The Annals of Statistics
2021, Vol. 49, No. 3, 1239–1266
https://doi.org/10.1214/20-AOS1980

A SHRINKAGE PRINCIPLE FOR HEAVY-TAILED DATA: HIGH-DIMENSIONAL ROBUST LOW-RANK MATRIX RECOVERY

BY JIANQING FAN[1,*], WEICHEN WANG[1,†] AND ZIWEI ZHU[2]

[1]*Department of Operations Research and Financial Engineering, Princeton University,* [*]*jqfan@princeton.edu;* [†]*nickweichwang@gmail.com*

[2]*Department of Statistics, University of Michigan, Ann Arbor, ziweiz@umich.edu*

This paper introduces a simple principle for robust statistical inference via appropriate shrinkage on the data. This widens the scope of high-dimensional techniques, reducing the distributional conditions from subexponential or sub-Gaussian to more relaxed bounded second or fourth moment. As an illustration of this principle, we focus on robust estimation of the low-rank matrix $\mathbf{\Theta}^*$ from the trace regression model $Y = \mathrm{Tr}(\mathbf{\Theta}^{*\top}\mathbf{X}) + \varepsilon$. It encompasses four popular problems: sparse linear model, compressed sensing, matrix completion and multitask learning. We propose to apply the penalized least-squares approach to the appropriately truncated or shrunk data. Under only bounded $2 + \delta$ moment condition on the response, the proposed robust methodology yields an estimator that possesses the same statistical error rates as previous literature with sub-Gaussian errors. For sparse linear model and multitask regression, we further allow the design to have only bounded fourth moment and obtain the same statistical rates. As a byproduct, we give a robust covariance estimator with concentration inequality and optimal rate of convergence in terms of the spectral norm, when the samples only bear bounded fourth moment. This result is of its own interest and importance. We reveal that under high dimensions, the sample covariance matrix is not optimal whereas our proposed robust covariance can achieve optimality. Extensive simulations are carried out to support the theories.

1. Introduction. Heavy-tailed distribution is ubiquitous in modern statistical analysis and machine learning problems. It is a stylized feature of high-dimensional data, which may be caused by chance of extreme events or by the complex data generating process. It has been widely known that financial returns and macroeconomic variables exhibit heavy tails for rare events, and large-scale imaging datasets in biological studies are corrupted by heavy-tailed noises due to limited measurement precision. Figure 1 provides some empirical evidence on this which is pandemic to high-dimensional data. These phenomena contradict the popular assumption of sub-Gaussian or subexponential noises in the theoretical analysis of standard statistical procedures. They also have adverse impact on the popularly used methods. Simple and effective principles are needed for dealing with heavy tailed data.

Recent years have witnessed increasing literature on the robust mean estimation when the population distribution is heavy-tailed. Catoni (2012) proposed a novel approach through minimizing a robust empirical loss, and more so after our initial submission of the paper in 2016. Unlike the traditional ℓ_2 loss, the robust loss function therein penalizes large deviations, thereby making the correspondent M-estimator insensitive to extreme values. It turns out that when the population has only finite second moment, the estimator has exponential concentration around the true mean and enjoys the same rate of statistical consistency as the sample average for sub-Gaussian distributions. Brownlees, Joly and Lugosi (2015) pursued

Received May 2019; revised March 2020.
MSC2020 subject classifications. Primary 62F35; secondary 62J05.
Key words and phrases. Robust statistics, shrinkage, heavy-tailed data, trace regression, low-rank matrix recovery, high-dimensional statistics.

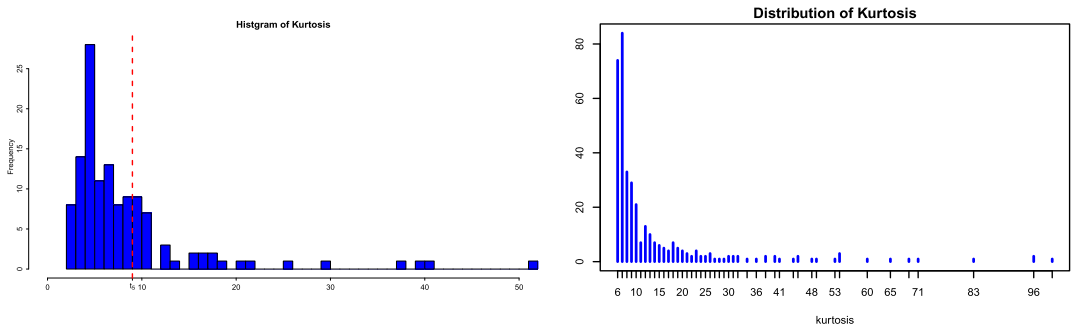


FIG. 1. Distributions of kurtosis of macroeconomic variables and gene expressions. *Red dashline marks variables with empirical kurtosis equals to that of t_5-distribution. Left panel*: *For* 131 *macroeconomics variables in Stock and Watson* (2002). *Right panel*: *For logarithm of expression profiles of* 383 *genes based on RNA-seq for autism data* (*Gupta et al.* (2014)), *whose kurtosis is bigger than that of t_5 among* 19,122 *genes.*

the Catoni's mean estimator further by applying it to empirical risk minimization. Fan, Li and Wang (2017) utilized the Huber loss with diverging threshold, called robust approximation to quadratic (RA-quadratic), in a sparse regression problem and showed that the derived M-estimator can also achieve the minimax statistical error rate. Loh (2017) studied the statistical consistency and asymptotic normality of a general robust M-estimator and provided a set of sufficient conditions to achieve the minimax rate in the high-dimensional regression problem.

Another effective approach to handle heavy-tailed distribution is the so-called "median of means" approach, which can be traced back to Nemirovsky and Yudin (1982). The main idea is to first divide the whole samples into several parts and take the median of the means from all pieces of subsamples as the final estimator. This "median of means" estimator also enjoys exponential large deviation bound around the true mean. Hsu and Sabato (2016) and Minsker (2015) generalized this idea to multivariate cases and applied it to robust PCA, high-dimensional sparse regression and matrix regression, achieving minimax optimal rates up to logarithmic factors.

In this paper, we propose a simple and effective principle: truncation of univariate data and more generally shrinkage of multivariate data to achieve the robustness. We will illustrate our ideas through a general model: the trace regression

$$Y = \operatorname{Tr}(\boldsymbol{\Theta}^{*\top}\mathbf{X}) + \varepsilon =: \langle \boldsymbol{\Theta}^*, \mathbf{X} \rangle + \varepsilon,$$

where for any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$, $\langle \mathbf{A}, \mathbf{B} \rangle := \operatorname{Tr}(\mathbf{A}^\top \mathbf{B})$. This model embraces linear regression, matrix or vector compressed sensing, matrix completion and multitask regression as specific examples. The goal is to estimate the coefficient matrix $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$, which is assumed to have a nearly low-rank structure in the sense that its Schatten norm is constrained: $\sum_{i=1}^{\min(d_1,d_2)} \sigma_i(\boldsymbol{\Theta}^*)^q \le \rho$ for $0 \le q < 1$, where $\sigma_i(\boldsymbol{\Theta}^*)$ is the ith singular value of $\boldsymbol{\Theta}^*$, that is, the square-root of the ith eigenvalue of $\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*$. In other words, the singular values of $\boldsymbol{\Theta}^*$ decay fast enough so that $\boldsymbol{\Theta}^*$ can be well approximated by a low-rank matrix. We always consider the high-dimensional setting where the sample size $n \ll d_1 d_2$. As we shall see, appropriate data shrinkage allows us to recover $\boldsymbol{\Theta}^*$ with only bounded second and fourth moment conditions on noise and design, respectively.

As the most simple and important example of the low-rank trace regression, sparse linear regression and compressed sensing have become a hot topic in statistics research in the past two decades. See, for example, Tibshirani (1996), Chen, Donoho and Saunders (2001), Fan and Li (2001), Donoho (2006), Candes and Tao (2006, 2007), Candes (2008), Figueiredo, Nowak and Wright (2007), Fan and Lv (2008), Zou and Li (2008), Bickel, Ritov and Tsybakov (2009), Zhang (2010), Negahban and Wainwright (2012), Donoho, Johnstone and

Montanari (2013). These pioneering papers explore the sparsity to achieve accurate signal recovery in high dimensions.

Recently significant progresses have been made on low-rank matrix recovery under high-dimensional settings. One of the most well-studied approaches is the penalized least-squares method. Negahban and Wainwright (2011) analyzed the nuclear norm penalization in estimating nearly low-rank matrices under the trace regression model. Specifically, they derived nonasymptotic estimation error bounds in terms of the Frobenius norm when the noise is sub-Gaussian. Rohde and Tsybakov (2011) proposed to use a Schatten-p quasi-norm penalty where $p \leq 1$, and they derived nonasymptotic bounds on the prediction risk and Schatten-q risk of the estimator, where $q \in [p, 2]$. Another effective method is through nuclear norm minimization under affine fitting constraint. Other important contributions include Recht, Fazel and Parrilo (2010), Candes and Plan (2011), Cai and Zhang (2014, 2015), etc. When the true low-rank matrix $\mathbf{\Theta}^*$ satisfies certain restricted isometry property (RIP) or similar properties, this approach can exactly recover $\mathbf{\Theta}^*$ under the noiseless setting and enjoy sharp statistical error rate with sub-Gaussian and subexponential noise.

There has also been great amount of work on matrix completion. Candes and Recht (2009) considered matrix completion under noiseless settings and gave conditions under which exact recovery is possible. Candes and Plan (2010) proposed to fill in the missing entries of the matrix by nuclear-norm minimization subject to data constraints, and showed that $rd \log^2 d$ noisy samples suffice to recover a $d \times d$ rank-r matrix with error that is proportional to the noise level. Recht (2011) improves the results of Candes and Recht (2009) on the number of observed entries required to reconstruct an unknown low-rank matrix. Negahban and Wainwright (2012) instead used nuclear-norm penalized least squares to recover the matrix. They derived the statistical error of the corresponding M-estimator and showed that it matched the information-theoretic lower bound up to logarithmic factors. Additional results and references can be found in the recent papers Chen et al. (2019, 2020) where the optimality and statistical inferences of the nuclear-norm penalization method are developed.

Our work aims to handle the presence of heavy-tailed noises, possibly with asymmetrical and heteroscedastic distributions, in the general trace regression. Based on the shrinkage principle, we develop a new loss function called the robust quadratic loss, which is constructed by plugging robust covariance estimators in the ℓ_2 risk function. Then we obtain the estimator $\widehat{\mathbf{\Theta}}$ by minimizing this new robust quadratic loss plus nuclear-norm penalty. By tailoring the analysis of Negahban et al. (2012) to this new loss, we can establish statistical rates in estimating $\mathbf{\Theta}^*$ just as those in Negahban et al. (2012) for the sub-Gaussian distributions, while relaxing the assumptions on the noise and design to allow heavy tails. This result is very generic and applicable to all four specific aforementioned problems.

Our robust approach is particularly simple and intuitive: it truncates or shrinks the response variables, depending on whether the responses are univariate or multivariate. Under the setting of sub-Gaussian design, large responses are very likely to be due to the outliers of noises. This explains why we need to truncate the responses when we have light-tailed covariates. If the covariates are also heavy-tailed, we need to truncate the designs as well. It turns out that appropriate truncation does not induce significant bias or hurt the restricted strong convexity of the loss function. This data robustfication enables us to apply the penalized least-squares method to recover sparse vectors or low-rank matrices. Under only bounded moment conditions for either noise or covariates, our robust estimator achieves the same statistical error rate as that under the case of the sub-Gaussian design and noise. The crucial component in our analysis is to obtain the sharp spectral-norm convergence rate for the robust covariance matrices based on the shrunk data. Note that other robust covariance estimation method, such as the RA-covariance estimation in Fan, Li and Wang (2017), are also possible to enjoy similar error rates, but this has not yet been fully studied for trace regression. So we only focus

on the shrinkage method, as it is computationally efficient, straightforward to analyze and always gives semipositive definite estimated covariance.

The successful application of the shrinkage sample covariance in multi-task regression also inspires us to study its statistical error in covariance estimation. It turns out that as long as the random samples $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ have bounded fourth moment in the sense that $\sup_{\mathbf{v}\in\mathcal{S}^{d-1}} \mathbb{E}(\mathbf{v}^\top \mathbf{x}_i)^4 \le R$, where $\mathcal{S}^{d-1}$ is the d-dimensional unit sphere, our ℓ_4-norm shrinkage sample covariance $\widetilde{\mathbf{\Sigma}}_n$ achieves the statistical error rate of order $O_{\mathbb{P}}(\sqrt{d\log d/n})$ under the spectral norm. This rate is optimal up to logarithmic term. In comparison, the naive sample covariance matrix $\overline{\mathbf{\Sigma}}_n$ only achieves rate of order $O_{\mathbb{P}}(\sqrt{d/n} \vee (d/n))$ according to Theorem 5.39 in Vershynin (2010) under sub-Gaussian data assumption. So sample covariance matrix itself surprisingly does not achieve optimality when dimension is high ($p \gg n$) even with Gaussian data. We will show in simulations that under the high-dimensional regime, $\widetilde{\mathbf{\Sigma}}$ indeed outperforms $\overline{\mathbf{\Sigma}}_n$ with sub-Gaussian samples. Therefore, shrinkage not only overcomes heavy-tailed corruption, but also mitigates curse of dimensionality. In terms of the elementwise max-norm, it is not hard to show that appropriate elementwise truncation of the data delivers a truncated sample covariance with statistical error rate of order $O_{\mathbb{P}}(\sqrt{\log d/n})$. This estimator can further be regularized if the true covariance has sparsity or other specific structure. See, for example, Meinshausen and Bühlmann (2006), Bickel and Levina (2008), Lam and Fan (2009), Cai and Liu (2011), Cai and Zhou (2012), Fan, Liao and Mincheva (2013), among others.

The current paper is organized as follows. In Section 2, we introduce the trace regression model and its four well-known examples: the linear model, matrix compressed sensing, matrix completion and multitask regression. Then we develop the generalized ℓ_2 loss, the truncated and shrinkage sample covariance and corresponding M-estimators. In Section 3, we present our main theoretical results. We first demonstrate through Theorem 1 the conditions required on the robust covariance inputs to ensure the desired statistical error rates of the M-estimator. Then we apply this theorem to all the specific aforementioned problems and explicitly derive the specific error rates. Section 4 studies the convergence properties of the shrinkage covariance estimator under the spectral norm. It should be of its own interest. Finally, we present simulation analyses in Section 5, which demonstrate the advantage of our robust estimators over the standard ones. The associated optimization algorithms are also discussed. All the proofs are relegated to the Appendix in the Supplementary Material (Fan, Wang and Zhu (2021)) for references.

Before presenting the detailed model and methodology, we first collect the general notation used in the paper. We follow the common convention of using boldface letters for vectors and matrices and using regular letters for scalars. For a vector $\mathbf{x}$, define $\|\mathbf{x}\|_q$ to be its ℓ_q norm; specifically, $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_2$ denote the ℓ_1 norm and ℓ_2 norm of $\mathbf{x}$, respectively. We use $\mathbb{R}^{d_1 d_2}$ to denote the space of $d_1 d_2$-dimensional real vectors, and use $\mathbb{R}^{d_1\times d_2}$ to denote the space of d_1-by-d_2 real matrices. For a matrix $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$, define $\|\mathbf{X}\|_{\mathrm{op}}$, $\|\mathbf{X}\|_N$, $\|\mathbf{X}\|_F$ and $\|\mathbf{X}\|_{\max}$ to be its operator norm, nuclear norm, Frobenius norm and elementwise max norm, respectively. We use $\mathrm{vec}(\mathbf{X})$ to denote vectorized version of $\mathbf{X}$, that is, $\mathrm{vec}(\mathbf{X}) = (\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_{d_2}^\top)^\top$, where $\mathbf{X}_j$ is the jth column of $\mathbf{X}$. Conversely, for a vector $\mathbf{x} \in \mathbb{R}^{d_1 d_2}$, we use $\mathrm{mat}(\mathbf{x})$ to denote the d_1-by-d_2 matrix constructed by $\mathbf{x}$, where $(x_{(j-1)d_1+1}, \ldots, x_{jd_1})^\top$ is the j^{th} column of $\mathrm{mat}(\mathbf{x})$. For any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1\times d_2}$, define the inner product $\langle \mathbf{A}, \mathbf{B}\rangle := \mathrm{Tr}(\mathbf{A}^\top \mathbf{B})$ where Tr is the trace operator. We denote $\mathrm{diag}(\mathbf{M}_1, \ldots, \mathbf{M}_n)$ to be the block diagonal matrix with the diagonal blocks as $\mathbf{M}_1, \ldots, \mathbf{M}_n$. For two Hilbert spaces $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} \perp \mathbf{B}$ if $\mathbf{A}$ and $\mathbf{B}$ are orthogonal to each other. For two scalar series $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we say $a_n \asymp_C b_n$ if there exist constants $0 < c_1 < c_2$ such that $c_1 a_n \le b_n \le c_2 a_n$ for $1 \le n < \infty$ where c_1, c_2 depend on C. For a random variable X, define its sub-Gaussian norm $\|X\|_{\psi_2} := \sup_{p\ge 1}(\mathbb{E}|X|^p)^{1/p}/\sqrt{p}$ and its subexponential norm $\|X\|_{\psi_1} := \sup_{p\ge 1}(\mathbb{E}|X|^p)^{1/p}/p$. For a

random vector $\mathbf{x} \in \mathbb{R}^d$, we define its sub-Gaussian norm $\|\mathbf{x}\|_{\psi_2} := \sup_{\mathbf{v}\in\mathcal{S}^{d-1}} \|\mathbf{v}^\top \mathbf{x}\|_{\psi_2}$ and subexponential norm $\|\mathbf{x}\|_{\psi_1} := \sup_{\mathbf{v}\in\mathcal{S}^{d-1}} \|\mathbf{v}^\top \mathbf{x}\|_{\psi_1}$. Given $x, y \in \mathbb{R}$, denote $\max(x, y)$ and $\min(x, y)$ by $x \vee y$ and $x \wedge y$, respectively. Let $\mathbf{e}_j$ be the unit vector with the jth element 1 and 0 elsewhere.

2. Model and methodology.

2.1. *Trace regression*. In this paper, we consider the trace regression model. Suppose we have N matrices $\{\mathbf{X}_i \in \mathbb{R}^{d_1\times d_2}\}_{i=1}^N$ and responses $\{Y_i \in \mathbb{R}\}_{i=1}^N$. We say $\{(Y_i, \mathbf{X}_i)\}_{i=1}^N$ follow the trace regression model if

$$Y_i = \langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle + \varepsilon_i, \tag{2.1}$$

where $\mathbf{\Theta}^* \in \mathbb{R}^{d_1\times d_2}$ is the true coefficient matrix, $\mathbb{E}\,\mathbf{X}_i = \mathbf{0}$ and $\{\varepsilon_i\}_{i=1}^N$ are independent noises satisfying $\mathbb{E}(\varepsilon_i|\mathbf{X}_i) = 0$. Note that here we do not require ε_i to be independent of $\mathbf{X}_i$. Model (2.1) includes the following specific cases:

- *Linear regression*: $d_1 = d_2 = d$, and $\{\mathbf{X}_i\}_{i=1}^N$ and $\mathbf{\Theta}^*$ are diagonal. Let $\mathbf{x}_i$ and $\boldsymbol{\theta}^*$ denote the vectors of diagonal elements of $\mathbf{X}_i$ and $\mathbf{\Theta}^*$, respectively, that is, $\mathbf{X}_i = \operatorname{diag}(x_{i1}, \ldots, x_{id})$ and $\mathbf{\Theta}^* = \operatorname{diag}(\theta_1^*, \ldots, \theta_d^*)$. Then, (2.1) reduces to familiar linear model: $Y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + \varepsilon_i$. Having a low-rank $\mathbf{\Theta}^*$ is then equivalent to having a sparse $\boldsymbol{\theta}^*$.
- *Compressed sensing*: For matrix compressed sensing, entries of $\mathbf{X}_i$ jointly follow the Gaussian distribution or other ensembles. For vector compressed sensing, we can take $\mathbf{X}_i$ and $\mathbf{\Theta}^*$ as diagonal matrices.
- *Matrix completion*: $\mathbf{X}_i$ is a singleton, that is, $\mathbf{X}_i = \mathbf{e}_{j(i)}\mathbf{e}_{k(i)}^\top$ for $1 \le j(i) \le d_1$ and $1 \le k(i) \le d_2$. In other words, a random entry of the matrix $\mathbf{\Theta}$ is observed along with noise for each sample.
- *Multitask learning*: The multitask (reduced-rank) regression assumes

$$\mathbf{y}_j = {\mathbf{\Theta}^*}^\top \mathbf{x}_j + \boldsymbol{\varepsilon}_j, \quad j = 1, \ldots, n, \tag{2.2}$$

  where $\mathbf{x}_j \in \mathbb{R}^{d_1}$ is the covariate vector, $\mathbf{y}_j \in \mathbb{R}^{d_2}$ is the response vector, $\mathbf{\Theta}^* \in \mathbb{R}^{d_1\times d_2}$ is the coefficient matrix and $\boldsymbol{\varepsilon}_j \in \mathbb{R}^{d_2}$ is the noise with each entry independent of each other. See, for example, Kim and Xing (2012) and Reinsel and Velu (2013). Each sample $(\mathbf{y}_j, \mathbf{x}_j)$ consists of d_2 responses and is equivalent to d_2 data points in (2.1), that is, $\{(Y_{(j-1)d_2+i} = y_{ji}, \mathbf{X}_{(j-1)d_2+i} = \mathbf{x}_j\mathbf{e}_i^\top)\}_{i=1}^{d_2}$. Therefore, n samples in (2.2) correspond to $N = nd_2$ observations in (2.1).

In this paper, we impose rank constraint on the coefficient matrix $\mathbf{\Theta}^*$. Rank constraint can be viewed as a generalized sparsity constraint for two-dimensional matrices. For linear regression, rank constraint is equivalent to the sparsity constraint since $\mathbf{\Theta}^*$ is diagonal. The rank constraint reduces the effective number of parameters in $\mathbf{\Theta}^*$ and arises frequently in many applications. Consider the Netflix problem, for instance, where $\mathbf{\Theta}_{ij}^*$ is the intrinsic score of film j given by customer i and we would like to recover the entire $\mathbf{\Theta}^*$ with only partial observations. Given that movies of similar types or qualities should receive similar scores from viewers, columns of $\mathbf{\Theta}^*$ should share colinearity, thus delivering a low-rank structure of $\mathbf{\Theta}^*$. The rationale of the model can also be understood from the celebrated factor model in finance and economics (Fan and Yao (2015)), which assumes that several market risk factors drive the returns of a large panel of stocks. Consider $N \times T$ matrix $\mathbf{Y}$ of N stock returns (like movies) over T days (like viewers). These financial returns are driven by K factors $\mathbf{F}$ ($K \times T$ matrix, representing K risk factors realized on T days or K attributes realized on T

movies) with a loading matrix $\mathbf{B}$ ($N \times K$ matrix, reflecting individual's preference on these attritributes). The factor model admits the following form:

$$\mathbf{Y} = \mathbf{BF} + \mathbf{E},$$

where $\mathbf{E}$ is idiosyncratic noise. Since $\mathbf{BF}$ has a small rank K, $\mathbf{BF}$ can be regarded as the low-rank matrix $\mathbf{\Theta}^*$ in the matrix completion problem. If all movies were rated by all viewers in the Netflix problem, the ratings should also be modeled as a low-rank matrix plus noise, namely, there should be several latent factors that drive ratings of movies. The major challenge of the matrix completion problem is that there are many missing entries.

Exact low-rank may be too stringent to model the real-world situations. Instead, we consider near low-rank $\mathbf{\Theta}^*$ satisfying

$$\mathcal{B}_q(\mathbf{\Theta}^*) := \sum_{i=1}^{d_1 \wedge d_2} \sigma_i(\mathbf{\Theta}^*)^q \le \rho, \tag{2.3}$$

where $0 \le q \le 1$. Note that when $q = 0$, the constraint (2.3) is the exact rank constraint. Restriction on $\mathcal{B}_q(\mathbf{\Theta}^*)$ ensures that the singular values decay fast enough; it is more general and natural than the exact low-rank assumption. In the analysis, we can allow ρ to grow with dimensionality and sample size.

A popular method for estimating $\mathbf{\Theta}^*$ is the penalized empirical loss that solves $\widehat{\mathbf{\Theta}} \in \operatorname{argmin}_{\mathbf{\Theta} \in \mathcal{S}} \mathcal{L}(\mathbf{\Theta}) + \lambda_N \mathcal{P}(\mathbf{\Theta})$, where $\mathcal{S}$ is a convex set in $\mathbb{R}^{d_1 \times d_2}$, $\mathcal{L}(\mathbf{\Theta})$ is the loss function, λ_N is the tuning parameter and $\mathcal{P}(\mathbf{\Theta})$ is a rank penalization function. Most of the previous work, for example, Koltchinskii, Lounici and Tsybakov (2011) and Negahban and Wainwright (2011), chose $\mathcal{L}(\mathbf{\Theta}) = \sum_{1 \le i \le N} (Y_i - \langle \mathbf{\Theta}, \mathbf{X}_i \rangle)^2$ and $\mathcal{P}(\mathbf{\Theta}) = \|\mathbf{\Theta}\|_N$, and derived the rate for $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ under the assumption of sub-Gaussian or subexponential noise. However, the ℓ_2 loss is sensitive to outliers and is unable to handle the data with moderately heavy or heavy tails.

2.2. *Robustifying ℓ_2 loss.* We aim to accommodate heavy-tailed noise and design for the near low-rank matrix recovery by robustifying the traditional ℓ_2 loss. We first notice that the ℓ_2 risk can be expressed as

$$\begin{aligned} R(\mathbf{\Theta}) &= \mathbb{E}\,\mathcal{L}(\mathbf{\Theta}) = \mathbb{E}\big(Y_i - \langle \mathbf{\Theta}, \mathbf{X}_i \rangle\big)^2 \\ &= \mathbb{E}\, Y_i^2 - 2\langle \mathbf{\Theta}, \mathbb{E}\, Y_i \mathbf{X}_i \rangle + \operatorname{vec}(\mathbf{\Theta})^\top \mathbb{E}\big(\operatorname{vec}(\mathbf{X}_i) \operatorname{vec}(\mathbf{X}_i)^\top\big) \operatorname{vec}(\mathbf{\Theta}) \\ &\equiv \mathbb{E}\, Y_i^2 - 2\langle \mathbf{\Theta}, \mathbf{\Sigma}_{Y\mathbf{X}} \rangle + \operatorname{vec}(\mathbf{\Theta})^\top \mathbf{\Sigma}_{\mathbf{XX}} \operatorname{vec}(\mathbf{\Theta}). \end{aligned} \tag{2.4}$$

Ignoring $\mathbb{E}\, Y_i^2$, if we substitute $\mathbf{\Sigma}_{Y\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{XX}}$ by their corresponding sample covariances, we recover the empirical ℓ_2 loss. This inspires us to define a generalized ℓ_2 loss as

$$\mathcal{L}(\mathbf{\Theta}) := -\langle \widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}, \mathbf{\Theta} \rangle + \frac{1}{2} \operatorname{vec}(\mathbf{\Theta})^\top \widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \operatorname{vec}(\mathbf{\Theta}), \tag{2.5}$$

where $\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ are estimators of $\mathbb{E}(Y_i \mathbf{X}_i)$ and $\mathbb{E}\{\operatorname{vec}(\mathbf{X}_i) \operatorname{vec}(\mathbf{X}_i)^\top\}$, respectively. Equation (2.5) suggests that one can first seek reliable covariance and cross-covariance estimators $\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ to construct an empirical ℓ_2 risk function $\mathcal{L}(\mathbf{\Theta})$, and then recover the parameter of interest $\mathbf{\Theta}^*$ by minimizing $\mathcal{L}(\mathbf{\Theta})$. Previous works exploited similar ideas to obtain sharp statistical results in high-dimensional linear models. For instance, under the matrix completion setup, Koltchinskii, Lounici and Tsybakov (2011) incorporated the (known) covariance of $\mathbf{X}_i$ into the ℓ_2 risk, and showed that the resulting nuclear-norm penalized estimator enjoys both a simple explicit form and minimax optimal rate. Another example is Bellec

et al. (2018) who focus on the adaptations of the supervised lasso estimator to the semisupervised learning and transductive learning settings. Therein both the labeled and unlabeled features are employed to estimate the feature covariance matrix $\mathbf{\Sigma}$ in the loss function, for the purpose of achieving sharper statistical accuracy than the lasso estimator based on only the labeled data. Our work also replaces the population covariance in the ℓ_2 risk with the empirical estimates. The difference is that we use truncated or shrunk sample covariance in (2.5) to guard against heavy-tailed corruptions. This explains the reason we allow bounded-moment design and response, while Bellec et al. (2018), say, assume almost sure bounds on the data.

In this paper, we study the following M-estimator of $\mathbf{\Theta}^*$ with the generalized ℓ_2 loss:

$$\widehat{\mathbf{\Theta}} \in \operatorname*{argmin}_{\mathbf{\Theta}\in\mathcal{S}} -\langle \widehat{\mathbf{\Sigma}}_{YX}, \mathbf{\Theta}\rangle + \frac{1}{2}\operatorname{vec}(\mathbf{\Theta})^\top \widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \operatorname{vec}(\mathbf{\Theta}) + \lambda_N \|\mathbf{\Theta}\|_N, \tag{2.6}$$

where $\mathcal{S}$ is a convex set in $\mathbb{R}^{d_1\times d_2}$. To handle heavy-tailed noise and design, we need to employ robust estimators $\widehat{\mathbf{\Sigma}}_{YX}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$. For ease of presentation, we always first consider the case where the design is sub-Gaussian and the response is heavy-tailed, and then further allow the design to have heavy-tailed distribution if it is appropriate for the specific problem setup.

We now introduce the robust covariance estimators to be plugged in (2.6) by the principle of truncation, or more generally shrinkage. The intuition is that shrinkage reduces sensitivity of the estimator to the heavy-tailed corruption. However, shrinkage induces bias. Our theories revolve around finding appropriate shrinkage level so as to ensure the induced bias is not too large and the final statistical error rate is sharp. Different problem setups have different forms of $\widehat{\mathbf{\Sigma}}_{YX}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$, but the principle of shrinkage of data is universal. For the linear regression, matrix compressed sensing and matrix completion, in which the response is univariate, $\widehat{\mathbf{\Sigma}}_{YX}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ take the following forms:

$$\widehat{\mathbf{\Sigma}}_{YX} = \widehat{\mathbf{\Sigma}}_{\widetilde{Y}\widetilde{\mathbf{X}}} = \frac{1}{N}\sum_{i=1}^N \widetilde{Y}_i\widetilde{\mathbf{X}}_i \quad \text{and} \quad \widehat{\mathbf{\Sigma}}_{\mathbf{XX}} = \widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}} = \frac{1}{N}\sum_{i=1}^N \operatorname{vec}(\widetilde{\mathbf{X}}_i)\operatorname{vec}(\widetilde{\mathbf{X}}_i)^\top, \tag{2.7}$$

where tilde notation means truncated versions of the random variables if they have heavy tails and equals the original random variables (truncation threshold is infinite) if they have light tails.

For the multitask regression, similar idea continues to apply. However, writing (2.2) in the general form of (2.1) requires adaptation of more complicated notation. We choose

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_{YX} &= \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^{d_2} \widetilde{Y}_{ij}\widetilde{\mathbf{x}}_i\mathbf{e}_j^\top = \frac{1}{d_2}\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}} \quad \text{and} \\
\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} &= \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^{d_2} \operatorname{vec}\big(\widetilde{\mathbf{x}}_i\mathbf{e}_j^\top\big)\operatorname{vec}\big(\widetilde{\mathbf{x}}_i\mathbf{e}_j^\top\big)^\top = \frac{1}{d_2}\operatorname{diag}(\underbrace{\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}},\ldots,\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}}_{d_2}),
\end{aligned} \tag{2.8}$$

where

$$\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}} = \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{x}}_i\widetilde{\mathbf{y}}_i^\top \quad \text{and} \quad \widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} = \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top$$

and $\widetilde{\mathbf{y}}_i$ and $\widetilde{\mathbf{x}}_i$ are again transformed versions of $\mathbf{y}_i$ and $\mathbf{x}_i$. The tilde notation means shrinkage for heavy-tailed variables and identity mapping (no shrinkage) for light-tailed variables. The factor d_2^{-1} is due to the fact that n independent samples under model (2.2) are treated as $N = nd_2$ samples in (2.1). As we shall see, under only bounded moment assumptions of the design and noise, the proposed truncated or shrinkage covariance enjoys desired convergence

rate to its population counterpart. This leads to a sharp $\mathcal{M}$-estimator $\widehat{\mathbf{\Theta}}$, whose statistical error rates match those established in Negahban and Wainwright (2011, 2012) under the setting of sub-Gaussian design and noise.

Note that using truncated or shrinkage covariance in the generalized ℓ_2 loss is equivalent to evaluating the traditional ℓ_2 loss on the truncated or shrunk data. Nevertheless, instead of directly analyzing the quadratic loss with shrunk data, we study the robust covariance first and then derive the error rate of the corresponding M-estimator. This analytical framework is modular and allows other potential robust covariance estimators to be plugged in, for instance, those based on Kendall's tau (Fan, Liu and Wang (2018)), median of means (Minsker (2015)), etc. As long as the recruited $\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ satisfy the set of sufficient conditions given by Theorem 1, the corresponding $\widehat{\mathbf{\Theta}}$ achieves the desired convergence. One recent work Loh and Tan (2018) took a similar strategy in estimating the high-dimensional precision matrix. The authors proposed to plug in appropriately chosen robust covariance estimators into graphical LASSO and CLIME and established sharp error bounds for the corresponding estimators of the precision matrix.

Finally, we conjecture that minimizing Huber loss, Tukey's biweight loss or other robust but maybe nonconvex losses (Loh (2017), Fan, Li and Wang (2017)) with nuclear norm regularization can also achieve nearly minimax optimal rate. However, these papers typically focus on high-dimensional sparse regression and so far we have not seen any statistical guarantee for these methods under the trace regression for the low-rank recovery. Moreover, our method is simple to implement with guaranteed optimization efficiency, while the other methods lack reliable and fast algorithms with convergence guarantee. Since our method is equivalent to applying the standard method to the truncated or shrunk data, the optimization is still a least square problem with nuclear-norm penalization. This is amenable to efficient algorithms such as the Peaceman–Rachford splitting method (PRSM) as described in Section 5.

3. Main results. In this section, we derive the statistical error rate of $\widehat{\mathbf{\Theta}}$ defined by (2.6). We always assume $d_1, d_2 \geq 2$ and $\rho > 1$ in (2.3). We first present the following general theorem that gives the estimation errors $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ and $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_N$.

THEOREM 1. *Define* $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$, *where* $\mathbf{\Theta}^*$ *satisfies* $\mathcal{B}_q(\mathbf{\Theta}^*) \leq \rho$. *Suppose* $\operatorname{vec}(\widehat{\mathbf{\Delta}})^\top \widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \operatorname{vec}(\widehat{\mathbf{\Delta}}) \geq \kappa_{\mathcal{L}} \|\widehat{\mathbf{\Delta}}\|_F^2$, *where* $\kappa_{\mathcal{L}}$ *is a positive constant that does not depend on* $\widehat{\mathbf{\Delta}}$. *Choose* $\lambda_N \geq 2\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \operatorname{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \operatorname{vec}(\mathbf{\Theta}^*))\|_{\mathrm{op}}$. *Then we have that*

$$\|\widehat{\mathbf{\Delta}}\|_F^2 \leq C_1 \rho \left(\frac{\lambda_N}{\kappa_{\mathcal{L}}}\right)^{2-q} \quad \text{and} \quad \|\widehat{\mathbf{\Delta}}\|_N \leq C_2 \rho \left(\frac{\lambda_N}{\kappa_{\mathcal{L}}}\right)^{1-q},$$

where C_1 *and* C_2 *are two universal constants*.

First of all, the above result is deterministic and nonasymptotic. As we can see from the theorem above, the statistical performance of $\widehat{\mathbf{\Theta}}$ relies on the restricted eigenvalue (RE) property of $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$, which was first studied by Bickel, Ritov and Tsybakov (2009). When the design is sub-Gaussian, we choose $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ to be the traditional sample covariance, whose RE property has been well established (e.g., Rudelson and Zhou (2013), Negahban and Wainwright (2011, 2012)). We will specify these results when we need them in the sequel. When the design only satisfies bounded moment conditions, we choose $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} = \widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}}$ to be the sample covariance of shrunk data. We show that with appropriate level of shrinkage, $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}}$ still retains the RE property, thus satisfying the conditions of the theorem.

Second, the conclusion of the theorem says that $\|\widehat{\mathbf{\Delta}}\|_F^2$ and $\|\widehat{\mathbf{\Delta}}\|_N$ are proportional to λ_N^{2-q} and λ_N^{1-q}, respectively, but we require $\lambda_N \geq \|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \operatorname{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \operatorname{vec}(\mathbf{\Theta}^*)\|_{\mathrm{op}}$. This implies that

$\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*)\|_{\mathrm{op}}$ is crucial to the statistical error of $\widehat{\mathbf{\Theta}}$. In the following subsections, we will derive the rate of $\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*)\|_{\mathrm{op}}$ for all the aforementioned specific problems with only bounded moment conditions on the response, and in some cases also on the design. Under such weak assumptions, we show that the proposed robust M-estimator possesses the same rates as those presented in Negahban and Wainwright (2011, 2012) with sub-Gaussian assumptions on the design and noise.

Finally, we emphasize one key technical contribution of our work: the bias-and-variance tradeoff through tuning of the truncation level τ. As we explained above, $\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*)\|_{\mathrm{op}}$ is crucial to the estimation accuracy of $\widehat{\mathbf{\Theta}}$. In our analysis, we decompose $\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*)$ into the following three terms:

$$\begin{aligned}
&\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*)\\
&\quad= \underbrace{\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \mathbb{E}[\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}]}_{\text{concentration term 1}} + \underbrace{\mathbb{E}[\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}] - \mathbb{E}\big[\mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*)\big]}_{\text{bias term}}\\
&\qquad+ \underbrace{\mathbb{E}\big[\mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*)\big] - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*)}_{\text{concentration term 2}}.
\end{aligned}$$

Choosing $\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ to be the truncated or shrinkage sample covariance, we will show that the truncation level τ only contributes to high-order terms in both concentration terms above. This allows us to strike a perfect balance between the bias term and the concentration terms and establish the optimal rate for $\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*))\|_{\mathrm{op}}$. In contrast, if we simply treat the truncated response as a sub-Gaussian random variable bounded by τ, τ will contribute to the leading terms in the concentration bounds, which implies suboptimal results. This observation also inspired us to construct the ℓ_4-norm shrinkage sample covariance and establish its (near) minimax optimal rate in Section 4. This new robust covariance estimator is employed in the multitasking regression with heavy-tailed data and leads to a minimax optimal MLE of $\mathbf{\Theta}^*$.

3.1. *Linear model.* For the linear regression problem, $\mathbf{\Theta}^*$ and $\{\mathbf{X}_i\}_{i=1}^N$ are $d \times d$ diagonal matrices. We denote the diagonals of $\mathbf{\Theta}^*$ and $\{\mathbf{X}_i\}_{i=1}^N$ by $\boldsymbol{\theta}^*$ and $\{\mathbf{x}_i\}_{i=1}^N$, respectively for ease of presentation. The optimization problem in (2.6) reduces to

$$\widehat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} -\widehat{\mathbf{\Sigma}}_{Y\mathbf{x}}^{\top}\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^{\top}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}\boldsymbol{\theta} + \lambda_N\|\boldsymbol{\theta}\|_1, \tag{3.1}$$

where $\widehat{\mathbf{\Sigma}}_{Y\mathbf{x}} = \widehat{\mathbf{\Sigma}}_{\widetilde{Y}\widetilde{\mathbf{x}}} = N^{-1}\sum_{i=1}^N \widetilde{Y}_i\widetilde{\mathbf{x}}_i$, $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}} = \widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} = N^{-1}\sum_{i=1}^N \widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\top}$. When the design is sub-Gaussian, we only need to truncate the response. Therefore, we choose the Winsorization $\widetilde{Y}_i = \widetilde{Y}_i(\tau) = \mathrm{sgn}(Y_i)(|Y_i| \wedge \tau)$ and $\widetilde{\mathbf{x}}_i = \mathbf{x}_i$, for some threshold τ. When the design is heavy-tailed, we choose $\widetilde{Y}_i(\tau_1) = \mathrm{sgn}(Y_i)(|Y_i| \wedge \tau_1)$ and $\widetilde{x}_{ij} = \mathrm{sgn}(x_{ij})(|x_{ij}| \wedge \tau_2)$, where τ_1 and τ_2 are both predetermined threshold values. To avoid redundancy, we will not repeat stating these choices in lemmas or theorems in this subsection.

To establish the statistical error rate of $\widehat{\boldsymbol{\theta}}$ in (3.1), in the following lemma, we derive the rate of $\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{x}} - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*))\|_{\mathrm{op}}$ in (2.6) for the sub-Gaussian design and bounded-moment (polynomial tail) design, respectively. Note here that

$$\big\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{x}} - \mathrm{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\,\mathrm{vec}(\mathbf{\Theta}^*))\big\|_{\mathrm{op}} = \big\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}\boldsymbol{\theta}^*\big\|_{\max}.$$

LEMMA 1. *Uniform convergence of cross covariance.*

(a) Sub-Gaussian design. *Consider the following conditions*:

(C1) $\{\mathbf{x}_i\}_{i=1}^N$ *are i.i.d. sub-Gaussian vectors with* $\|\mathbf{x}_i\|_{\psi_2} \le \kappa_0 < \infty$, $\mathbb{E}\,\mathbf{x}_i = \mathbf{0}$ *and* $\lambda_{\min}(\mathbb{E}\,\mathbf{x}_i\mathbf{x}_i^\top) \ge \kappa_{\mathcal{L}} > 0$;

(C2) $\forall i = 1, \ldots, N$, $\mathbb{E}\,|Y_i|^{2k} \le M < \infty$ *for some* $k > 1$.

Choose $\tau \asymp_{k,\kappa_0} \sqrt{M^{1/k} N / \log d}$. *There exists* $C > 0$, *depending only on* k, κ_0 *and* $\kappa_{\mathcal{L}}$, *such that as long as* $\log d < N$, *for any* $\xi > 1$,

$$\mathbb{P}\left(\left\| \widehat{\mathbf{\Sigma}}_{Y\mathbf{x}}(\tau) - \widehat{\mathbf{\Sigma}}_{\mathbf{xx}} \boldsymbol{\theta}^* \right\|_{\max} \ge C\xi \sqrt{\frac{M^{1/k} \log d}{N}} \right) \le 2d^{-(\xi-1)}. \tag{3.2}$$

(b) Bounded moment design. *Consider instead the following set of conditions*:

(C1′) $\|\boldsymbol{\theta}^*\|_1 \le R < \infty$;

(C2′) $\{\mathbf{x}_i\}_{i=1}^N$ *are i.i.d., and for any* $1 \le j \le d$, $\mathbb{E}\,|x_{1j}|^4 \le M < \infty$;

(C3′) $\forall i = 1, \ldots, N$, $\mathbb{E}\,|Y_i|^4 \le M < \infty$.

Choose $\tau_1, \tau_2 \asymp_R (MN/\log d)^{\frac{1}{4}}$. *For any* $\xi > 2$, *we have that*

$$\mathbb{P}\left(\left\| \widehat{\mathbf{\Sigma}}_{Y\mathbf{x}}(\tau_1, \tau_2) - \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}(\tau_2) \boldsymbol{\theta}^* \right\|_{\max} > C \sqrt{\frac{M\xi \log d}{N}} \right) \le 2d^{-(\xi-2)},$$

where C *only depends on* R.

REMARK 1. If we choose $\widehat{\mathbf{\Sigma}}_{Y\mathbf{x}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ to be the sample covariance, that is, $\widehat{\mathbf{\Sigma}}_{Y\mathbf{x}} = \overline{\mathbf{\Sigma}}_{Y\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N Y_i \mathbf{x}_i$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}} = \overline{\mathbf{\Sigma}}_{\mathbf{xx}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top$, Corollary 2 of Negahban et al. (2012) showed that under the sub-Gaussian noise and design,

$$\left\| \overline{\mathbf{\Sigma}}_{Y\mathbf{x}} - \overline{\mathbf{\Sigma}}_{\mathbf{xx}} \boldsymbol{\theta}^* \right\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log d}{N}} \right).$$

This is the same rate as what we achieved under only the bounded moment conditions on response and design. The assumption $\|\boldsymbol{\theta}^*\|_1 < \infty$ holds if the sparsity level ρ is finite. This condition is used to bound $|\widetilde{\mathbf{x}}_i' \boldsymbol{\theta}^*| \le \tau_2 \|\boldsymbol{\theta}^*\|_1$, and seems necessary for heavy-tailed design which lacks the nice structure of sub-Gaussian design, where the ψ_2 norm of $\mathbf{x}_i' \boldsymbol{\theta}^*$ is determined by $\|\boldsymbol{\theta}^*\|_2$.

Next, we establish the restricted strong convexity of the proposed robust ℓ_2 loss.

LEMMA 2. *Restricted strong convexity.*

(a) Sub-Gaussian design. *Under Condition* (C1) *of Lemma* 1, *it holds for any* $\xi > 1$ *that*

$$\begin{aligned} &\mathbb{P}\left(\mathbf{v}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{xx}} \mathbf{v} \ge \frac{1}{2} \mathbf{v}^\top \mathbf{\Sigma}_{\mathbf{xx}} \mathbf{v} - \frac{C\xi \log d}{N} \|\mathbf{v}\|_1^2, \forall \mathbf{v} \in \mathbb{R}^d \right) \\ &\quad \ge 1 - \frac{d^{-(\xi-1)}}{3} - 2d \exp(-cN), \end{aligned} \tag{3.3}$$

where C *depends only on* κ_0 *and* $\kappa_{\mathcal{L}}$, *and* c *is a universal constant*.

(b) Bounded moment design. *Assume that* $\mathbf{x}_i$ *satisfies Condition* (C2′) *of Lemma* 1. *Choosing* $\tau_2 \asymp_R (MN/\log d)^{\frac{1}{4}}$, *we have for any* $\xi > 2$,

$$\mathbb{P}\left(\mathbf{v}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}(\tau_2) \mathbf{v} \ge \mathbf{v}^\top \mathbf{\Sigma}_{\mathbf{xx}} \mathbf{v} - C \sqrt{\frac{M\xi \log d}{N}} \|\mathbf{v}\|_1^2, \forall \mathbf{v} \in \mathbb{R}^d \right) \le d^{-(\xi-2)}, \tag{3.4}$$

where C *depends only on* R.

REMARK 2. Comparing the results we get for sub-Gaussian and heavy-tailed design, we see that the coefficients before $\|\mathbf{v}\|_1^2$ are different. Under the sub-Gaussian design, that coefficient is of order $\log d/N$, while under the heavy-tailed design, the coefficient is of order $\sqrt{\log d/N}$. This difference leads to different scaling requirements for N, d and ρ in the sequel. As we shall see, the heavy-tailed design requires a stronger scaling condition to retain the same statistical rate as the sub-Gaussian design.

Finally we derive the statistical error rate of $\widehat{\boldsymbol{\theta}}$ as defined in (3.1).

THEOREM 2. *Assume* $\sum_{i=1}^d |\theta_i^*|^q \le \rho$, *where* $0 \le q \le 1$.

(a) Sub-Gaussian design. *Suppose Conditions* (C1) *and* (C2) *in Lemma* 1 *hold and* $\log d < N$. *Choose* $\tau \asymp_{k,\kappa_0} \sqrt{M^{1/k}N/\log d}$. *For any* $\xi > 1$, *let* $\lambda_N = 2C\xi\sqrt{M^{1/k}\log d/N}$, *where* C *and* ξ *are the same as in part* (*a*) *of Lemma* 1. *There exists* C_1, *depending only on* $\kappa_{\mathcal{L}}$ *and* κ_0, *such that as long as* $\rho\xi^{1-q}(\log d/N)^{1-(q/2)}M^{-q/(2k)} \le C_1$, *we have*

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\theta}}(\tau,\lambda_N)-\boldsymbol{\theta}^*\|_2^2 > C_2\rho\left(\frac{\xi^2 M^{1/k}\log d}{N}\right)^{1-(q/2)}\right\} \le \frac{7d^{-(\xi-1)}}{3} + 2d\exp(-cN)$$

and

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\theta}}(\tau,\lambda_N)-\boldsymbol{\theta}^*\|_1 > C_3\rho\left(\frac{\xi^2 M^{1/k}\log d}{N}\right)^{(1-q)/2}\right\} \le 3d^{-(\xi-1)} + 2d\exp(-cN),$$

where C_2, C_3 *only depend on* κ_0 *and* $\kappa_{\mathcal{L}}$.

(b) Bounded moment design. *Choose* τ_1, $\tau_2 \asymp_R (MN/\log d)^{1/4}$ *and* $\lambda_N = 2C \times \sqrt{M\xi\log d/N}$, *where* C *and* ξ *are the same as in part* (*b*) *of Lemma* 1. *Under Conditions* (C1′), (C2′) *and* (C3′), *there exists* C_1, *depending only on* R, *such that whenever* $\rho(M\xi\log d/N)^{(1-q)/2} \le C_1$,

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\theta}}(\tau_1,\tau_2,\lambda_N)-\boldsymbol{\theta}^*\|_2^2 \ge C_2\rho\left(\frac{M\xi\log d}{N}\right)^{1-(q/2)}\right\} \le 3d^{-(\xi-2)}$$

and

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\theta}}(\tau_1,\tau_2,\lambda_N)-\boldsymbol{\theta}^*\|_1 \ge C_3\rho\left(\frac{M\xi\log d}{N}\right)^{(1-q)/2}\right\} \le 3d^{-(\xi-2)},$$

where C_2 *and* C_3 *only depend on* R.

REMARK 3. Under both sub-Gaussian and heavy-tailed design, our proposed $\widehat{\boldsymbol{\theta}}$ achieves the minimax optimal rate of ℓ_2 norm established by Raskutti, Wainwright and Yu (2011). However, the difference lies in the scaling requirement on N, d and ρ. For sub-Gaussian design, we require $\rho(\log d/N)^{1-(q/2)}$ to be bounded, whereas for heavy-tailed design we need $\rho(\log d/N)^{(1-q)/2}$ to be bounded. Under the typical high-dimensional regime that $d \gg N \gg \log d$, the former is weaker. Therefore, heavy-tailed design requires stronger scaling than sub-Gaussian design to achieve the optimal statistical rates.

3.2. *Matrix compressed sensing.* For the matrix compressed sensing problem, since the design is chosen by users, we only consider the sub-Gaussian design. Hence, we keep the original design matrix and only truncate the response. In (2.7), choose $\widetilde{Y}_i = \operatorname{sgn}(Y_i)(|Y_i| \wedge \tau)$ and $\widetilde{\mathbf{X}}_i = \mathbf{X}_i$, then we have

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} &= \widehat{\mathbf{\Sigma}}_{\widetilde{Y}\widetilde{\mathbf{X}}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \operatorname{sgn}(Y_i)\big(|Y_i| \wedge \tau\big)\mathbf{X}_i \quad \text{and} \\ \widehat{\mathbf{\Sigma}}_{\mathbf{XX}} &= \frac{1}{N}\sum_{i=1}^{N} \operatorname{vec}(\mathbf{X}_i)\operatorname{vec}(\mathbf{X}_i)^{\top}. \end{aligned} \tag{3.5}$$

The following lemma explicitly derives the rate of $\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \operatorname{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\operatorname{vec}(\mathbf{\Theta}^*))\|_{\mathrm{op}}$. Note that here $\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}} - \operatorname{mat}(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\operatorname{vec}(\mathbf{\Theta}^*)) = \widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}(\tau) - \frac{1}{N}\sum_{i=1}^{N}\langle \mathbf{X}_i, \mathbf{\Theta}^*\rangle \mathbf{X}_i$.

LEMMA 3. *Consider the following conditions*:

(C1) $\{\operatorname{vec}(\mathbf{X}_i)\}_{i=1}^{N}$ *are i.i.d. sub-Gaussian vectors with* $\|\operatorname{vec}(\mathbf{X}_i)\|_{\psi_2} \le \kappa_0 < \infty$, $\mathbb{E}\,\mathbf{X}_i = \mathbf{0}$ *and* $\lambda_{\min}(\mathbb{E}\operatorname{vec}(\mathbf{X}_i)\operatorname{vec}(\mathbf{X}_i)^{\top}) \ge \kappa_{\mathcal{L}} > 0$.

(C2) $\forall i = 1, \ldots, N$, $\mathbb{E}\,|Y_i|^{2k} \le M < \infty$ *for some* $k > 1$.

Choose $\tau \asymp_{\kappa_0,\kappa_{\mathcal{L}},k} \sqrt{M^{1/k}N/(d_1+d_2)}$. *There exists* $C > 0$, *depending only on* κ_0, $\kappa_{\mathcal{L}}$ *and* k, *such that whenever* $d_1 + d_2 < N$, *for any* $\xi > \log 8$,

$$\begin{aligned} \mathbb{P}\left(\left\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}(\tau) - \frac{1}{N}\sum_{i=1}^{N}\langle \mathbf{X}_i, \mathbf{\Theta}^*\rangle \mathbf{X}_i\right\|_{\mathrm{op}} \ge C\sqrt{\frac{M^{1/k}(d_1+d_2)\xi}{N}}\right) \\ \le 4\exp\big((d_1+d_2)(\log 8 - \xi)\big), \end{aligned} \tag{3.6}$$

where C *only depends on* κ_0, $\kappa_{\mathcal{L}}$ *and* k.

REMARK 4. For the sample covariance $\overline{\mathbf{\Sigma}}_{Y\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N} Y_i\mathbf{X}_i$, Negahban and Wainwright (2011) showed that when the noise and design are sub-Gaussian,

$$\left\|\overline{\mathbf{\Sigma}}_{Y\mathbf{X}} - \frac{1}{N}\sum_{i=1}^{N}\langle \mathbf{X}_i, \mathbf{\Theta}^*\rangle \mathbf{X}_i\right\|_{\mathrm{op}} = \left\|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\mathbf{X}_i\right\|_{\mathrm{op}} = O_{\mathbb{P}}\big(\sqrt{(d_1+d_2)/N}\big).$$

Lemma 3 shows that $\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}(\tau)$ achieves the same rate for response with just bounded moments.

The following theorem gives the statistical error rate of $\widehat{\mathbf{\Theta}}$ in (2.6).

THEOREM 3. *Suppose Conditions* (C1) *and* (C2) *in Lemma* 3 *hold and* $\mathcal{B}_q(\mathbf{\Theta}^*) \le \rho$. *We further assume that* $\operatorname{vec}(\mathbf{X}_i)$ *is Gaussian. Choose* $\tau \asymp_{\kappa_0,\kappa_{\mathcal{L}},k} \sqrt{N/(d_1+d_2)}$ *and* $\lambda_N = 2C\sqrt{M^{1/k}\xi(d_1+d_2)/N}$, *where* C *is the same universal constant as in Lemma* 3. *There exist universal constants* $\{C_i\}_{i=1}^{3}$ *such that once* $\rho M^{-q/(2k)}((d_1+d_2)/N)^{1-(q/2)} \le C_1$, *we have that for any* $\xi > \log 8$,

$$\begin{aligned} &\mathbb{P}\left\{\|\widehat{\mathbf{\Theta}}(\tau,\lambda_N) - \mathbf{\Theta}^*\|_F^2 \ge C_2\rho\left(\frac{M^{1/k}(d_1+d_2)\xi}{N}\right)^{1-(q/2)}\right\} \\ &\quad\le 2\exp\left(-\frac{N}{32}\right) + 4\exp\big((d_1+d_2)(\log 8 - \xi)\big) \end{aligned}$$

and

$$\mathbb{P}\Big\{\big\|\widehat{\boldsymbol{\Theta}}(\tau,\lambda_N)-\boldsymbol{\Theta}^*\big\|_N \geq C_3\rho\Big(\frac{M^{1/k}(d_1+d_2)\xi}{N}\Big)^{(1-q)/2}\Big\}$$
$$\leq 2\exp\Big(-\frac{N}{32}\Big)+4\exp\big((d_1+d_2)(\log 8-\xi)\big).$$

REMARK 5. The Frobenius norm rate here is identical to the rate established under sub-Gaussian noise in Negahban and Wainwright (2011). When $q=0$, ρ is the upper bound of the rank of $\boldsymbol{\Theta}^*$ and the rate of convergence depends only on $\rho(d_1+d_2)\asymp\rho(d_1\vee d_2)$, the effective number of independent parameters in $\boldsymbol{\Theta}^*$, rather than the ambient dimension d_1d_2.

REMARK 6. The rate we derived in Theorem 3 is minimax optimal. Denote $\max(d_1,d_2)$ by d and define

$$\Psi(N,d,r,\rho):=\min\Big\{\frac{rd}{N},\rho\Big(\frac{d}{N}\Big)^{1-(q/2)},\rho^{2/q}\Big\}.$$

Under the restricted isometry condition, Rohde and Tsybakov (2011) gives a minimax lower bound on the statistical rate of recovering a low-rank matrix under trace regression:

$$\inf_{\widehat{\boldsymbol{\Theta}}}\sup_{\boldsymbol{\Theta}^*\in\mathcal{B}_q(\rho),\operatorname{rank}(\boldsymbol{\Theta}^*)\leq r}\mathbb{P}\big(\big\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\big\|_F^2\geq C\Psi(N,d,r,\rho)\big)\geq c>0,$$

where C is a constant independent of N, d, ρ and c is a universal constant. When r is very large or $q=0$, $\rho(d/N)^{1-q/2}$ becomes the dominant term in $\Psi(N,d,r,\rho)$. This matches the upper bound we derived in Theorem 3.

3.3. *Matrix completion.* In this section, we consider the matrix completion problem with heavy-tailed noise. Under a conventional setting, $\mathbf{X}_i$ is a singleton, $\|\boldsymbol{\Theta}^*\|_{\max}=O(1)$ and $\|\boldsymbol{\Theta}^*\|_F=O(\sqrt{d_1d_2})$. If we rescale the original model as

$$Y_i=\langle\mathbf{X}_i,\boldsymbol{\Theta}^*\rangle+\varepsilon_i=\langle\sqrt{d_1d_2}\mathbf{X}_i,\boldsymbol{\Theta}^*/\sqrt{d_1d_2}\rangle+\varepsilon_i$$

and treat $\sqrt{d_1d_2}\mathbf{X}_i$ as the new design $\check{\mathbf{X}}_i$ and $\boldsymbol{\Theta}^*/\sqrt{d_1d_2}$ as the new coefficient matrix $\check{\boldsymbol{\Theta}}^*$, then $\|\check{\mathbf{X}}_i\|_F=O(\sqrt{d_1d_2})$ and $\|\check{\boldsymbol{\Theta}}^*\|_F=O(1)$. Therefore, by rescaling, we can assume without loss of generality that $\boldsymbol{\Theta}^*$ satisfies $\|\boldsymbol{\Theta}^*\|_F\leq 1$ and $\mathbf{X}_i$ is uniformly sampled from $\{\sqrt{d_1d_2}\mathbf{e}_j\mathbf{e}_k^\top\}_{1\leq j\leq d_1,1\leq k\leq d_2}$.

In order to achieve consistent estimation of $\boldsymbol{\Theta}^*$, we require it not to be overly spiky, that is, $\|\boldsymbol{\Theta}^*\|_{\max}\leq R\|\boldsymbol{\Theta}^*\|_F/\sqrt{d_1d_2}\leq R/\sqrt{d_1d_2}$. We put this constraint in seeking the M-estimator of $\boldsymbol{\Theta}^*$:

$$\widehat{\boldsymbol{\Theta}}\in\operatorname*{argmin}_{\|\boldsymbol{\Theta}\|_{\max}\leq R/\sqrt{d_1d_2}}-\langle\widehat{\boldsymbol{\Sigma}}_{Y\mathbf{X}}(\tau),\boldsymbol{\Theta}\rangle+\frac{1}{2}\operatorname{vec}(\boldsymbol{\Theta})^\top\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\operatorname{vec}(\boldsymbol{\Theta})+\lambda_N\|\boldsymbol{\Theta}\|_N.\tag{3.7}$$

This spikiness condition is proposed by Negahban and Wainwright (2012) and it is required by the matrix completion problem per se instead of our robust estimation.

To derive robust estimation in matrix completion problem, we choose $\widetilde{Y}_i=\operatorname{sgn}(Y_i)(|Y_i|\wedge\tau)$ and $\widetilde{\mathbf{X}}_i=\mathbf{X}_i$ in (2.7). Then $\widehat{\boldsymbol{\Sigma}}_{Y\mathbf{X}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}$ are given by (3.5). Note that the design $\mathbf{X}_i$ here takes the singleton form, which leads to different scaling and consistency rates from the setting of matrix compressed sensing.

LEMMA 4. *Assume the following conditions hold*:

(C1) $\|\boldsymbol{\Theta}^*\|_F\leq 1$ *and* $\|\boldsymbol{\Theta}^*\|_{\max}\leq R/\sqrt{d_1d_2}$;

(C2) $\mathbf{X}_i$ *is uniformly sampled from* $\{\sqrt{d_1 d_2}\mathbf{e}_j\mathbf{e}_k^\top\}_{1\le j\le d_1, 1\le k\le d_2}$;
(C3) $\forall i = 1, \ldots, N$, $\mathbb{E}(\mathbb{E}(\varepsilon_i^2|\mathbf{X}_i)^k) \le M < \infty$, *where* $k > 1$;
(C4) *There exists a universal constant* $\gamma > 0$ *such that* $N > \gamma(d_1 \vee d_2)$.

Choose $\tau = \sqrt{(R^2 + M^{1/k})N/((d_1 \vee d_2)\log(d_1 + d_2))}$. *There exists* $C > 0$, *depending only on* γ, *such that for any* $\xi > 1$,

$$\mathbb{P}\Bigg(\Bigg\|\widehat{\mathbf{\Sigma}}_{Y\mathbf{X}}(\tau) - \frac{1}{N}\sum_{i=1}^{N}\langle\mathbf{X}_i, \mathbf{\Theta}^*\rangle\mathbf{X}_i\Bigg\|_{\mathrm{op}} > C\xi\sqrt{\frac{(R^2 + M^{1/k})(d_1 \vee d_2)\log(d_1 + d_2)}{N}}\Bigg) \le 4(d_1 + d_2)^{1-\xi}. \tag{3.8}$$

REMARK 7. Again, for $\overline{\mathbf{\Sigma}}_{Y\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N} Y_i\mathbf{X}_i$, Negahban and Wainwright (2012) proved that

$$\Bigg\|\overline{\mathbf{\Sigma}}_{Y\mathbf{X}} - \frac{1}{N}\sum_{i=1}^{N}\langle\mathbf{X}_i, \mathbf{\Theta}^*\rangle\mathbf{X}_i\Bigg\|_{\mathrm{op}} = O_{\mathbb{P}}\big(\sqrt{(d_1 + d_2)\log(d_1 + d_2)/N}\big)$$

for subexponential noise. Compared with this result, Lemma 4 achieves the same rate of convergence. By Jensen's inequality, condition (C3) is implied by $E\varepsilon_i^{2k} \le M < \infty$.

Now we present the theorem on the statistical error of $\widehat{\mathbf{\Theta}}$ defined in (3.7).

THEOREM 4. *Suppose that the conditions of Lemma 4 hold. Consider* $\mathcal{B}_q(\mathbf{\Theta}^*) \le \rho$ *with* $\|\mathbf{\Theta}^*\|_{\max}/\|\mathbf{\Theta}^*\|_F \le R/\sqrt{d_1 d_2}$. *For any* $\xi > 0$, *choose*

$$\tau = \sqrt{\frac{(R^2 + M^{1/k})N}{(d_1 \vee d_2)\log(d_1 + d_2)}}$$

and

$$\lambda_N = 2C\xi\sqrt{\frac{(R^2 + M^{1/k})(d_1 \vee d_2)\log(d_1 + d_2)}{N}}.$$

Under the same conditions as in Lemma 4, we have with probability at least $1 - 4(d_1 + d_2)^{1-\xi} - C_1\exp(-C_2(d_1 + d_2))$ *that*

$$\|\widehat{\mathbf{\Theta}}(\tau, \lambda_N) - \mathbf{\Theta}^*\|_F^2 \le C_3\max\Bigg\{\rho\Bigg(\frac{\xi(R^2 + M^{1/k})(d_1 \vee d_2)\log(d_1 + d_2)}{N}\Bigg)^{1-(q/2)}, \frac{R^2}{N}\Bigg\}$$

and

$$\|\widehat{\mathbf{\Theta}}(\tau, \lambda_N) - \mathbf{\Theta}^*\|_N \le C_4\max\Bigg\{\rho\Bigg(\frac{\xi(R^2 + M^{1/k})(d_1 \vee d_2)\log(d_1 + d_2)}{N}\Bigg)^{\frac{1-q}{2}}, \Bigg(\frac{\rho R^{2-2q}}{N^{1-q}}\Bigg)^{\frac{1}{2-q}}\Bigg\},$$

where $\{C_i\}_{i=1}^4$ *are universal constants.*

REMARK 8. We claim that the rate we derived in Theorem 4 is minimax optimal up to a logarithmic factor and a trailing term. Theorem 3 in Negahban and Wainwright (2012) shows that for matrix completion problems,

$$\inf_{\widehat{\boldsymbol{\Theta}}} \sup_{\mathcal{B}_q(\boldsymbol{\Theta})\le\rho} \mathbb{E}\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\|_F^2 \ge c\min\left\{\rho\left(\frac{d}{N}\right)^{1-q/2}, \frac{d^2}{N}\right\},$$

where c is some universal constant. As commented therein, as long as $\rho = O((d/N)^{q/2}d)$, $\rho(d/N)^{1-q/2}$ is the dominant term and this is what we established in Theorem 4 up to a logarithmic factor and a small trailing term.

REMARK 9. The spikiness condition that $\|\boldsymbol{\Theta}^*\|_{\max}/\|\boldsymbol{\Theta}^*\|_F \le R/\sqrt{d_1d_2}$ is an essential and nonremovable condition to obtain a sharp rate in matrix completion. Negahban and Wainwright (2012) have shown that when the true matrix is overly spiky, say a singleton, one needs to pay high sample complexity to accurately recover the matrix. Other works on matrix completion impose morally similar conditions. For instance, Koltchinskii, Lounici and Tsybakov (2011) and Minsker (2018) assume that $\|\boldsymbol{\Theta}^*\|_{\max}$ is bounded.

3.4. *Multitask learning*. Before presenting the theoretical results, we first simplify (2.6) under the setting of multitask regression. According to (2.8), (2.6) can be reduced to be

$$\widehat{\boldsymbol{\Theta}} \in \operatorname*{argmin}_{\boldsymbol{\Theta}\in\mathcal{S}} \frac{1}{d_2}\left(-\langle\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}}, \boldsymbol{\Theta}\rangle + \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{\Theta}^\top\widetilde{\mathbf{x}}_i\|_2^2\right) + \lambda_N\|\boldsymbol{\Theta}\|_N. \tag{3.9}$$

Recall here that n is the sample size in terms of (2.2) and $N = d_2n$. We also have that $\widehat{\boldsymbol{\Sigma}}_{Y\mathbf{X}} - \operatorname{mat}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\operatorname{vec}(\boldsymbol{\Theta}^*)) = (\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}} - \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}\boldsymbol{\Theta}^*)/d_2$.

Under the sub-Gaussian design, we only need to shrink the response vector $\mathbf{y}_i$. In (2.8), we choose $\widetilde{\mathbf{x}}_i = \mathbf{x}_i$ and $\widetilde{\mathbf{y}}_i = (\|\mathbf{y}_i\|_2 \wedge \tau)\mathbf{y}_i/\|\mathbf{y}_i\|_2$, where τ is some threshold. In words, we maintain the original design vectors, but shrink the Euclidean norm of the responses. Note that when $\mathbf{y}_i$ is one-dimensional, the shrinkage reduces to the truncation $y_i(\tau) = \operatorname{sgn}(y_i)(|y_i|\wedge\tau)$. When the design has only bounded moments, we need to shrink both the design vector $\mathbf{x}_i$ and response vector $\mathbf{y}_i$ by their ℓ_4 norm instead, that is, we choose $\widetilde{\mathbf{x}}_i = (\|\mathbf{x}_i\|_4 \wedge \tau_1)\mathbf{x}_i/\|\mathbf{x}_i\|_4$ and $\widetilde{\mathbf{y}}_i = (\|\mathbf{y}_i\|_4 \wedge \tau_2)\mathbf{y}_i/\|\mathbf{y}_i\|_4$, where τ_1 and τ_2 are two thresholds. Here, shrinking based on the fourth moment is uncommon, but it is crucial; it accelerates the convergence rate of the induced bias term so that it matches the desired statistical error rate. The details can be found in the proofs. Again, we will omit stating these choices in the following lemmas and theorems to ease presentation.

LEMMA 5. *Convergence rate of gradient of the robustified quadratic loss.*

(a) Sub-Gaussian design. *Assume the following conditions hold*:

(C1) $\lambda_{\max}(\mathbb{E}(\mathbf{y}_i\mathbf{y}_i^\top)) \le R < \infty$;

(C2) $\{\mathbf{x}_i\}_{i=1}^n$ *are i.i.d. sub-Gaussian vectors with* $\|\mathbf{x}_i\|_{\psi_2} \le \kappa_0 < \infty$, $\mathbb{E}\,\mathbf{x}_i = \mathbf{0}$ *and* $\lambda_{\min}(\mathbb{E}\,\mathbf{x}_i\mathbf{x}_i^\top) \ge \kappa_{\mathcal{L}} > 0$.

(C3) $\forall i = 1,\dots,n$, $j_1, j_2 = 1,\dots,d_2$ *and* $j_1 \ne j_2$, $\varepsilon_{ij_1} \perp\!\!\!\perp \varepsilon_{ij_2}|\mathbf{x}_i$, *and* $\forall j = 1,\dots,d_1$, $\mathbb{E}\{\mathbb{E}(\varepsilon_{ij}^2|\mathbf{x}_i)^k\} \le M < \infty$, *where* $k > 1$.

Choose $\tau \asymp_{k,\kappa_0} \sqrt{n(R+M^{1/k})/\log(d_1+d_2)}$. *Whenever* $(d_1+d_2)\log(d_1+d_2) < n$, *we have that for any* $\xi > 1$,

$$\mathbb{P}\Bigg(\Bigg\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\widetilde{\mathbf{y}}}(\tau) - \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{\Theta}^{*\top}\mathbf{x}_j\mathbf{x}_j^{\top}\Bigg\|_{\mathrm{op}} \geq C_1\xi\sqrt{\frac{(R+M^{1/k})(d_1+d_2)\log(d_1+d_2)}{n}}\Bigg) \leq 3(d_1+d_2)^{1-\xi},$$

where C_1 *depends only on* k, κ_0 *and* $\kappa_{\mathcal{L}}$.

(b) Bounded moment design. *Consider the following conditions*:

(C1′) $\forall \mathbf{v} \in \mathcal{S}^{d_1-1}$, $\mathbb{E}(\mathbf{v}^{\top}\mathbf{x}_i)^4 \leq \kappa_0 < \infty$ *and* $\lambda_{\min}(\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^{\top})) \geq \kappa_{\mathcal{L}} > 0$;
(C2′) $\forall \mathbf{v} \in \mathcal{S}^{d_2-1}$, $\mathbb{E}(\mathbf{v}^{\top}\mathbf{y}_i)^4 \leq M < \infty$.

Under Conditions (C1′) *and* (C2′), *for any* $\xi > 1$, *we have that*

$$\mathbb{P}\Bigg(\|\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}}(\tau_1,\tau_2) - \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}(\tau_1)\boldsymbol{\Theta}^*\|_{\mathrm{op}} \geq C_1 M^{1/4}\xi\sqrt{\frac{(d_1+d_2)\log(d_1+d_2)}{n}}\Bigg) \leq 4(d_1+d_2)^{1-\xi},$$

where $\tau_1 \asymp \{n\kappa_0/\log(d_1+d_2)\}^{1/4}$, $\tau_2 \asymp \{nM/\log(d_1+d_2)\}^{1/4}$ *and* C_1 *only depends on* κ_0 *and* $\kappa_{\mathcal{L}}$.

REMARK 10. When the noise and design are sub-Gaussian, Negahban and Wainwright (2011) used the covering argument to show that the regular sample covariance matrices $\overline{\boldsymbol{\Sigma}}_{\mathbf{xy}}$ and $\overline{\boldsymbol{\Sigma}}_{\mathbf{xx}}$ satisfy that

$$\|\overline{\boldsymbol{\Sigma}}_{\mathbf{xy}} - \overline{\boldsymbol{\Sigma}}_{\mathbf{xx}}\boldsymbol{\Theta}^*\|_{\mathrm{op}} = \Bigg\|\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{\varepsilon}_j\mathbf{x}_j^{\top}\Bigg\|_{\mathrm{op}} = O_{\mathbb{P}}\big(\sqrt{(d_1+d_2)/n}\big).$$

Lemma 5 shows that up to a logarithmic factor, the shrinkage sample covariance achieves nearly the same rate of convergence for noise and design with only bounded moments.

Finally, we establish the statistical error for the low-rank multitask learning.

THEOREM 5. *Assume* $\mathcal{B}_q(\boldsymbol{\Theta}^*) \leq \rho$.

(a) Sub-Gaussian design. *Suppose that Conditions* (C1), (C2) *and* (C3) *in Lemma 5 hold. For any* $\xi > 1$, *choose*

$$\tau \asymp_{k,\kappa_0} \sqrt{n\big(R+M^{1/k}\big)/\log(d_1+d_2)},$$

and

$$\lambda_N = \frac{2C_1\xi}{d_2}\sqrt{\frac{(R+M^{1/k})(d_1+d_2)\log(d_1+d_2)}{n}},$$

where C_1 *is the same as in part* (a) *of Lemma 5. There exist* C_2, C_3 *and* C_4, *depending only on* k, κ_0, $\kappa_{\mathcal{L}}$, *such that if* $n \geq C_2 d_1$, *then we have with probability at least* $1-3(d_1+d_2)^{1-\xi} - 2\exp(-cd_1)$ *that*

$$\|\widehat{\boldsymbol{\Theta}}(\tau,\lambda_N) - \boldsymbol{\Theta}^*\|_F^2 \leq C_2\rho\Bigg\{\frac{\xi^2(R+M^{1/k})(d_1+d_2)\log(d_1+d_2)}{n}\Bigg\}^{1-(q/2)}$$

and

$$\|\widehat{\mathbf{\Theta}}(\tau,\lambda_N)-\mathbf{\Theta}^*\|_N \le C_3\rho\left\{\frac{\xi^2(R+M^{1/k})(d_1+d_2)\log(d_1+d_2)}{n}\right\}^{(1-q)/2}.$$

(b) Bounded moment design. *Suppose instead that Conditions* (C1′) *and* (C2′) *in Lemma* 5 *hold. For any* $\xi>1$ *and* $\eta>0$, *choose*

$$\tau_1 \asymp \{n\kappa_0/\log(d_1+d_2)\}^{1/4}, \qquad \tau_2 \asymp \{nM/\log(d_1+d_2)\}^{1/4}$$

and

$$\lambda_N = \frac{2C_1M^{1/4}\xi}{d_2}\sqrt{\frac{(d_1+d_2)\log(d_1+d_2)}{n}},$$

where C_1 *is the same as in part* (b) *of Lemma* 5. *There exists* $\gamma>0$, *depending only on* $\kappa_{\mathcal{L}}$, *such that if* $\eta\kappa_0 d_1\log d_1/n<\gamma$, *then with probability at least* $1-3(d_1+d_2)^{1-\xi}-d_1^{1-C\eta}$,

$$\|\widehat{\mathbf{\Theta}}(\tau_1,\tau_2,\lambda_N)-\mathbf{\Theta}^*\|_F^2 \le C_3\rho\left(\frac{\xi^2\sqrt{M}(d_1+d_2)\log(d_1+d_2)}{n}\right)^{1-q/2}$$

and

$$\|\widehat{\mathbf{\Theta}}(\tau_1,\tau_2,\lambda_N)-\mathbf{\Theta}^*\|_N \le C_4\rho\left(\frac{\xi^2\sqrt{M}(d_1+d_2)\log(d_1+d_2)}{n}\right)^{(1-q)/2},$$

where C_3 *and* C_4 *are universal constants*.

REMARK 11. According to (C.47) in Fan, Wang and Zhu (2021), the multitask regression model satisfies the lower bound part of the RI condition in Rohde and Tsybakov (2011) with $\nu\asymp\sqrt{d_2}$. Substituting $\nu\asymp\sqrt{d_2}$, $\Delta=\rho^{1/q}/\nu$, $M=\max(d_1,d_2)$ and $N=nd_2$ into Theorem 5 in Rohde and Tsybakov (2011) yields that

$$\Psi(n,r,d_1,d_2,\rho)=\frac{1}{d_2}\min\left\{\frac{r\max(d_1,d_2)}{n},\rho\left(\frac{\max(d_1,d_2)}{n}\right)^{1-q/2},\rho^{2/q}\right\}.$$

Note that therein $C(\alpha,\nu)\asymp d_2$. Therefore, we have

$$\inf_{\widehat{\mathbf{\Theta}}}\sup_{\mathbf{\Theta}^*\in\mathcal{B}_q(\rho),\operatorname{rank}(\mathbf{\Theta}^*)\le r}\mathbb{P}\big(\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\|_F^2\ge Cd_2\Psi(n,r,d_1,d_2,\rho)\big)\ge c>0,$$

where C and c are constants. Under regular scaling assumptions, the dominant term in the minimax rate is $\rho(\max(d_1,d_2)/n)^{1-q/2}$, which matches our upper bound in Theorem 5 up to a logarithmic factor.

4. Robust covariance estimation. In multitask regression, the error bound derivation of $\|\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}}(\tau_1,\tau_2)-\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}(\tau_1)\mathbf{\Theta}^*\|_{\mathrm{op}}$ sheds light on applying the ℓ_4-norm shrinkage for robust covariance estimation. This topic is of its own interest, so we emphasize this interesting finding with a separate section. Here, we formulate the problem and the result, whose proof is relegated to Appendix C in the Supplementary Material (Fan, Wang and Zhu (2021)).

Suppose we have n i.i.d. d-dimensional random vectors $\{\mathbf{x}_i\}_{i=1}^n$ with $\mathbb{E}\,\mathbf{x}_i=\mathbf{0}$. Our goal is to estimate the covariance matrix $\mathbf{\Sigma}=\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)$ when the distribution of $\{\mathbf{x}_i\}_{i=1}^n$ has only bounded fourth moment. For any $\tau\in\mathbb{R}^+$, let $\widetilde{\mathbf{x}}_i:=(\|\mathbf{x}_i\|_4\wedge\tau)\mathbf{x}_i/\|\mathbf{x}_i\|_4$, where $\|\cdot\|_4$ is the ℓ_4-norm. We propose the following shrinkage sample covariance to estimate $\mathbf{\Sigma}$:

$$\widetilde{\mathbf{\Sigma}}_n(\tau)=\frac{1}{n}\sum_{i=1}^n\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top. \tag{4.1}$$

The following theorem establishes the statistical error rate of $\widetilde{\mathbf{\Sigma}}_n(\tau)$ with exponential deviation bound.

THEOREM 6. *Suppose* $\mathbb{E}(\mathbf{v}^\top \mathbf{x}_i)^4 \le R$ *for any* $\mathbf{v} \in \mathcal{S}^{d-1}$, *then it holds that for any* $\xi > 0$,

$$\mathbb{P}\left(\left\| \widetilde{\mathbf{\Sigma}}_n(\tau) - \mathbf{\Sigma} \right\|_{\mathrm{op}} \ge \sqrt{\frac{\xi R d \log d}{n}} \right) \le d^{1-C\xi}, \tag{4.2}$$

where $\tau \asymp \{nR/(\xi \log d)\}^{1/4}$ *and* C *is a universal constant.*

Below we also present a lower bound result, showing that our shrinkage sample covariance is minimax optimal up to a logarithmic factor.

THEOREM 7. *Define* $\mathbf{\Sigma}_{\mathbf{v}} := \mathbf{v}\mathbf{v}^\top + \mathbf{I}$. *Suppose* $\{\mathbf{x}_i\}_{i=1}^n$ *are* n *i.i.d.* d*-dimensional random vectors with mean zero and covariance* $\mathbf{\Sigma}_{\mathbf{v}}$. *When* $d \ge 34$, *it holds that*

$$\inf_{\widehat{\mathbf{\Sigma}}} \max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbb{P}\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{\mathbf{v}}\|_{\mathrm{op}} \ge \frac{1}{48}\sqrt{\frac{6d}{n}} \right) \ge \frac{1}{3}.$$

We have several comments for the proposed shrinkage sample covariance. First of all, to understand its behavior, we compare it with the sample covariance $\overline{\mathbf{\Sigma}}_n$ under the Gaussian data setting. Suppose $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Then we have $\|\mathbf{x}_i\|_4^4 = \sum_{j=1}^d x_{ij}^4 = O_{\mathbb{P}}(d)$. On the other hand, $\sup_{\mathbf{v}\in\mathcal{S}^{d-1}} \mathbb{E}(\mathbf{v}^\top \mathbf{x}_i)^4 = 3$, and thus $\tau^4 \asymp n/\log d$. Therefore, depending on whether n is greater or smaller than $d \log d$, one expects either all the vectors or none of them to be shrunk, and the shrinkage sample covariance to be either equal to the sample covariance or close to a multiple of it, with scaling of order $\tau/d^{1/4}$. In other words, in the low-dimensional regime, there is no need to shrink Gaussian random vectors for covariance estimation, while in the high-dimensional regime, we need to shrink the sample covariance matrix toward zero.

Note that $\widetilde{\mathbf{\Sigma}}_n(\tau)$ outperforms the sample covariance $\overline{\mathbf{\Sigma}}_n$ even with Gaussian samples when $\mathrm{tr}(\mathbf{\Sigma}) \asymp d\|\mathbf{\Sigma}\|_{\mathrm{op}}$ and d is large. According to Koltchinskii and Lounici (2017) or Vershynin ((2010), Theorem 5.39), when $\mathrm{tr}(\mathbf{\Sigma}) \asymp d\|\mathbf{\Sigma}\|_{\mathrm{op}}$,

$$\|\overline{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_{\mathrm{op}} \asymp \|\mathbf{\Sigma}\|_{\mathrm{op}} \max\left(\sqrt{\frac{d}{n}}, \frac{d}{n} \right).$$

When d/n is large, the d/n term will dominate $\sqrt{d/n}$, thus delivering statistical error of order d/n for Gaussian sample covariance. However, our shrinkage sample covariance always attains the rate of $\sqrt{d \log d/n}$ regardless of relationship between the dimension and the sample size. Theorem 7 shows that the minimax optimal rate of estimating the covariance in terms of the operator norm is $\sqrt{d/n}$. Hence, the shrinkage sample covariance is minimax optimal up to a logarithmic factor, whereas traditional sample covariance is not in high dimensions. Therefore, shrinkage overcomes not only heavy-tailed corruption, but also curse of dimensionality. In Section 5.4, we conduct simulations to further illustrate this point.

If we are concerned with an error bound in the elementwise max norm, we need to naturally apply elementwise truncation rather than the ℓ_4-norm shrinkage. Define $\check{\mathbf{x}}_i$ such that $\check{x}_{ij} = \mathrm{sgn}(x_{ij})(|x_{ij}| \wedge \tau)$ for $1 \le j \le d_1$ and $\check{\mathbf{\Sigma}}_n(\tau) = n^{-1}\sum_{i=1}^n \check{\mathbf{x}}_i \check{\mathbf{x}}_i^\top$. It is not hard to derive that with $\tau \asymp (n/\log d)^{\frac{1}{4}}$, $\|\check{\mathbf{\Sigma}}_n(\tau) - \mathbf{\Sigma}\|_{\max} = O_{\mathbb{P}}(\sqrt{\log d/n})$ as in Fan, Li and Wang (2017). Note that the elementwise-truncated sample covariance is not necessarily positive semidefinite. Besides, with this choice of τ, $\check{\mathbf{\Sigma}}_n(\tau)$ will equal to $\overline{\mathbf{\Sigma}}_n$ with high probability when data are Gaussian. To see this, again suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The maximum of $|x_{ij}|$ would be sharply concentrated around a term of order $\sqrt{\log(nd)}$. Therefore, if $n \ge \log^2(nd)\log(d)$, with high probability the threshold would not be reached. Hence,

$\check{\mathbf{x}}_i = \mathbf{x}_i$ for all i and $\check{\mathbf{\Sigma}}_n = \overline{\mathbf{\Sigma}}_n$. Further regularization can be applied to $\check{\mathbf{\Sigma}}_n$ if the true covariance is sparse. See, for example, Meinshausen and Bühlmann (2006), Bickel and Levina (2008), Lam and Fan (2009), Cai and Liu (2011), Cai and Zhou (2012), Fan, Liao and Mincheva (2013), among others.

Our method is designed to robustify tail behavior rather than ε-contamination in the adversarial learning, where an ε-contamination model allows an adversary to arbitrarily corrupt ε-fraction of the observations. To illustrate the difference between these two perspectives of robustness, below we compare our setup and results with those of Diakonikolas et al. (2016) and Loh and Tan (2018); both of them consider robust estimation of high-dimensional covariance matrices under an ε-contamination model.

- Model setup: Neither Loh and Tan (2018) or Diakonikolas et al. (2016) imposes assumptions on the distribution of contamination. Therefore, under the ε-contamination model, our bounded moment conditions are not satisfied, and the truncation or shrinkage approach might be sub-optimal. On the other hand, both Diakonikolas et al. (2016) and Loh and Tan (2018) assume the original uncorrupted data to be Gaussian to ensure the required concentration results, which will not hold if these data are heavy-tailed as in our Theorem 6.
- Methodology: To obtain robust estimation of the covariance matrix, Loh and Tan (2018) use the median absolute deviation (MAD) to estimate the standard deviation, and use Kendall's tau and Spearman's rho to estimate the correlation. Note that the consistency of these estimators relies on the Gaussian assumption of the original uncorrupted data. Without Gaussianity, the desired statistical rate as shown in (11a), (11b), (12a) and (12b) of Loh and Tan (2018) might not be achieved. Diakonikolas et al. (2016) develop an approximated separation oracle for the set of the covariance estimates that satisfy a desired error guarantee, and propose to use the ellipsoid method to find an element of this set. Such a separation oracle incurs higher computational cost than the simple truncation or shrinkage, and it is well known that the ellipsoid method converges slowly in practice.
- Error guarantee: Both Diakonikolas et al. (2016) and Loh and Tan (2018) are particularly concerned with the dependence of the statistical error on the proportion of the corrupted data ε. Their error guarantees scale (nearly) linearly with ε. In contrast, our goal is to achieve the minimax rate that was established under the light-tail setup for heavy-tailed data. Besides, in terms of covariance estimation, Diakonikolas et al. (2016) and Loh and Tan (2018) study the Frobenius and max-norm error respectively, while we focus on the operator norm.

Finally, we notice a recent independent work by Minsker (2018) that studies the mean of random matrices with bounded moments. There, the proposed estimator admits sub-Gaussian or subexponential concentration around the unknown mean in terms of the operator norm. Applying their general results to the covariance estimation setting achieves the same statistical error rate as we obtain here under bounded fourth-moment conditions of $\mathbf{x}_i$'s.

5. Simulation study. In this section, we mainly compare the numerical performance of our shrinkage procedure with that of the standard procedure under various problem setups. In linear regression, we also investigate two alternative robust approaches: the ℓ_1-regularized Huber approach and the median of means approach, whose details are given in the sequel. We do not investigate the nuclear-norm regularized Huber approach for the matrix estimation setups, because so far we are not aware of any algorithm that solves the corresponding optimization problem with rigorous convergence guarantee.

As for the data distribution, in each of the three matrix problems, we investigate three noise settings: log-normal noise, Cauchy noise and Gaussian noise. They represent heavy-tailed asymmetric distributions, extremely heavy-tailed symmetric distributions and light-tailed symmetric distributions respectively. We set the design to be sub-Gaussian in matrix

compressed sensing and multitask regression. In linear regression, we investigate both light-tailed and heavy-tailed designs, while we only investigate heavy-tailed noise there for conciseness.

Now we elucidate the tuning procedure in our simulations. We set $\lambda = C_\lambda f(N, d)$ and $\tau = C_\tau g(N, d)$, where C_λ and C_τ are constants that do not depend on N and d, and where f and g follow the rates we derived for each problem setup. We exhaustively tune C_τ and C_λ in the setup of the minimal sample size for best estimation error, and maintain these constants as N and d increase. Through this, we can validate the established rates of λ and τ. This tuning procedure is feasible in the linear regression setup, given that the computational cost is relatively cheap. However, similar exhaustive search is computationally intimidating in the matrix setups. To address this, we propose the robust cross-validation (RCV) procedure as described below:

1. Tune λ with K-fold robust CV using the original data ($\tau = \infty$). Here, in order to guard against heavy-tailed corruption in CV, we truncate the original response by its η-quantile when calculating the CV score on a validation set. Note that η is only applied to the validation set and should be chosen a priori and independently with τ. In all of our simulations, we choose $\eta = 0.95$ and $K = 5$.
2. Choose λ by the famous one-standard-error rule to improve sparsity. We calculate the standard error (SE) of the K CV scores for each λ. We pick the λ corresponding to the minimal mean CV score, and then increase λ until the mean CV score hits the minimal mean CV score plus its corresponding standard error.
3. Tune τ with K-fold robust cross-validation with the chosen λ. The robust CV scores as above are calculated for each validation set and for each τ.
4. Choose τ by the one-standard-error rule but this time to improve robustness. We first find the τ with the minimal mean CV score, and then decrease τ until the mean CV score hits the minimal mean CV score plus its corresponding one standard error.

In a nutshell, RCV first tunes λ with $\tau = \infty$ and then tunes τ with the fixed λ, thus avoiding heavy computation in the exhaustive grid tuning. It is also robust in the sense that we clip an a priori percentage of data in the validation set to ensure that the CV scores are insensitive to extreme values. We show in the subsequent sections that RCV yields satisfactory numerical results in all of our matrix problem setups.

The main message is that with appropriate truncation or shrinkage, our robust procedure outperforms the standard one under the setting with heavy-tailed noise, and it performs equally well as the standard procedure under the Gaussian noise. In other words, our procedure is adaptive to the tail of noise. The simulations are based on 100 independent Monte Carlo experiments. Besides presenting the numerical results, we also elaborate the optimization algorithms we exploited, which might be of interest to readers concerned with implementation.

Last but not least, we compare the numerical performance of the regular sample covariance and the shrinkage sample covariance as proposed in (4.1) in estimating the true covariance. Simulation results show superiority of the shrinkage sample covariance over the regular sample covariance under both Gaussian noise and t_3 noise. Therefore, the shrinkage not only overcomes the heavy-tailed corruption, but also mitigates the curse of high dimensionality.

5.1. *Linear model.* In this section, we focus on a high-dimensional sparse linear model:

$$Y = \mathbf{x}^\top \boldsymbol{\theta}^* + \varepsilon, \tag{5.1}$$

where $\mathbf{x}$ is valued in $\mathbb{R}^{100}$ ($d = 100$) and has i.i.d. entries, where $\boldsymbol{\theta}^* = (1, 1, 1, 0, \ldots, 0)^\top$, and where ε is independent of $\mathbf{x}$. We investigate two designs: a light-tailed design, where

$x_j \sim_{\text{i.i.d.}} \mathcal{N}(0, 0.25)$ for $j = 1, \ldots, 100$, and a heavy-tailed design, where $x_j \sim_{\text{i.i.d.}} t_{2.1}/\sqrt{21}$. We assume that $\varepsilon \sim t_{2.1}$. Besides comparing the performance of LASSO on original data and truncated data, we also assess the performance of ℓ_1-regularized Huber loss minimization and the median-of-means approach. In the following, we elucidate how we implement these four methods and choose the tuning parameters.

1. LASSO with truncated data: We solve

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} (\widetilde{Y}_i - \widetilde{\mathbf{x}}_i^\top \boldsymbol{\theta})^2 + \lambda \|\boldsymbol{\theta}\|_1,$$

where the definition of $\widetilde{Y}_i$ and $\widetilde{\mathbf{x}}_i$ follow Section 3.1. The rate of the thresholds in terms of N and d follow Lemma 1.

2. Huber loss minimization with ℓ_1-regularization: Define the Huber loss

$$h_\tau(x) := \begin{cases} -\tau(x+\tau) + \tau^2/2, & x < -\tau, \\ x^2/2, & |x| \le \tau, \\ \tau(x-\tau) + \tau^2/2, & x > \tau. \end{cases}$$

We solve the following optimization problem:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} h_\tau(Y_i - \boldsymbol{\theta}^\top \mathbf{x}_i) + \lambda \|\boldsymbol{\theta}\|_1,$$

where $\tau \asymp \sqrt{N/\log d}$ and $\lambda \asymp \sqrt{\log d/N}$, as established in Sun, Zhou and Fan ((2020), Theorem 3). One recent work Wang et al. (2018) proposed a principle for tuning-free Huber regression. Nevertheless, to be fair in assessing the approaches, here we stick with the exhaustive search for tuning λ and τ.

3. Median of means: Randomly and evenly divide the whole dataset into K subsets $\{\mathcal{D}_k\}_{k=1}^K$ and calculate a standard LASSO estimator $\widehat{\boldsymbol{\theta}}^{(k)}$ on each subdataset $\mathcal{D}_k$. Then we take the geometric median of $\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K$ as the final estimator, which is mathematically defined as

$$\widehat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{k=1}^{K} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^{(k)}\|_2.$$

K is chosen to be of order $\log N$.

4. Standard LASSO:

$$\widehat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2 + \lambda \|\boldsymbol{\theta}\|_1,$$

where $\lambda \asymp \sqrt{\log d/N}$.

We let N grow geometrically from 100 to 1000 and compare $\log(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2)$ using four different approaches based on 500 independent Monte Carlo simulations. The constants before the rates of the tuning parameters are tuned for optimal performance.

From Figure 2, one can see that (i) under the light-tailed design, our approach and the Huber approach perform similarly, and they significantly outperform the other two; (ii) under the heavy-tailed design, our approach achieves the best performance among the four methods, in particular when the sample size is large. Note that the median of means and standard method have exactly the same performance because the optimal number of sample splits K turns out to be always one in the simulation.

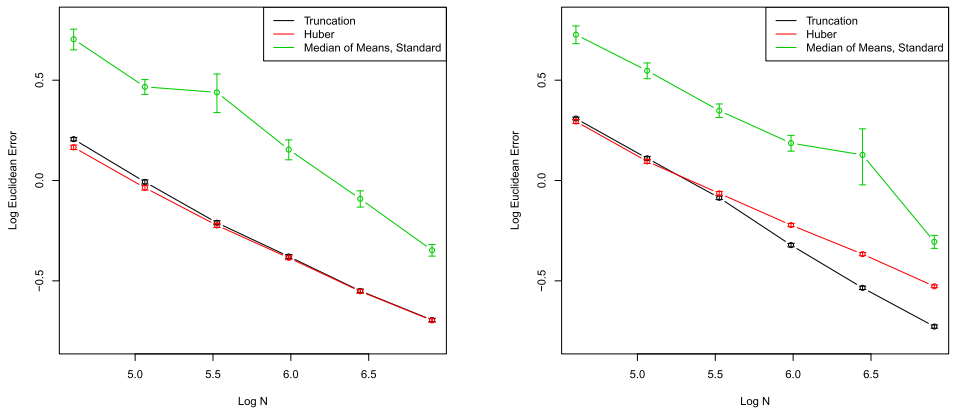


FIG. 2. *The logarithm of Euclidean statistical errors with standard error bars. The left panel corresponds to the light-tailed design, while the right panel corresponds to the heavy-tailed design. Two ends of the error bar correspond to* $\log(\mathrm{MEAN}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2) - \mathrm{SE}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2))$ *and* $\log(\mathrm{MEAN}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2) + \mathrm{SE}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2))$.

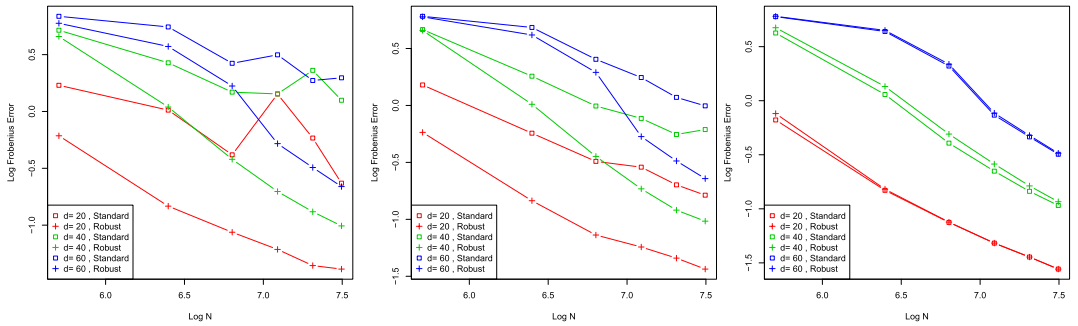
An interesting setup that was suggested by one referee is the case where the design is heavy-tailed, but the noise is light-tailed. Intuitively, clipping large values of design there should induce bias and worsen the estimation. This intuition is further confirmed by our simulation. We still assume (5.1), but this time we set $x_j \sim_{\text{i.i.d.}} t_{2.1}$ and $\varepsilon \sim N(0, 4)$. Let $N = 1000$. We use RCV to tune λ and the truncation level τ for the design. Given that the noise is sub-Gaussian, there is no need to clip the response; we thus set $\eta = 1$ in RCV. It turns out that in all the 500 independent Monte Carlo experiments, RCV chooses not to truncate any value on the design. This implies that truncation on the design is unnecessary in this setup, and that RCV is adaptive to the tail of the design.

Finally, we compare our truncation method and the standard method when the design has collinearity. The take-away message is that truncation of the response still dramatically improves the statistical performance when the condition number of the population covariance is large. To save the space, we relegate the details to Appendix A in the Supplementary Material (Fan, Wang and Zhu (2021)).

5.2. *Compressed sensing.* We first specify the parameters in the true model: $Y_i = \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle + \varepsilon_i$. In the simulation, we set $d_1 = d_2 = d$ and construct $\boldsymbol{\Theta}^*$ to be $\sum_{i=1}^{5} \mathbf{v}_i \mathbf{v}_i^\top$, where $\mathbf{v}_i$ is the ith top eigenvector of the sample covariance of 100 i.i.d. centered isotropic Gaussian vectors, so that $\mathrm{rank}(\boldsymbol{\Theta}^*) = 5$ and $\|\boldsymbol{\Theta}^*\|_F \approx \sqrt{5}$. The design matrix $\mathbf{X}_i$ has i.i.d. standard Gaussian entries. The noise distributions are characterized as follows:

- Log-normal: $\varepsilon_i \overset{d}{=} (Z - \mathbb{E}\, Z)/1000$, where $\ln Z \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 = 9$;
- Cauchy: $\varepsilon_i \overset{d}{=} Z/64$, where Z follows Cauchy distribution;
- Gaussian: $\varepsilon_i \sim N(0, \sigma^2)$, where $\sigma^2 = 0.25$.

The constants above are chosen to ensure appropriate signal-to-noise ratio for better presentation. We let N grow from 300 to 1800 and choose $d = 20, 40, 60$. We present the numerical results in Figure 3. As we can observe from the plots, the robust estimator has much smaller statistical error than the standard estimator under the heavy-tailed log-normal and Cauchy noises. Also, robust procedures deliver sharper estimation as the sample size increases, while the standard procedure does not necessarily do so under the heavy-tailed noise. Under Gaussian noise, the robust estimator has nearly the same statistical performance as the standard one, so it does not hurt to apply the truncation under the light-tailed noise setting. The details



Log-normal noise Cauchy noise Gaussian noise

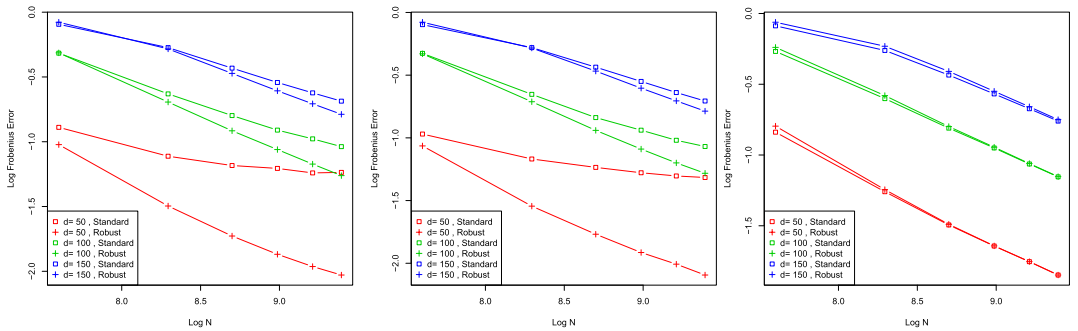
FIG. 3. *Statistical errors of* $\ln\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ *v.s. logarithmic sample size* $\ln N$ *for different dimensions* d *in matrix compressed sensing.*

of our implementation algorithm can be found in Appendix B the Supplementary Material (Fan, Wang and Zhu (2021)).

5.3. *Matrix completion.* We again set $d_1 = d_2 = d$ and construct $\mathbf{\Theta}^*$ as $\sum_{i=1}^5 \mathbf{v}_i \mathbf{v}_i^\top / \sqrt{5}$, where $\mathbf{v}_i$ is the ith top eigenvector of the sample covariance of 100 i.i.d. centered Gaussian random vectors with identity covariance. Each design matrix $\mathbf{X}_i$ takes the singleton form, which is uniformly sampled from $\{\mathbf{e}_j \mathbf{e}_k^\top\}_{1 \le j,k \le d}$. The three noise distributions considered are identical to those in the previous section.

We let N grow from 2000 to 12,000 and choose $d = 50, 100, 150$. Again, we used the robust cross-validation for tuning and only tune for the minimal sample size $N = 2000$. The numerical results are presented in Figure 4. Analogous to the matrix compressed sensing, we can observe from the figure that compared with the standard procedure, the robust procedure has significantly smaller statistical error in estimating $\mathbf{\Theta}^*$ under the log-normal and Cauchy noise and nearly the same performance under the Gaussian noise. The implementation algorithm is again given in Appendix B in the Supplementary Material (Fan, Wang and Zhu (2021)).

5.4. *Multitask regression.* We choose $d_1 = d_2 = d$ and construct $\mathbf{\Theta}^*$ as in Section 5.2. The design vectors $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The noise distributions are characterized as follows:



Log-normal noise Cauchy noise Gaussian noise

FIG. 4. *Statistical errors of* $\ln\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ *v.s. logarithmic sample size* $\ln N$ *for different dimensions* d *in matrix completion.*

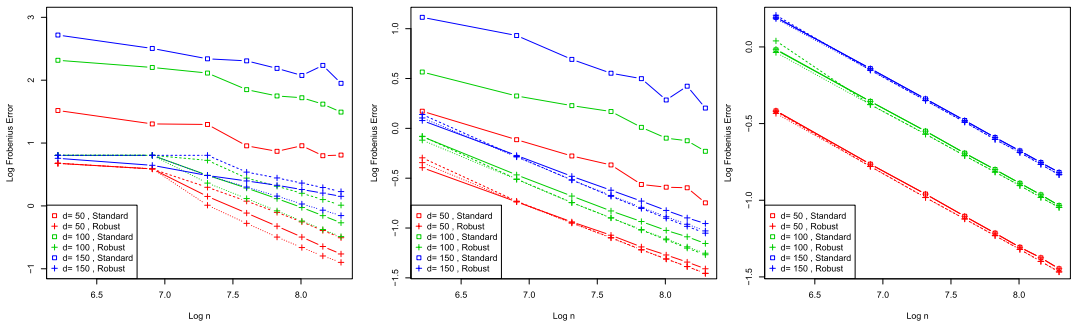


FIG. 5. *Statistical errors of* $\ln\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ *versus logarithmic sample size* $\ln N$ *for different dimensions* d *in multitask regression. For the robust method, the solid lines correspond to* $\tau = \tau_{\mathrm{RCV}}$, *the dashed lines correspond to* $\tau = 0.8\tau_{\mathrm{RCV}}$ *and the dotted lines correspond to* $\tau = 1.2\tau_{\mathrm{RCV}}$.

- Log-normal: $\varepsilon_i \overset{d}{=} (Z - \mathbb{E}\, Z)/500$, where $\ln Z \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 = 7.84$;
- Cauchy: $\varepsilon_i \overset{d}{=} Z/128$, where Z follows Cauchy distribution;
- Gaussian: $\varepsilon_i \sim N(0, \sigma^2)$, where $\sigma^2 = 0.25$.

In this simulation, we choose n from 500 to 4000 and $d = 50, 100, 150$. We present the numerical performance of our robust method and the standard method in Figure 5. Similar to the two examples before, the robust procedure tuned by RCV yields sharper accuracy in estimating $\mathbf{\Theta}^*$ under the two heavy-tailed noises, while maintaining competitive under the Gaussian noise. In addition, we show the statistical error of our robust approach with $\tau = 0.8\tau_{\mathrm{RCV}}$ and $\tau = 1.2\tau_{\mathrm{RCV}}$. As one can observe, even if the shrinkage threshold is perturbed by 20%, the statistical performance of the robust method remains nearly the same, which demonstrates that our method is not sensitive to the shrinkage threshold. Please refer to Appendix B in the Supplementary Material (Fan, Wang and Zhu (2021)) for the detailed algorithm for solving multitask learning.

5.5. *Covariance estimation.* In this subsection, we investigate the statistical error of the shrinkage sample covariance $\widetilde{\mathbf{\Sigma}}_n(\tau)$ proposed in Section 4, compared with the regular sample covariance $\overline{\mathbf{\Sigma}}_n$. We consider two common distributions: Gaussian and Student's t_3 random samples. The dimension is set to be proportional to sample size, that is, $d/n = \alpha$ with α being 0.2, 0.5, 1. The sample size n will range from 100 to 500 for each case. Regardless of how large the dimension d is, the true covariance $\mathbf{\Sigma}$ is always set to be a diagonal matrix with the first diagonal element equal to 4 and all the other diagonal elements equal to 1. The statistical errors are measured in terms of the operator norm, and our simulation is based on 100 independent Monte Carlo replications.

Unlike the supervised learning problems discussed above, covariance estimation does not involve response. Therefore, RCV is not applicable here. In the following, we investigate the statistical performance of the proposed ℓ_4-norm shrinkage sample covariance under different levels of shrinkage. We define the "shrinkage ratio" η as follows:

$$\eta := 1 - \frac{1}{n}\sum_{i=1}^{n} 1 \wedge (\tau/\|\mathbf{x}_i\|_4).$$

One can see that a smaller η implies less shrinkage, and that $\eta = 0$ implies no shrinkage at all. For $n = 100$ and $\alpha = 1$, we choose C in $\tau = C(n/\log d)^{1/4}$ such that $\eta = 0.2$, and use the same C for other combinations of n and α. The results are presented in Figure 6. We

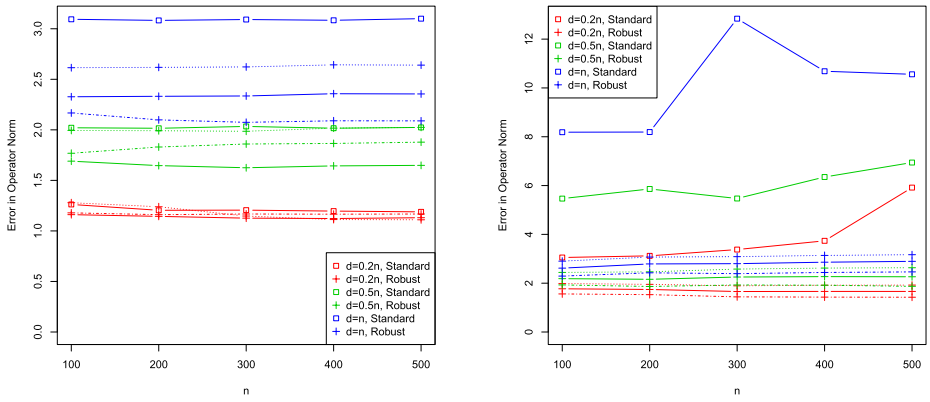


FIG. 6. $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}$ *v.s. n with different dimensions d in covariance estimation. For the robust methods, the solid lines correspond to the shrinkage ratio* $\eta = 0.2$*, while we also report results for* $\eta = 0.3$ (*dotted lines*) *and* $\eta = 0.1$ (*dot–dashed lines*) *for sensitivity assessment.*

also plot the results for $\eta = 0.3$ (dotted lines) and $\eta = 0.1$ (dash–dotted lines) for sensitivity assessment.

As we can see, for Gaussian samples, as long as we fix d/n, the statistical error of both $\overline{\boldsymbol{\Sigma}}_n$ and $\widetilde{\boldsymbol{\Sigma}}_n(\tau)$ does not change, which is consistent with Theorem 5.39 in Vershynin (2010) and Theorem 6 in our paper. Also, the higher the dimension is, the more significant the superiority of $\widetilde{\boldsymbol{\Sigma}}_n(\tau)$ is over $\overline{\boldsymbol{\Sigma}}_n$. This validates our remark after Theorem 6 that the shrinkage ameliorates the curse of dimensionality. Even for Gaussian data, shrinkage is meaningful and provides significant improvement.

For t_3 distribution, since it is heavy-tailed, the regular sample covariance does not maintain constant statistical error for a fixed d/n; instead the error increases as the sample size increases. In contrast, our shrinkage sample covariance still retains consistent performance and enjoys much higher accuracy. This strongly supports the sharp statistical error rate we derived for $\widetilde{\boldsymbol{\Sigma}}_n(\tau)$ in Theorem 6. In addition, the performance of the shrinkage sample covariance is not sensitive to the choice of the tuning parameter.

6. Discussion. In this paper, we proposed a general shrinkage principle for trace regression with heavy-tailed data. We show that in low-rank matrix recovery, *truncating properly* your heavy-tailed data before solving empirical risk minimization *works as if* you have sub-Gaussian data to start with. We investigated four setups: linear regression, matrix compressed sensing, matrix completion and multitask learning, where we may apply elementwise, ℓ_2-norm or ℓ_4-norm truncation depending on whether design or noise is heavy-tailed or not. This shrinkage principle is modular and does not require any algorithmic adaptation, and the resulting estimator is shown to achieve the (nearly) optimal statistical rate.

Note that different problem setups have different requirements for the shrinkage principle to take effect. One interesting observation is that the heavy-tailed design typically requires stronger assumptions than the sub-Gaussian design. Specifically, in linear regression, we required bounded $\|\boldsymbol{\theta}^*\|_1$ and a stronger scaling condition under heavy-tailed design. It could be interesting to see if those conditions can be further relaxed.

We hypothesized that other robust approaches as investigated in Section 5.1 may be able to achieve the same statistical rate as our proposed method. Nevertheless, our shrinkage method is clearly among the most computationally feasible approaches. As for tuning of truncation parameters, we proposed the RCV procedure, which performed well in our simulation studies. Although the results demonstrated the great potential of our method, we would like to

remind our readers that in real data analysis, truncation and shrinkage should be combined with prior knowledge, useful data transformation, and proper normalization to handle heterogeneity across different features.

Acknowledgments. Since its availability in arxiv.org in 2015, the paper has gone through many revisions, thanks to various useful comments by referees, AE and Editors.

Funding. The research was supported by NSF Grants DMS-1662139, DMS-1712591, DMS-2015366 and NIH Grant R01-GM072611-14.

SUPPLEMENTARY MATERIAL

Supplement to "A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery" (DOI: 10.1214/20-AOS1980SUPP; .pdf). The supplement includes additional simulation results, detailed description of the numerical algorithms and technical proof.

REFERENCES

BELLEC, P. C., DALALYAN, A. S., GRAPPIN, E. and PARIS, Q. (2018). On the prediction loss of the lasso in the partially labeled setting. *Electron. J. Stat.* **12** 3443–3472. MR3864589 https://doi.org/10.1214/18-EJS1457

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann. Statist.* **43** 2507–2536. MR3405602 https://doi.org/10.1214/15-AOS1350

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949 https://doi.org/10.1198/jasa.2011.tm10560

CAI, T. T. and ZHANG, A. (2014). Sparse representation of a polytope and recovery in sparse signals and low-rank matrices. *IEEE Trans. Inf. Theory* **60** 122–132. MR3150915 https://doi.org/10.1109/TIT.2013.2288639

CAI, T. T. and ZHANG, A. (2015). ROP: Matrix recovery via rank-one projections. *Ann. Statist.* **43** 102–138. MR3285602 https://doi.org/10.1214/14-AOS1267

CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607 https://doi.org/10.1214/12-AOS998

CANDES, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris* **346** 589–592. MR2412803 https://doi.org/10.1016/j.crma.2008.03.014

CANDES, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

CANDES, E. J. and PLAN, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory* **57** 2342–2359. MR2809094 https://doi.org/10.1109/TIT.2011.2111771

CANDES, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

CANDES, E. J. and TAO, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory* **52** 5406–5425. MR2300700 https://doi.org/10.1109/TIT.2006.885507

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43** 129–159. MR1854649 https://doi.org/10.1137/S003614450037906X

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proc. Natl. Acad. Sci. USA* **116** 22931–22937. MR4036123 https://doi.org/10.1073/pnas.1910053116

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 J. Optim.* **30** 3098–3121. MR4167625 https://doi.org/10.1137/19M1290000

DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2016). Robust estimators in high dimensions without the computational intractability. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 655–664. IEEE Computer Soc., Los Alamitos, CA. MR3631028 https://doi.org/10.1109/FOCS.2016.85

DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inf. Theory* **52** 1289–1306. MR2241189 https://doi.org/10.1109/TIT.2006.871582

DONOHO, D. L., JOHNSTONE, I. and MONTANARI, A. (2013). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inf. Theory* **59** 3396–3433. MR3061255 https://doi.org/10.1109/TIT.2013.2239356

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972 https://doi.org/10.1111/rssb.12166

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653 https://doi.org/10.1111/rssb.12016

FAN, J., LIU, H. and WANG, W. (2018). Large covariance estimation through elliptical factor models. *Ann. Statist.* **46** 1383–1414. MR3819104 https://doi.org/10.1214/17-AOS1588

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

FAN, J., WANG, W. and ZHU, Z. (2021). Supplement to "A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery." https://doi.org/10.1214/20-AOS1980SUPP

FAN, J. and YAO, Q. (2015). *Elements of Financial Econometrics*. Science Press, Beijing.

FIGUEIREDO, M. A. T., NOWAK, R. D. and WRIGHT, S. J. (2007). Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signal Process.* **1** 586–597.

GUPTA, S., ELLIS, S. E., ASHAR, F. N., MOES, A., BADER, J. S., ZHAN, J., WEST, A. B. and ARKING, D. E. (2014). Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5** 5748. https://doi.org/10.1038/ncomms6748

HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40. MR3491112

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL mapping. *Ann. Appl. Stat.* **6** 1095–1117. MR3012522 https://doi.org/10.1214/12-AOAS549

KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459 https://doi.org/10.1214/09-AOS720

LOH, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403 https://doi.org/10.1214/16-AOS1471

LOH, P.-L. and TAN, X. L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ε-contamination. *Electron. J. Stat.* **12** 1429–1467. MR3804842 https://doi.org/10.1214/18-EJS1427

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468 https://doi.org/10.3150/14-BEJ645

MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348 https://doi.org/10.1214/10-AOS850

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697. MR2930649

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133 https://doi.org/10.1214/12-STS400

NEMIROVSKY, A. S. and YUDIN, D. B. (1982). *Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience Publication*. Wiley, New York. MR0702836

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

RECHT, B. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12** 3413–3430. MR2877360

RECHT, B., FAZEL, M. and PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501. MR2680543 https://doi.org/10.1137/070697835

REINSEL, G. C. and VELU, R. P. (2013). *Multivariate Reduced-Rank Regression*: *Theory and Applications*. *Lecture Notes in Statistics* **136**. Springer, New York. MR1719704 https://doi.org/10.1007/978-1-4757-2853-8

ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342 https://doi.org/10.1214/10-AOS860

RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inf. Theory* **59** 3434–3447. MR3061256 https://doi.org/10.1109/TIT.2013.2243201

STOCK, J. H. and WATSON, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econom. Statist.* **20** 147–162. MR1963257 https://doi.org/10.1198/073500102317351921

SUN, Q., ZHOU, W.-X. and FAN, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265. MR4078461 https://doi.org/10.1080/01621459.2018.1543124

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Preprint. Available at arXiv:1011.3027.

WANG, L., ZHENG, C., ZHOU, W. and ZHOU, W.-X. (2018). A new principle for tuning-free Huber regression. Preprint.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802