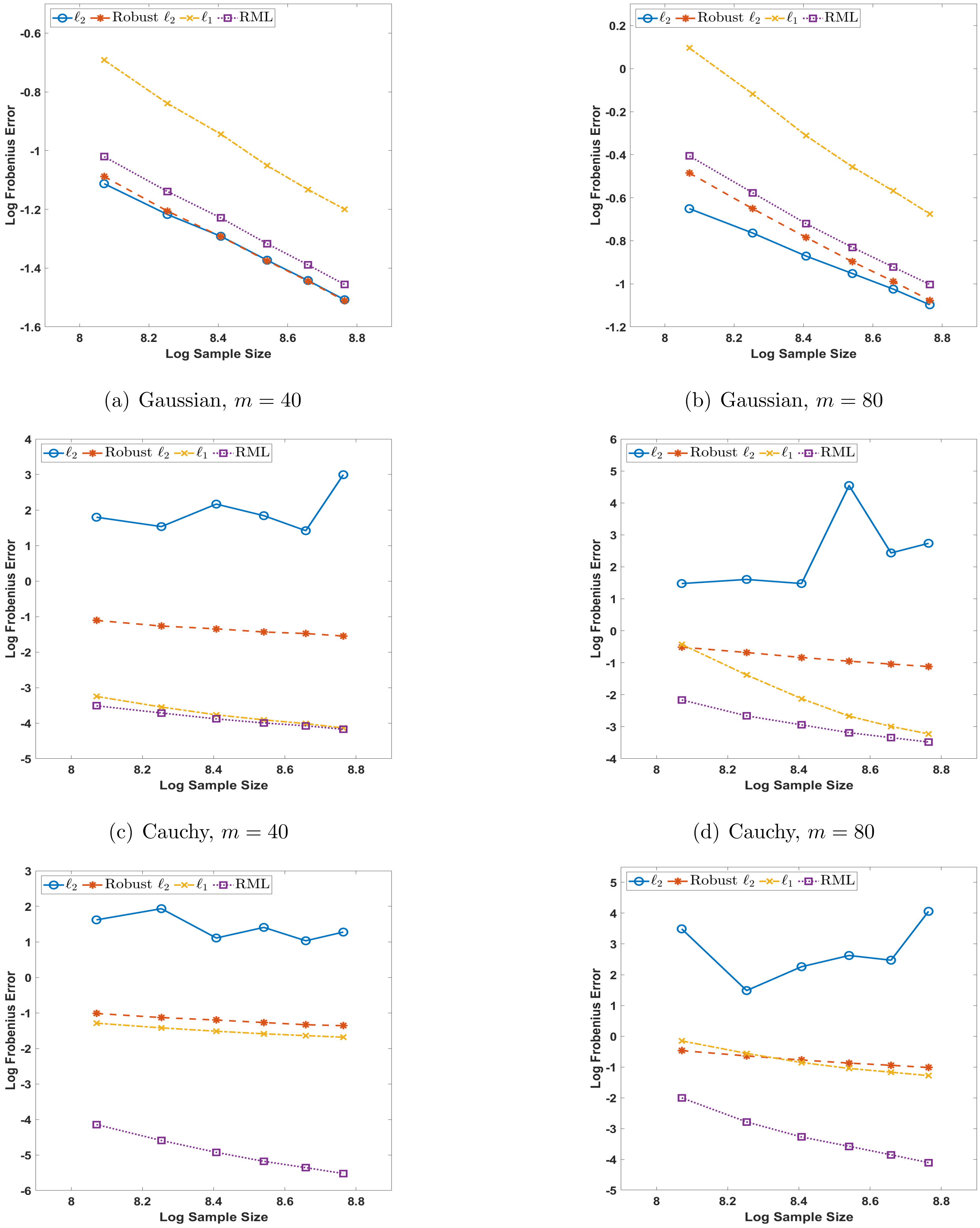
In Figure 2, under the same setting as Figure 1, we compare our method with the state-of-art alternate projection method called Scaled Gradient Descent (SGD) in Tong et al. (2021a) and Scaled Subgradient Methods (SSM) in Tong et al. (2021b). The SGD and SSM consider the ℓ_2 loss and ℓ_1 loss under matrix 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. 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 be best or is close to the best. In fact, the best performances of SGD and SSM are not easily to be attained in practice because we do not know the true rank, which is also a manifestation of the usefulness of our method.

4.2 Multivariate regression

In this subsection we consider the multivariate regression model (2.3). 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 Section 4.1, but the setting considered in this section 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:

- Identity covariance. $\mathbf{\Sigma} = \mathbf{I}_m$, which gives i.i.d. normal components for the random vectors. In this case our Assumption 3.2.2 is met with $\kappa_0 = 1$ and $b_3 = 1$.

- 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



(a) Gaussian, $m = 40$

(b) Gaussian, $m = 80$

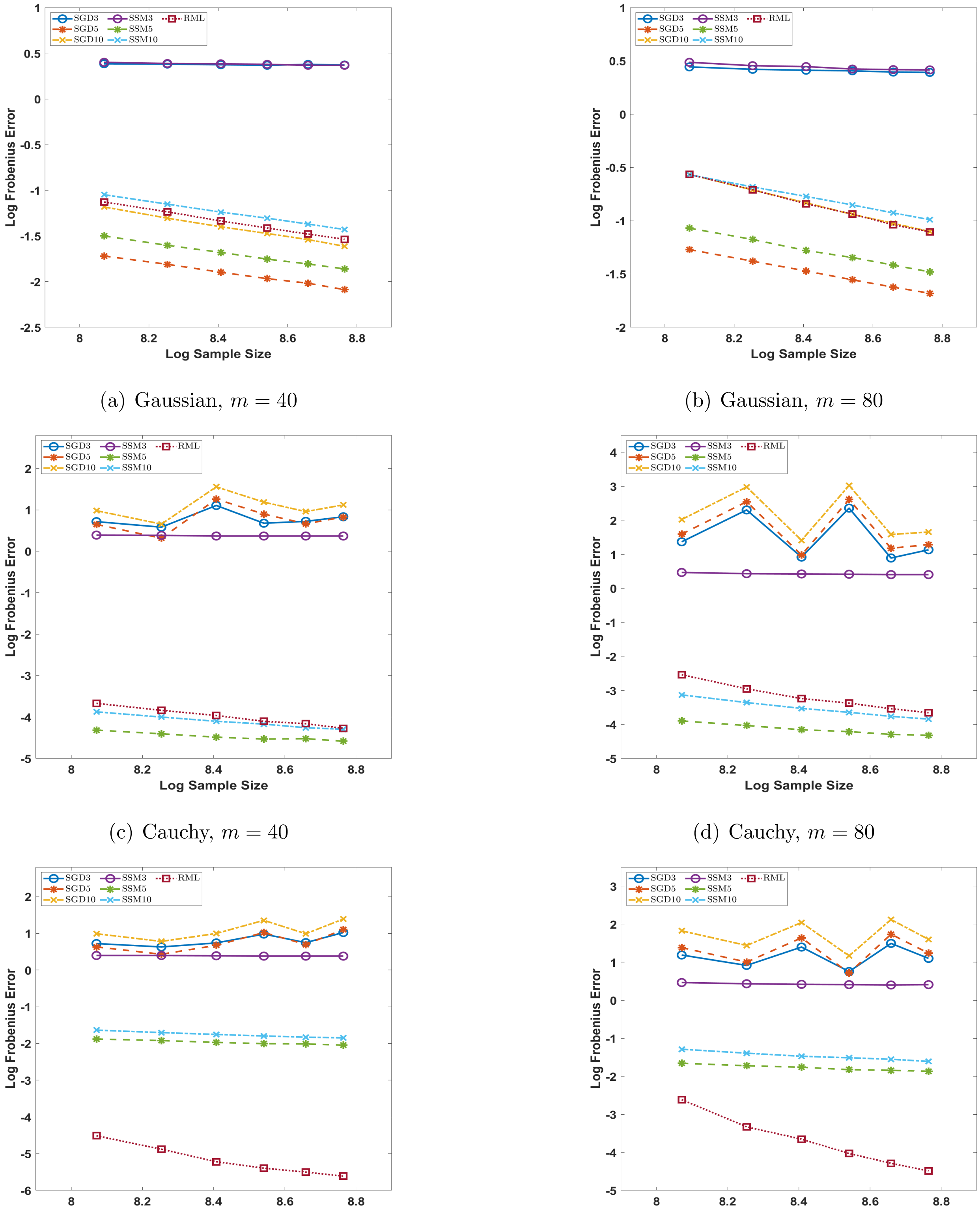
(c) Cauchy, $m = 40$

(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

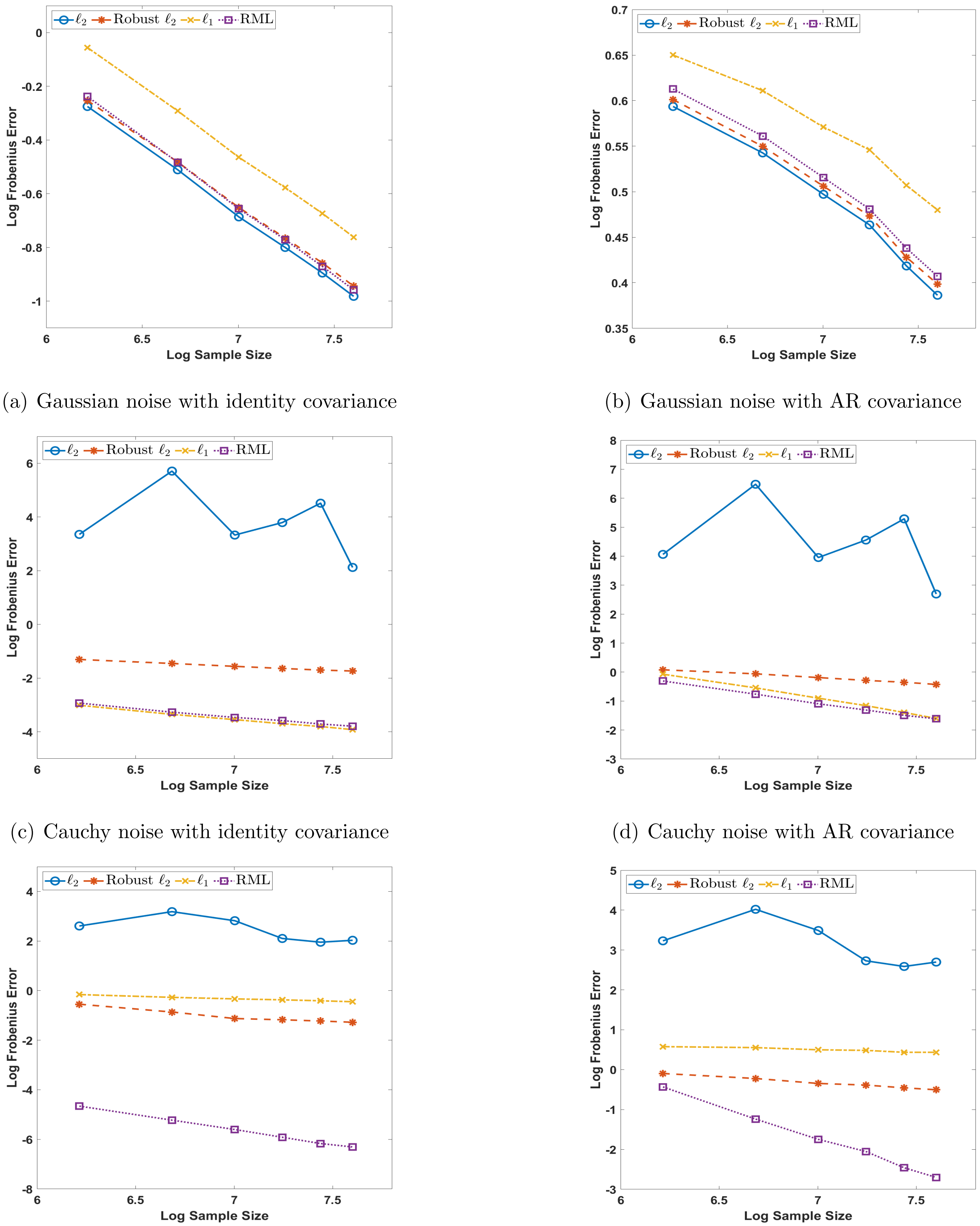
Szegö (1958), in this case Assumption 3.2.2 is satisfied with $\kappa_0 = 1/9$ and $b_3 = 9$.

We compare our method with other nuclear norm penalization estimators and numerical results are presented in Figure 3(a)-3(f). The proposed method shows a quite competitive performance within the four candidates and yields sharp accuracy in estimating the ground truth under both designs.

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 Section 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.

4.3 Matrix completion

For the last section, we study the matrix completion model in Example 2.3. $\mathbf{A}_0$ 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 Section 4.1, and the sample size ranges from 3200 to 6400. The results are presented in Figure 5. In particular, based on our comments in Section 3.3, we apply the random symmetrization here for our methodology.



(a) Gaussian noise with identity covariance

(b) Gaussian noise with AR covariance

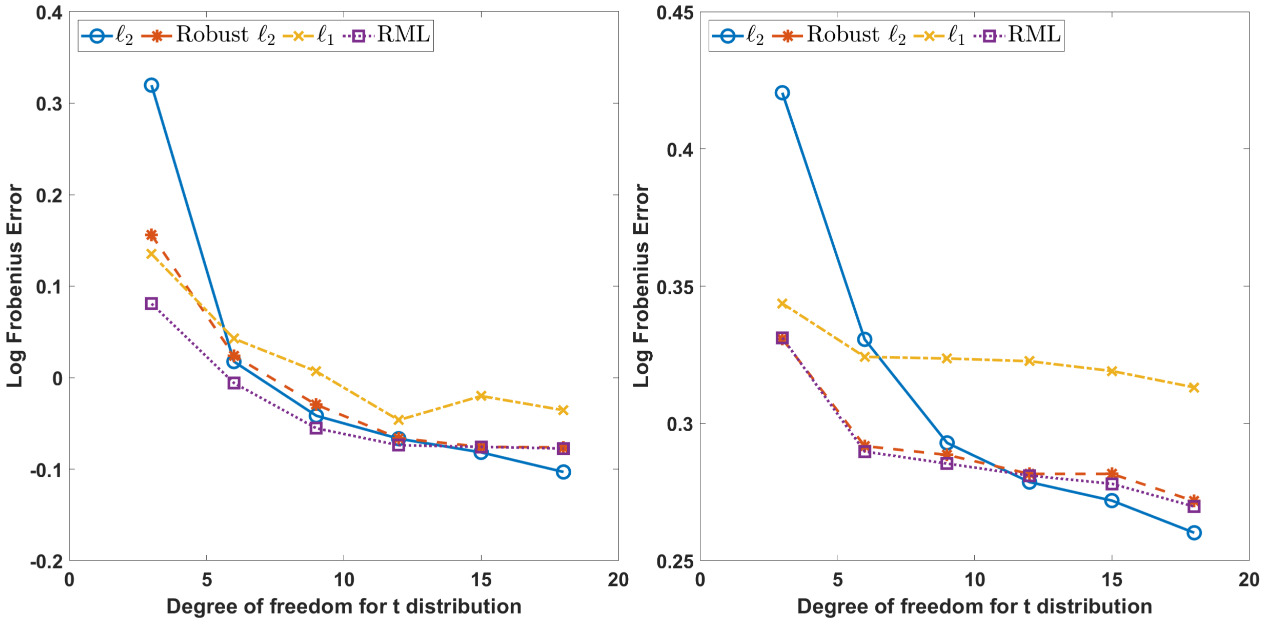
(c) Cauchy noise with identity covariance

(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): $m_1 = 40$, $m_2 = 80$, $n = 60$

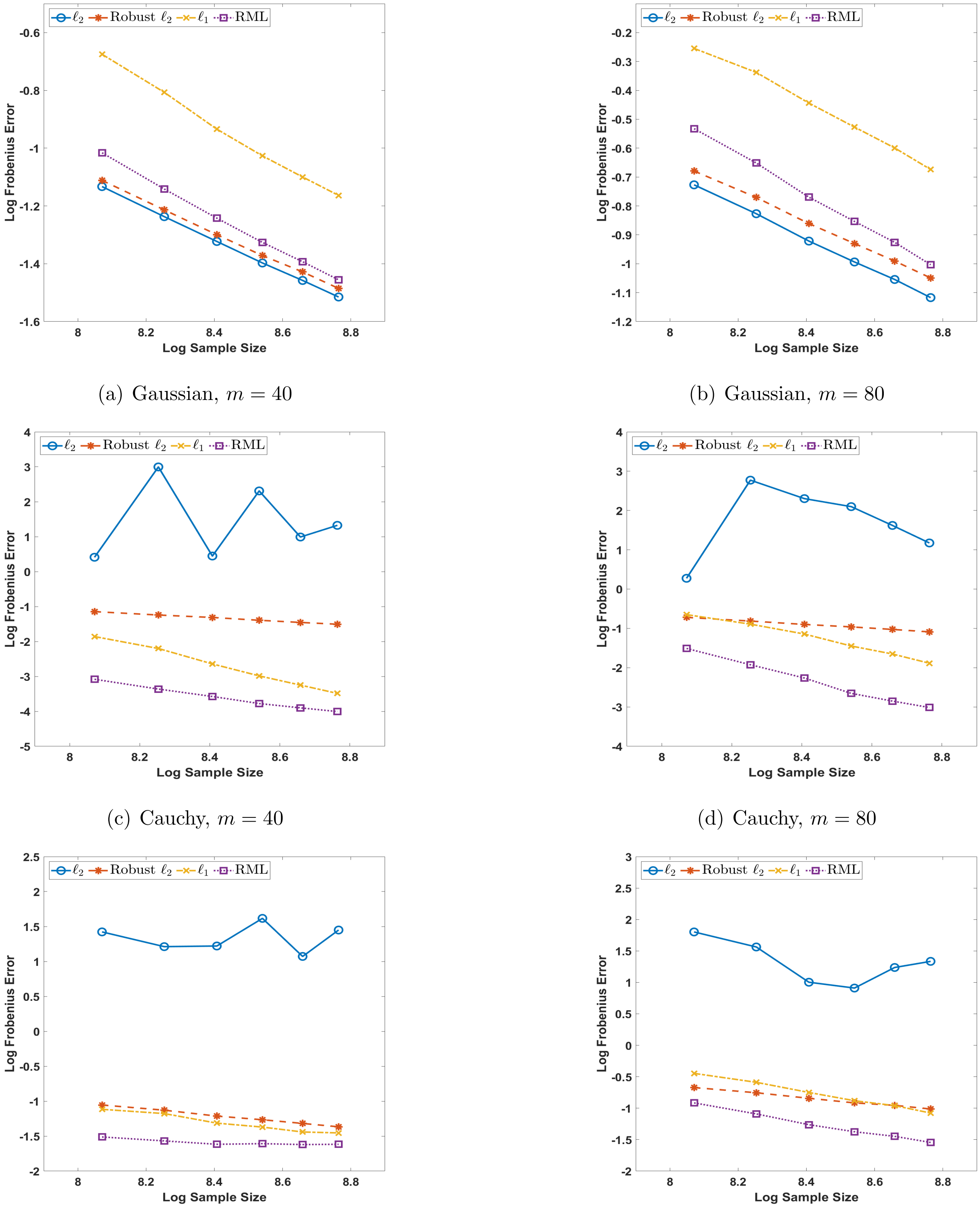
Figure 4: Log Frobenius Errors of different estimators for multivariate regression model under t noises with varying degrees of freedom

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 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.

Real-life application typically involves computation of large matrices(m_1, m_2 as high as 10^4), discussion on how to adapt our algorithm to large-scale problems and the corresponding numerical results is presented in Appendix G3 of the Supplementary Material.

4.4 Computation time

The computational merit of our Rank Matrix LASSO is that it is tuning-free 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 Section 4.1 when $m_1 = m_2 = 40$ and $n = 3200$, (II)



(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

multivariate regression model with the same settings as Section 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 Section 4.3 when $m_1 = m_2 = 40$ and $n = 3200$. To save the space, we only consider Fan et al. (2021b)'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 $\left\|\widehat{\mathbf{A}} - \mathbf{A}_0\right\|_F^2$, 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.

4.5 Real data analysis: Arabidopsis thaliana genetic data

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 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 4.2, trained over the first 80% of the data, $\mathbf{Y}_{train}$, and report 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,

$$\text{MAD} = \frac{1}{m_1 m_2}\|\mathbf{Y}_{pre} - \mathbf{Y}_{test}\|_{1,1}, \ \text{MSE} = \frac{1}{m_1 m_2}\|\mathbf{Y}_{pre} - \mathbf{Y}_{test}\|_F^2.$$

Here $\|\cdot\|_{1,1}$ simply gives the summation of the absolute values of all the entries for a given matrix. For Matrix LASSO, Robustified Matrix LASSO and regularized LAD, we determine the tuning parameter using robust cross validation (Fan et al., 2021b), and for our RML we apply the pivotal tuning procedure. The result is summarized in Table 2.

Four methods generate comparable testing errors. Table 2 implies a trade-off between

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

Model	Error	Estimator	$\left\|\widehat{\mathbf{A}} - \mathbf{A}_0\right\|_F^2$	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

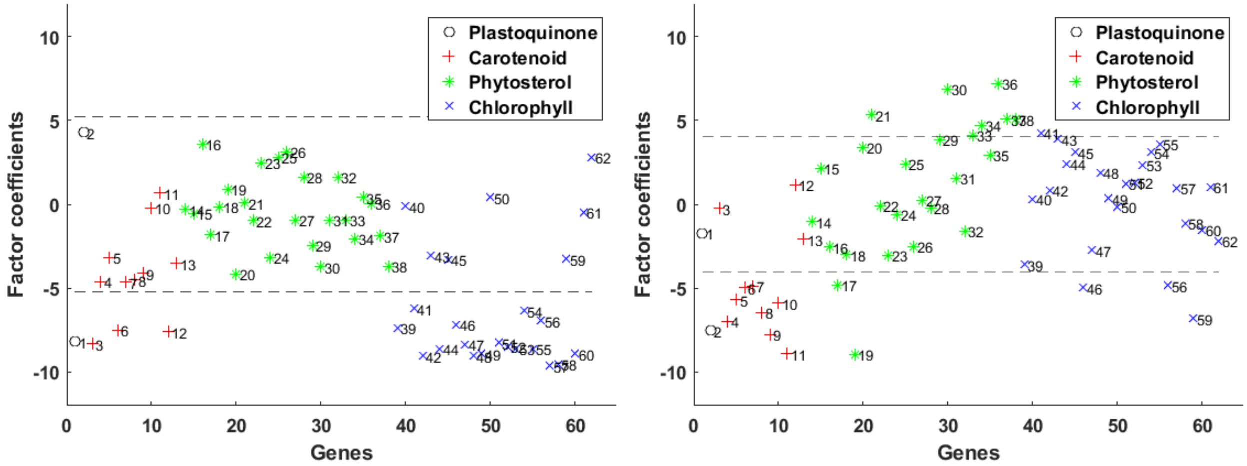
Table 2: Prediction accuracy for the Arabidopsis thaliana generic data

Method	ℓ_2	Robust-ℓ_2	ℓ_1	RML
MSE	0.521	0.617	0.531	0.576
MAD	0.515	0.578	0.526	0.540
Estimated rank	14	7	10	5

the prediction accuracy and the model size. Matrix LASSO (ℓ_2) and regularized LAD (ℓ_1) achieved a low testing error, but end up with a less parsimonious model, leading to a sacrifice in interpretability. Robustified Matrix LASSO and our Rank Matrix LASSO, on the other hand, produced estimators with smaller rank, which is beneficial for some follow-up analysis such as principal component analysis or exploratory factor analysis (EFA).

Next, we perform an EFA based on the Rank Matrix LASSO 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.



(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

5 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.

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. (2021b) 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/strucute information.

Supplementary Material

Technical proofs of all Theorems and Corollaries and additional simulations results are included to the Supplementary Material.

Acknowledgments

The authors thank the Editor, Associate Editor and anonymous referees for their many helpful comments that have resulted in significant improvements in the article.

References

Baglama, J. and Reichel, L. (2005), "Augmented implicitly restarted Lanczos bidiagonalization methods," *SIAM Journal on Scientific Computing*, 27, 19–42.

Becker, S., Fadili, J., and Ochs, P. (2019), "On quasi-Newton forward-backward splitting: Proximal calculus and convergence," *SIAM Journal on Optimization*, 29, 2445–2481.

Belloni, A. and Chernozhukov, V. (2011), "ℓ_1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, 39, 82–130.

Belloni, A., Chernozhukov, V., and Wang, L. (2011), "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, 98, 791–806.

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

Bian, W. and Wu, F. (2021), "Accelerated forward-backward method with fast convergence rate for nonsmooth convex optimization beyond differentiability," *arXiv preprint arXiv:2110.01454*.

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

Bunea, F., She, Y., and Wegkamp, M. H. (2011), "Optimal selection of reduced rank estimators of high-dimensional matrices," *The Annals of Statistics*, 39, 1282–1309.

— (2012), "Joint variable and rank selection for parsimonious estimation of high-dimensional matrices," *The Annals of Statistics*, 40, 2359–2388.

Candès, E. J. and Recht, B. (2009), "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, 9, 717–772.

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

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020), "Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization," *SIAM Journal on Optimization*, 30, 3098–3121.

Davenport, M. A., Plan, Y., Van Den Berg, E., and Wootters, M. (2014), "1-bit matrix completion," *Information and Inference: A Journal of the IMA*, 3, 189–223.

Elsener, A. and van de Geer, S. (2018), "Robust low-rank matrix estimation," *The Annals of Statistics*, 46, 3481–3509.

Fan, J., Li, Q., and Wang, Y. (2017), "Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions," *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79, 247.

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

Fan, J., Liu, H., Sun, Q., and Zhang, T. (2018), "I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error," *The Annals of Statistics*, 46, 814.

Fan, J., Wang, K., Zhong, Y., and Zhu, Z. (2021a), "Robust high-dimensional factor models with applications to statistical machine learning," *Statistical Science*, 36, 303–327.

Fan, J., Wang, W., and Zhu, Z. (2021b), "A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery," *The Annals of statistics*, 49, 1239.

Fazel, M., Candes, E., Recht, B., and Parrilo, P. (2008), "Compressed sensing and robust recovery of low rank matrices," in *2008 42nd Asilomar Conference on Signals, Systems and Computers*, IEEE, pp. 1043–1047.

Golbabaee, M. and Vandergheynst, P. (2012), "Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2741–2744.

Grenander, U. and Szegö, G. (1958), *Toeplitz forms and their applications*, Berkeley and Los Angeles: University of California Press.

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

Hettmansperger, T. and McKean, J. (1998), *Robust Nonparametric Statistical Methods*, London: Arnold.

Jaeckel, L. A. (1972), "Estimating regression coefficients by minimizing the dispersion of the residuals," *The Annals of Mathematical Statistics*, 1449–1458.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013), "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674.

Ji, S. and Ye, J. (2009), "An accelerated gradient method for trace norm minimization," in *Proceedings of the 26th annual international conference on machine learning*, pp. 457–464.

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

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011), "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, 39, 2302–2329.

Larsen, R. M. (2004), "Propack-software for large and sparse svd calculations," Available at `http://sun.stanford.edu/rmunk/PROPACK`.

Law, M., Ritov, Y., Zhang, R., and Zhu, Z. (2021), "Rank-Constrained Least-Squares: Prediction and Inference," *arXiv preprint arXiv:2111.14287*.

Lee, J. D., Sun, Y., and Saunders, M. A. (2014), "Proximal Newton-type methods for minimizing composite functions," *SIAM Journal on Optimization*, 24, 1420–1443.

Lovász, L. and Vempala, S. (2007), "The geometry of logconcave functions and sampling algorithms," *Random Structures & Algorithms*, 30, 307–358.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2018), "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion," in *International Conference on Machine Learning*, PMLR, pp. 3345–3354.

Ma, J. and Fattahi, S. (2021), "Implicit Regularization of Sub-Gradient Method in Robust Matrix Recovery: Don't be Afraid of Outliers," *arXiv preprint arXiv:2102.02969*.

Negahban, S. and Wainwright, M. J. (2011), "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, 1069–1097.

— (2012), "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The Journal of Machine Learning Research*, 13, 1665–1697.

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

— (2013), "Gradient methods for minimizing composite functions," *Mathematical Programming*, 140, 125–161.

Nguyen, L. T., Kim, J., Kim, S., and Shim, B. (2019), "Localization of IoT networks via low-rank matrix completion," *IEEE Transactions on Communications*, 67, 5833–5847.

Parzen, M. I., Wei, L.-J., and Ying, Z. (1994), "A resampling method based on pivotal estimating functions," *Biometrika*, 81, 341–350.

Ramlatchan, A., Yang, M., Liu, Q., Li, M., Wang, J., and Li, Y. (2018), "A survey of matrix completion methods for recommendation systems," *Big Data Mining and Analytics*, 1, 308–323.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011), "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Transactions on Information Theory*, 57, 6976–6994.

Recht, B., Fazel, M., and Parrilo, P. A. (2010), "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, 52, 471–501.

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

Saumard, A. and Wellner, J. A. (2014), "Log-concavity and strong log-concavity: a review," *Statistics Surveys*, 8, 45.

She, Y. (2017), "Selective factor extraction in high dimensions," *Biometrika*, 104, 97–110.

She, Y. and Chen, K. (2017), "Robust reduced-rank regression," *Biometrika*, 104, 633–647.

She, Y. and Tran, H. (2019), "On cross-validation for sparse reduced rank regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 145–161.

Sun, Q., Zhou, W.-X., and Fan, J. (2020), "Adaptive huber regression," *Journal of the American Statistical Association*, 115, 254–265.

Sun, R. and Luo, Z.-Q. (2016), "Guaranteed matrix completion via non-convex factorization," *IEEE Transactions on Information Theory*, 62, 6535–6579.

Tan, K. M., Sun, Q., and Witten, D. (2018), "Robust sparse reduced rank regression in high dimensions," *arXiv preprint arXiv:1810.07913*.

Toh, K.-C. and Yun, S. (2010), "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, 6, 15.

Tong, T., Ma, C., and Chi, Y. (2021a), "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent," *Journal of Machine Learning Research*, 22, 1–63.

— (2021b), "Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number," *IEEE Transactions on Signal Processing*, 69, 2396–2409.

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

Wang, L. and Li, R. (2009), "Weighted Wilcoxon-type smoothly clipped absolute deviation method," *Biometrics*, 65, 564–571.

Wang, L., Peng, B., Bradic, J., Li, R., and Wu, Y. (2020), "A tuning-free robust and efficient approach to high-dimensional regression," *Journal of the American Statistical Association*, 115, 1–44.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., von Rohr, P., Thiele, L., et al. (2004), "Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana," *Genome Biology*, 5, 1–13.

Yu, J., Vishwanathan, S., Günter, S., and Schraudolph, N. N. (2010), "A quasi-Newton approach to nonsmooth convex optimization problems in machine learning," *The Journal of Machine Learning Research*, 11, 1145–1200.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007), "Dimension reduction and coefficient estimation in multivariate linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 329–346.

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

Zheng, J., Qin, M., Yu, H., and Wang, W. (2018), "An efficient truncated nuclear norm constrained matrix completion for image inpainting," in *Proceedings of Computer Graphics International 2018*, pp. 97–106.