

# A Shrinkage Principle for Heavy-Tailed Data: High-Dimensional Robust Low-Rank Matrix Recovery\*

Jianqing Fan, Weichen Wang, Ziwei Zhu

Department of Operations Research and Financial Engineering

Princeton University.

## Abstract

This paper introduces a simple principle for robust high-dimensional statistical inference via an appropriate shrinkage on the data. This widens the scope of high-dimensional techniques, reducing the moment conditions from sub-exponential or sub-Gaussian distributions to merely bounded second or fourth moment. As an illustration of this principle, we focus on robust estimation of the low-rank matrix  $\Theta^*$  from the trace regression model  $Y = \text{Tr}(\Theta^{*T} X) + \epsilon$ . It encompasses four popular problems: sparse linear models, compressed sensing, matrix completion and multi-task regression. We propose to apply penalized least-squares approach to 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 models and multi-tasking regression, we further allow the design to have only bounded fourth moment and obtain the same statistical rates, again, by appropriate shrinkage of the design matrix. As a byproduct, we give a robust covariance matrix estimator and establish its concentration inequality in terms of the spectral norm when the random samples have only bounded fourth moment. Extensive simulations have been carried out to support our theories.

**Keywords:** Robust Statistics, Shrinkage, Heavy-Tailed Data, Trace Regression, Low-Rank Matrix Recovery, High-Dimensional Statistics.

---

\*The research was supported by NSF grants DMS-1206464 and DMS-1406266 and NIH grant R01-GM072611-12.

# 1 Introduction

Heavy-tailed distributions are ubiquitous in modern statistical analysis and machine learning problems. They are stylized features of high-dimensional data. By chance alone, some of observable variables in high-dimensional datasets can have heavy or moderately heavy tails (see right panel of Figure 1). It has been widely known that financial returns and macroeconomic variables exhibit heavy tails, and large-scale imaging datasets in biological studies are often corrupted by heavy-tailed noises due to limited measurement precisions. Figure 1 provides some empirical evidence on this which is pandemic to high-dimensional data. These stylized features and phenomena contradict the popular assumption of sub-Gaussian or sub-exponential noises in the theoretical analysis of standard statistical procedures. They also have adverse impacts on the methods that are popularly used. Simple and effective principles are needed for dealing with moderately heavy or heavy tailed data.

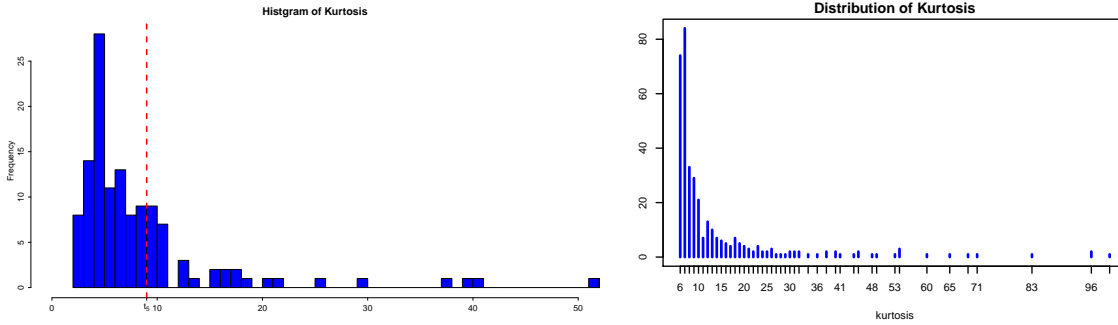


Figure 1: **Distributions of kurtosis of macroeconomic variables and gene expressions.** Red dashline marks variables with empirical kurtosis equals to that of  $t_5$ -distribution. Left panel: For 131 macroeconomics variables in Stock and Watson (2002). Right panel: For logarithm of expression profiles of 383 genes based on RNA-seq for autism data (Gupta et al., 2014), whose kurtosis is bigger than that of  $t_5$  among 19122 genes.

Recent years have witnessed increasing literature on the robust mean estimation when the population distribution is heavy-tailed. Catoni (2012) proposed a novel approach that is through minimizing a robust empirical loss. Unlike the traditional  $\ell_2$  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.

Brownlee et al. (2015) pursued the Catoni’s mean estimator further by applying it to empirical risk minimization. Fan et al. (2016) 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 (2015) 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 et al. (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 sub-samples 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 called the trace regression

$$Y = \text{Tr}(\Theta^{*T} \mathbf{X}) + \epsilon,$$

which embraces linear regression, matrix or vector compressed sensing, matrix completion and multi-tasking regression as specific examples. The goal is to estimate the coefficient matrix  $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ , which is assumed to have a nearly low-rank structure in the sense that its Schatten norm is constrained:  $\sum_{i=1}^{\min(d_1, d_2)} \sigma_i(\Theta^*)^q \leq \rho$  for  $0 \leq q < 1$ , where  $\sigma_i(\Theta^*)$  is the  $i^{th}$  singular value of  $\Theta^*$ , i.e., the square-root of the  $i^{th}$  eigenvalue of  $\Theta^{*T} \Theta^*$ . In other words, the singular values of  $\Theta^*$  decay fast enough so that  $\Theta^*$  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  $\Theta^*$  with only bounded moment conditions on noise and design.

As the most simple and important example of 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 et al. (2001), Fan and Li (2001), Donoho (2006), Candes and Tao (2006), Candes and Tao (2007), Candes (2008), Nowak et al. (2007), Fan and Lv (2008), Zou and Li (2008), Bickel et al. (2009), Zhang (2010), Negahban and Wainwright (2012), Donoho et al. (2013). These pioneering works 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 non-asymptotic 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 non-asymptotic 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 et al. (2010), Candes and Plan (2011), Cai and Zhang (2014), Cai and Zhang (2015), etc. When the true low-rank matrix  $\Theta^*$  satisfies certain restricted isometry property (RIP) or similar properties, this approach can exactly recover  $\Theta^*$  under the noiseless setting and enjoy sharp statistical error rate with sub-Gaussian and sub-exponential noise.

There has also been great amount of work on matrix completion. Candès 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 Candès 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.

Our work aims to handle the presence of heavy-tailed, asymmetrical and heteroscedastic noises in the general trace regression. Based on the shrinkage of data, we developed a new loss function called the robust quadratic loss, which is constructed by plugging robust covariance estimators in the  $\ell_2$  risk function. Then we obtain the estimator  $\hat{\Theta}$  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 the matrix  $\Theta^*$  that are the same as those in Negahban et al. (2012) for the sub-Gaussian distributions, while allowing the noise and design to have much heavier tails. This result is very generic and applicable to all four specific aforementioned examples.

Our robust approach is particularly simple: it truncates or shrinks appropriately the response variables, depending on whether the responses are univariate or multivariate. Under the setting of sub-Gaussian design, unusually 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. Under the setting of heavy-tailed covariates, 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. With these data robustifications, we can then apply 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 the sharp spectral-norm convergence rate of robust covariance matrices based on data shrinkage. Of course, other robustifications of estimated covariance matrices, such as the RA-covariance estimation in Fan et al. (2016), are also possible to enjoy similar statistical error rates, but we will only focus on the shrinkage method, as it is easier to analyze and always semi-positive definite.

It is worth emphasis that the successful application of the shrinkage sample covariance in multi-tasking regression inspires us to also 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  $\sup_{\mathbf{v} \in \mathcal{S}^{d-1}} \mathbb{E}(\mathbf{v}^T \mathbf{x}_i)^4 \leq R < \infty$ , where  $\mathcal{S}^{d-1}$  is the  $d$ -dimensional unit sphere, our  $\ell_4$ -norm shrinkage sample covariance  $\tilde{\Sigma}_n$  achieves the statistical error rate of order  $O_P(\sqrt{d \log d/n})$  in terms of the spectral norm. This rate is the same, up to a logarithmic term, as that of the standard sample covariance matrix  $\bar{\Sigma}_n$  with sub-Gaussian samples under the low-dimensional regime. Under the high-dimensional regime,  $\tilde{\Sigma}$  even outperforms  $\bar{\Sigma}_n$  for sub-Gaussian random samples, since now the error rate of  $\bar{\Sigma}_n$  deteriorates to  $O_P(d/n)$  while the error rate of  $\tilde{\Sigma}$  is still  $O_P(\sqrt{d \log d/n})$ . This means even with light-tailed data, standard sample covariance can be inadmissible in terms of convergence rate when dimension is high. 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 sample covariance with statistical error rate of order  $O_P(\sqrt{\log d/n})$ . This estimator can further be regularized if the true covariance has sparsity and other 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 et al. (2013), among others.

The 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 multi-tasking regression. Then we develop the generalized  $\ell_2$  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 on the robust covariance inputs to ensure the statistical error rate of the M-estimator. Then we apply this theorem to all the four specific aforementioned problems and derive explicitly the statistical error rate of our M-estimators. Section 4 derives the statistical error of the shrinkage covariance estimator in terms of the spectral norm. Finally we present simulation studies in Section 5, which demonstrate the advantage of our robust estimator over the standard one. The associated optimization algorithms are also discussed there. All the proofs are relegated to the Appendix A.

## 2 Models and methodology

We first collect the general notation before formulating the model and methodology.

### 2.1 Generic Notations

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  $\ell_q$  norm; specifically,  $\|\mathbf{x}\|_1$  and  $\|\mathbf{x}\|_2$  denote the  $\ell_1$  norm and  $\ell_2$  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  $\|\mathbf{X}\|_{op}$ ,  $\|\mathbf{X}\|_N$ ,  $\|\mathbf{X}\|_F$  and  $\|\mathbf{X}\|_{\max}$  to be its operator norm, nuclear norm, Frobenius norm and elementwise max norm respectively. We use  $\text{vec}(\mathbf{X})$  to denote vectorized version of  $\mathbf{X}$ , i.e.,  $\text{vec}(\mathbf{X}) = (\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_{d_2}^T)^T$ , where  $\mathbf{X}_j$  is

the  $j^{\text{th}}$  column of  $\mathbf{X}$ . Conversely, for a vector  $\mathbf{x} \in \mathbb{R}^{d_1 d_2}$ , we use  $\text{mat}(\mathbf{x})$  to denote the  $d_1$ -by- $d_2$  matrix constructed by  $\mathbf{x}$ , where  $(x_{(j-1)d_1+1}, \dots, x_{jd_1})^T$  is the  $j^{\text{th}}$  column of  $\text{mat}(\mathbf{x})$ . For any two matrices  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$ , define the inner product  $\langle \mathbf{A}, \mathbf{B} \rangle := \text{Tr}(\mathbf{A}^T \mathbf{B})$  where  $\text{Tr}$  is the trace operator. We denote  $\text{diag}(\mathbf{M}_1, \dots, \mathbf{M}_n)$  to be the block diagonal matrix with the diagonal blocks as  $\mathbf{M}_1, \dots, \mathbf{M}_n$ . 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  $a_n \asymp b_n$  if there exist constants  $0 < c_1 < c_2$  such that  $c_1 a_n \leq b_n \leq c_2 a_n$  for  $1 \leq n < \infty$ . For a random variable  $X$ , define its sub-Gaussian norm  $\|X\|_{\psi_2} := \sup_{p \geq 1} (\mathbb{E} |X|^p)^{\frac{1}{p}} / \sqrt{p}$  and its sub-exponential norm  $\|X\|_{\psi_1} := \sup_{p \geq 1} (\mathbb{E} |X|^p)^{\frac{1}{p}} / p$ . For a random vector  $\mathbf{x} \in \mathbb{R}^d$ , we define its sub-Gaussian norm  $\|\mathbf{x}\|_{\psi_2} := \sup_{\mathbf{v} \in \mathcal{S}^{d-1}} \|\mathbf{v}^T \mathbf{x}\|_{\psi_2}$  and sub-exponential norm  $\|\mathbf{x}\|_{\psi_1} := \sup_{\mathbf{v} \in \mathcal{S}^{d-1}} \|\mathbf{v}^T \mathbf{x}\|_{\psi_1}$ . Given  $x, y \in \mathbb{R}$ , we 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  $j^{\text{th}}$  element 1 and other elements 0.

## 2.2 Trace Regression

In this paper, we consider the trace regression, a general model that encompasses the linear regression, compressed sensing, matrix completion, multi-tasking regression, etc. 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

$$Y_i = \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle + \epsilon_i, \quad (2.1)$$

where  $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$  is the true coefficient matrix,  $\mathbb{E} \mathbf{X}_i = \mathbf{0}$  and  $\{\epsilon_i\}_{i=1}^N$  are independent noises satisfying  $\mathbb{E}(\epsilon_i | \mathbf{X}_i) = 0$ . Note that here we do not assume  $\{\mathbf{X}_i\}_{i=1}^N$  are independent to each other nor assume  $\epsilon_i$  is independent to  $\mathbf{X}_i$ . Model (2.1) includes the following specific cases.

- **Linear regression:**  $d_1 = d_2 = d$ , and  $\{\mathbf{X}_i\}_{i=1}^N$  and  $\boldsymbol{\Theta}^*$  are diagonal. Let  $\mathbf{x}_i$  and  $\boldsymbol{\theta}^*$  denote the vectors of diagonal elements of  $\mathbf{X}_i$  and  $\boldsymbol{\Theta}^*$  respectively in the context of linear regression, i.e.,  $\mathbf{x}_i = \text{diag}(\mathbf{X}_i)$  and  $\boldsymbol{\theta}^* = \text{diag}(\boldsymbol{\Theta}^*)$ . Then, (2.1) reduces to familiar linear model:  $Y_i = \mathbf{x}_i^T \boldsymbol{\theta} + \epsilon_i$ . Having a low-rank  $\boldsymbol{\Theta}^*$  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  $\mathbf{X}$  and  $\Theta^*$  as diagonal matrices.
- **Matrix completion:**  $\mathbf{X}_i$  is a singleton, i.e.,  $\mathbf{X}_i = \mathbf{e}_{j(i)} \mathbf{e}_{k(i)}^T$  for  $1 \leq j(i) \leq d_1$  and  $1 \leq k(i) \leq d_2$ . In other words, a random entry of the matrix  $\Theta$  is observed along with noise for each sample.
- **Multi-tasking regression:** The multi-tasking (reduced-rank) regression model is

$$\mathbf{y}_j = \Theta^{*T} \mathbf{x}_j + \epsilon_j, \quad j = 1, \dots, n, \quad (2.2)$$

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,  $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$  is the coefficient matrix and  $\epsilon_j \in \mathbb{R}^{d_2}$  is the noise with each entry independent to each other. See, for example, Kim and Xing (2010) and Velu and Reinsel (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), i.e.,  $\{(Y_{(j-1)d_2+i} = y_{ji}, \mathbf{X}_{(j-1)d_2+i} = \mathbf{x}_j \mathbf{e}_i^T)\}_{i=1}^{d_2}$ . 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  $\Theta^*$ . 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  $\Theta^*$  is diagonal. The rank constraint reduces the effective number of parameters in  $\Theta^*$  and arises frequently in many applications. Consider the Netflix problem for instance, where  $\Theta_{ij}^*$  is the intrinsic score of film  $j$  given by customer  $i$  and we would like to recover the entire  $\Theta^*$  with only partial observations. Given that movies of similar types or qualities should receive similar scores from viewers, columns of  $\Theta^*$  should share colinearity, thus delivering a low-rank structure of  $\Theta^*$ . The rationale of the model can also be understood from the celebrated factor model in finance and econometrics (Fan and Yao, 2015), which assumes that several market risk factors drive the returns of a large panel of stocks. Consider  $N \times T$  returns  $\mathbf{Y}$  of  $N$  stocks (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) with a loading matrix  $\mathbf{B}$  ( $N \times K$  matrix), where  $K$  is much smaller than  $N$  or  $T$ . The factor model admits the



following form:

$$\mathbf{Y} = \mathbf{B}\mathbf{F} + \mathbf{E}$$

where  $\mathbf{E}$  is idiosyncratic noise. Since  $\mathbf{B}\mathbf{F}$  has a small rank  $K$ ,  $\mathbf{B}\mathbf{F}$  can be regarded as the low-rank matrix  $\Theta^*$  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.

Being exactly low-rank is still too stringent to model the real-world situations. We instead consider  $\Theta^*$  satisfying

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

where  $0 \leq q \leq 1$ . Note that when  $q = 0$ , the constraint (2.3) is an exact rank constraint. Restriction on  $\mathcal{B}_q(\Theta^*)$  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  $\rho$  to grow with dimensionality and sample size.

A popular method for estimating  $\Theta^*$  is the penalized empirical loss that solves  $\hat{\Theta} \in \operatorname{argmin}_{\Theta \in \mathcal{S}} \mathcal{L}(\Theta) + \lambda_N \mathcal{P}(\Theta)$ , where  $\mathcal{S}$  is a convex set in  $\mathbb{R}^{d_1 \times d_2}$ ,  $\mathcal{L}(\Theta)$  is a loss function,  $\lambda_N$  is a tuning parameter and  $\mathcal{P}(\Theta)$  is a rank penalization function. Most of the previous work, e.g., Koltchinskii et al. (2011) and Negahban and Wainwright (2011), chose  $\mathcal{L}(\Theta) = \sum_{1 \leq i \leq N} (Y_i - \langle \Theta, \mathbf{X}_i \rangle)^2$  and  $\mathcal{P}(\Theta) = \|\Theta\|_N$ , and derived the rate for  $\|\hat{\Theta} - \Theta^*\|_F$  under the assumption of sub-Gaussian or sub-exponential noise. However, the  $\ell_2$  loss is sensitive to outliers and is unable to handle the data with moderately heavy or heavy tails.

### 2.3 Robustifying $\ell_2$ Loss

We aim to accomodate heavy-tailed noise and design for the nearly low-rank matrix recovery by robustifying the traditional  $\ell_2$  loss. We first notice that the  $\ell_2$  risk can be expressed as

$$\begin{aligned} R(\Theta) &= \mathbb{E} \mathcal{L}(\Theta) = \mathbb{E} (Y_i - \langle \Theta, \mathbf{X}_i \rangle)^2 \\ &= \mathbb{E} Y_i^2 - 2\langle \Theta, \mathbb{E} Y_i \mathbf{X}_i \rangle + \operatorname{vec}(\Theta)^T \mathbb{E} (\operatorname{vec}(\mathbf{X}_i) \operatorname{vec}(\mathbf{X}_i)^T) \operatorname{vec}(\Theta) \\ &\equiv \mathbb{E} Y_i^2 - 2\langle \Theta, \Sigma_{Y\mathbf{X}} \rangle + \operatorname{vec}(\Theta)^T \Sigma_{\mathbf{X}\mathbf{X}} \operatorname{vec}(\Theta). \end{aligned} \quad (2.4)$$

Ignoring  $E Y_i^2$ , if we substitute  $\Sigma_{Y\mathbf{X}}$  and  $\Sigma_{\mathbf{X}\mathbf{X}}$  by their corresponding sample covariances, we obtain the empirical  $\ell_2$  loss. This inspires us to define a generalized  $\ell_2$  loss via

$$\mathcal{L}(\Theta) = -\langle \hat{\Sigma}_{Y\mathbf{X}}, \Theta \rangle + \frac{1}{2} \text{vec}(\Theta)^T \hat{\Sigma}_{\mathbf{X}\mathbf{X}} \text{vec}(\Theta), \quad (2.5)$$

where  $\hat{\Sigma}_{Y\mathbf{X}}$  and  $\hat{\Sigma}_{\mathbf{X}\mathbf{X}}$  are estimators of  $E Y_i \mathbf{X}_i$  and  $E \text{vec}(\mathbf{X}_i) \text{vec}(\mathbf{X}_i)^T$  respectively.

In this paper, we study the following M-estimator of  $\Theta^*$  with the generalized  $\ell_2$  loss:

$$\hat{\Theta} \in \underset{\Theta \in \mathcal{S}}{\text{argmin}} -\langle \hat{\Sigma}_{Y\mathbf{X}}, \Theta \rangle + \frac{1}{2} \text{vec}(\Theta)^T \hat{\Sigma}_{\mathbf{X}\mathbf{X}} \text{vec}(\Theta) + \lambda_N \|\Theta\|_N, \quad (2.6)$$

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  $\hat{\Sigma}_{Y\mathbf{X}}$  and  $\hat{\Sigma}_{\mathbf{X}\mathbf{X}}$ . 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 used 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  $\hat{\Sigma}_{Y\mathbf{X}}$  and  $\hat{\Sigma}_{\mathbf{X}\mathbf{X}}$ , 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,  $\hat{\Sigma}_{Y\mathbf{X}}$  and  $\hat{\Sigma}_{\mathbf{X}\mathbf{X}}$  take the following forms:

$$\hat{\Sigma}_{Y\mathbf{X}} = \hat{\Sigma}_{\tilde{Y}\tilde{\mathbf{X}}} = \frac{1}{N} \sum_{i=1}^N \tilde{Y}_i \tilde{\mathbf{X}}_i \quad \text{and} \quad \hat{\Sigma}_{\mathbf{X}\mathbf{X}} = \hat{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} = \frac{1}{N} \sum_{i=1}^N \text{vec}(\tilde{\mathbf{X}}_i) \text{vec}(\tilde{\mathbf{X}}_i)^T, \quad (2.7)$$

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. Note that this construction of generalized quadratic loss is equivalent to truncating the data first and then employing the usual quadratic loss.

For the multi-tasking 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

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

where

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

and  $\widetilde{\mathbf{y}}_i$  and  $\widetilde{\mathbf{x}}_i$  are again transformed versions of  $\mathbf{y}_i$  and  $\mathbf{x}_i$ . The tilde means shrinkage for heavy-tailed variables and means 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  $nd_2$  samples in (2.1). As we shall see, under only bounded moment assumptions of the design and noise, the generalized  $\ell_2$  loss equipped with the proposed robust covariance estimators yields a sharp  $\mathcal{M}$ -estimator  $\widehat{\Theta}$ , whose statistical error rates match those established in Negahban and Wainwright (2011) and Negahban and Wainwright (2012) under the setting of sub-Gaussian design and noise.

### 3 Main results

Our goal is to derive the statistical error rate of  $\widehat{\Theta}$  defined by (2.6). In the theoretical results of this section, 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 error  $\|\widehat{\Theta} - \Theta^*\|_F$ .

**Theorem 1.** *Define  $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$ , where  $\Theta^*$  satisfies  $\mathcal{B}_q(\Theta^*) \leq \rho$ . Suppose  $\text{vec}(\widehat{\Delta})^T \widehat{\Sigma}_{XX} \text{vec}(\widehat{\Delta}) \geq \kappa_{\mathcal{L}} \|\widehat{\Delta}\|_F^2$ , where  $\kappa_{\mathcal{L}}$  is a positive constant that does not depend on  $\widehat{\Delta}$ . Choose  $\lambda_N \geq 2\|\widehat{\Sigma}_{YX} - \text{mat}(\widehat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op}$ . Then we have for some constants  $C_1$  and  $C_2$ ,*

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

First of all, the above result is deterministic and nonasymptotic. As we can see from the theorem above, the statistical performance of  $\widehat{\Theta}$  relies on the restricted eigenvalue (RE)

property of  $\widehat{\Sigma}_{\mathbf{X}\mathbf{X}}$ , which was first studied by Bickel et al. (2009). When the design is sub-Gaussian, we choose  $\widehat{\Sigma}_{\mathbf{X}\mathbf{X}}$  to be the traditional sample covariance, whose RE property has been well established (e.g., Rudelson and Zhou (2013), Negahban and Wainwright (2011) and Negahban and Wainwright (2012)). We will specify these results when we need them in the sequel. When the design only satisfies bounded moment conditions, we choose  $\widehat{\Sigma}_{\mathbf{X}\mathbf{X}} = \widehat{\Sigma}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}}$ , i.e., the sample covariance of shrunk  $\mathbf{X}$ . We show that with appropriate level of shrinkage,  $\widehat{\Sigma}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}}$  still retains the RE property, thus satisfying the conditions of the theorem.

Secondly, the conclusion of the theorem says that  $\|\widehat{\Delta}\|_F^2$  and  $\|\widehat{\Delta}\|_N$  are proportional to  $\lambda_N^{2-q}$  and  $\lambda_N^{1-q}$  respectively, but we require  $\lambda_N \geq \|\widehat{\Sigma}_{Y\mathbf{X}} - \text{mat}(\widehat{\Sigma}_{\mathbf{X}\mathbf{X}}\text{vec}(\Theta^*))\|_{op}$ . This implies that the rate of  $\|\widehat{\Sigma}_{Y\mathbf{X}} - \text{mat}(\widehat{\Sigma}_{\mathbf{X}\mathbf{X}}\text{vec}(\Theta^*))\|_{op}$  is crucial to the statistical error of  $\widehat{\Theta}$ . In the following subsections, we will derive the rate of  $\|\widehat{\Sigma}_{Y\mathbf{X}} - \text{mat}(\widehat{\Sigma}_{\mathbf{X}\mathbf{X}}\text{vec}(\Theta^*))\|_{op}$  for all the aforementioned specific problems with only bounded moment conditions on the response and 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.

### 3.1 Linear Model

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

$$\widehat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\text{argmin}} -\widehat{\Sigma}_{Y\mathbf{X}}^T \theta + \frac{1}{2} \theta^T \widehat{\Sigma}_{\mathbf{X}\mathbf{X}} \theta + \lambda_N \|\theta\|_1, \quad (3.1)$$

where  $\widehat{\Sigma}_{Y\mathbf{X}} = \widehat{\Sigma}_{\widetilde{Y}\widetilde{\mathbf{X}}} = N^{-1} \sum_{i=1}^N \widetilde{Y}_i \widetilde{\mathbf{x}}_i$ ,  $\widehat{\Sigma}_{\mathbf{X}\mathbf{X}} = \widehat{\Sigma}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}} = N^{-1} \sum_{i=1}^N \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^T$ . When the design is sub-Gaussian, we only need to truncate the response. Therefore, we choose  $\widetilde{Y}_i = \widetilde{Y}_i(\tau) = \text{sgn}(Y_i)(|Y_i| \wedge \tau)$  and  $\widetilde{\mathbf{x}}_i = \mathbf{x}_i$ , for some threshold  $\tau$ . When the design is heavy-tailed, we choose  $\widetilde{Y}_i(\tau) = \text{sgn}(Y_i)(|Y_i| \wedge \tau_1)$  and  $\widetilde{x}_{ij} = \text{sgn}(x_{ij})(|x_{ij}| \wedge \tau_2)$ , where  $\tau_1$  and  $\tau_2$  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  $\widehat{\theta}$  in (3.1), in the following lemma, we derive the rate of  $\|\widehat{\Sigma}_{Y\mathbf{X}} - \text{mat}(\widehat{\Sigma}_{\mathbf{X}\mathbf{X}}\text{vec}(\Theta^*))\|_{op}$  in (2.6) for the sub-Gaussian design and bounded-

moment (polynomial tail) design respectively. Note here that

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

**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  $\|\mathbf{x}_i\|_{\psi_2} \leq \kappa_0 < \infty$ ,  $\mathbb{E} \mathbf{x}_i = \mathbf{0}$  and  $\lambda_{\min}(\mathbb{E} \mathbf{x}_1 \mathbf{x}_1^T) \geq \kappa_{\mathcal{L}} > 0$ ;  
(C2)  $\forall i = 1, \dots, N$ ,  $\mathbb{E} |Y_i|^{2k} \leq M < \infty$  for some  $k > 1$ .

Choose  $\tau \asymp \sqrt{N/\log d}$ . For any  $\delta > 0$ , there exists a constant  $\gamma_1 > 0$  such that as long as  $\log d/N < \gamma_1$ , we have

$$P\left(\|\widehat{\Sigma}_{Y\mathbf{x}}(\tau) - \widehat{\Sigma}_{\mathbf{x}\mathbf{x}} \boldsymbol{\theta}^*\|_{\max} \geq \nu_1 \sqrt{\frac{\delta \log d}{N}}\right) \leq 2d^{1-\delta}, \quad (3.2)$$

where  $\nu_1$  is a universal constant.

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

- (C1')  $\|\boldsymbol{\theta}^*\|_1 \leq R < \infty$ ;  
(C2')  $\mathbb{E} |x_{ij_1} x_{ij_2}|^2 \leq M < \infty$ ,  $1 \leq j_1, j_2 \leq d$ ;  
(C3')  $\forall i = 1, \dots, N$ ,  $\mathbb{E} |Y_i|^4 \leq M < \infty$ .

Choose  $\tau_1, \tau_2 \asymp (N/\log d)^{\frac{1}{4}}$ . For any  $\delta > 0$ , it holds that

$$P\left(\|\widehat{\Sigma}_{Y\mathbf{x}}(\tau_1, \tau_2) - \widehat{\Sigma}_{\mathbf{x}\mathbf{x}}(\tau_2) \boldsymbol{\theta}^*\|_{\max} > \nu_2 \sqrt{\frac{\delta \log d}{N}}\right) \leq 2d^{1-\delta},$$

where  $\nu_2$  is a universal constant.

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

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

This is the same rate as what we achieved under only the bounded moment conditions on response and design.

Next we establish the restricted strong convexity of the proposed robust  $\ell_2$  loss.

**Lemma 2.** *Restricted strong convexity.*

(a) **Sub-Gaussian design.** *Under Condition (C1) of Lemma 1, it holds for certain constants  $C_1, C_2$  and any  $\eta_1 > 1$  that*

$$P\left(\mathbf{v}^T \widehat{\Sigma}_{\mathbf{xx}} \mathbf{v} \geq \frac{1}{2} \mathbf{v}^T \Sigma_{\mathbf{xx}} \mathbf{v} - \frac{C_1 \eta_1 \log d}{N} \|\mathbf{v}\|_1^2, \quad \forall \mathbf{v} \in \mathbb{R}^d\right) \geq 1 - \frac{d^{1-\eta_1}}{3} - 2d \exp(-C_2 N). \quad (3.3)$$

(b) **Bounded moment design.** *If  $\mathbf{x}_i$  satisfies Condition (C2') of Lemma 1, then it holds for some constant  $C_3 > 0$  and any  $\eta_2 > 2$  that*

$$P\left(\mathbf{v}^T \widehat{\Sigma}_{\mathbf{xx}}(\tau_2) \mathbf{v} \geq \mathbf{v}^T \Sigma_{\mathbf{xx}} \mathbf{v} - C_3 \eta_2 \sqrt{\frac{\log d}{N}} \|\mathbf{v}\|_1^2, \quad \forall \mathbf{v} \in \mathbb{R}^d\right) \leq d^{2-\eta_2}, \quad (3.4)$$

as long as  $\tau_2 \asymp (N/\log d)^{\frac{1}{4}}$ .

**Remark 2.** Comparing the results we get for sub-Gaussian design and heavy-tailed design, we can find that the coefficients before  $\|\mathbf{v}\|_1^2$  are different. Under the sub-Gaussian design, that coefficient is of the order  $\log d/N$ , while under the heavy-tailed design, the coefficient is of order  $\sqrt{\log d/N}$ . This difference will lead to different scaling requirements for  $N, d$  and  $\rho$  in the sequel. As we shall see, the heavy-tailed design requires stronger scaling conditions to retain the same statistical error rate as the sub-Gaussian design for the linear model.

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

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

(a) **Sub-Gaussian design:** Suppose Conditions (C1) and (C2) in Lemma 1 hold. For any  $\delta > 0$ , choose  $\tau \asymp \sqrt{N/\log d}$  and  $\lambda_N = 2\nu_1 \sqrt{\delta \log d/N}$ , where  $\nu_1$  and  $\delta$  are the same as in part (a) of Lemma 1. There exist positive constants  $\{C_i\}_{i=1}^3$  such that as long as  $\rho(\log d/N)^{1-\frac{q}{2}} \leq C_1$ , it holds that

$$P\left(\|\widehat{\boldsymbol{\theta}}(\tau, \lambda_N) - \boldsymbol{\theta}^*\|_2^2 > C_2 \rho \left(\frac{\delta \log d}{N}\right)^{1-\frac{q}{2}}\right) \leq 3d^{1-\delta}$$

and

$$P\left(\|\widehat{\boldsymbol{\theta}}(\tau, \lambda_N) - \boldsymbol{\theta}^*\|_1 > C_3 \rho \left(\frac{\delta \log d}{N}\right)^{\frac{1-q}{2}}\right) \leq 3d^{1-\delta}.$$

(b) **Bounded moment design:** For any  $\delta > 0$  choose  $\tau_1, \tau_2 \asymp (N/\log d)^{\frac{1}{4}}$  and  $\lambda_N = 2\nu_2\sqrt{\delta \log d/N}$ , where  $\nu_2$  and  $\delta$  are the same as in part (b) of Lemma 1. Under Conditions (C1'), (C2') and (C3'), there exist constants  $\{C_i\}_{i=4}^6$  such that as long as  $\rho(\log d/N)^{\frac{1-q}{2}} \leq C_4$ , we have

$$P\left(\|\widehat{\boldsymbol{\theta}}(\tau_1, \tau_2, \lambda_N) - \boldsymbol{\theta}^*\|_2^2 > C_5 \rho \left(\frac{\delta \log d}{N}\right)^{1-\frac{q}{2}}\right) \leq 3d^{1-\delta}$$

and

$$P\left(\|\widehat{\boldsymbol{\theta}}(\tau_1, \tau_2, \lambda_N) - \boldsymbol{\theta}^*\|_1 > C_6 \rho \left(\frac{\delta \log d}{N}\right)^{\frac{1-q}{2}}\right) \leq 3d^{1-\delta}.$$

**Remark 3.** Under both sub-Gaussian and heavy-tailed design, our proposed  $\widehat{\boldsymbol{\theta}}$  achieves the minimax optimal rate of  $\ell_2$  norm established by Raskutti et al. (2011). However, the difference lies in the scaling requirement on  $N$ ,  $d$  and  $\rho$ . For sub-Gaussian design, we require  $\rho(\log d/N)^{1-\frac{q}{2}} \leq C_1$ , whereas for heavy-tailed design we need  $\rho(\log d/N)^{\frac{1-q}{2}} \leq C_4$ . Under the 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 rate.

## 3.2 Matrix Compressed Sensing

For the matrix compressed sensing problem, since the design is chosen by users, we consider solely the most popular design: the Gaussian design. We thus keep the original design matrix and only truncate the response. In (2.7), choose  $\widetilde{Y}_i = \text{sgn}(Y_i)(|Y_i| \wedge \tau)$  and  $\widetilde{\mathbf{X}}_i = \mathbf{X}_i$ , then we have

$$\widehat{\boldsymbol{\Sigma}}_{Y\mathbf{X}} = \widehat{\boldsymbol{\Sigma}}_{\widetilde{Y}\widetilde{\mathbf{X}}}(\tau) = \frac{1}{N} \sum_{i=1}^N \text{sgn}(Y_i)(|Y_i| \wedge \tau) \mathbf{X}_i \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{X}} = \frac{1}{N} \sum_{i=1}^N \text{vec}(\mathbf{X}_i) \text{vec}(\mathbf{X}_i)^T. \quad (3.5)$$

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

**Lemma 3.** *Consider the following conditions:*

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

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

There exists a constant  $\gamma > 0$  such that as long as  $(d_1 + d_2)/N < \gamma$ , it holds that

$$P\left(\|\widehat{\Sigma}_{Y\mathbf{X}}(\tau) - \frac{1}{N} \sum_{i=1}^N \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle \mathbf{X}_i\|_{op} \geq \nu \sqrt{\frac{d_1 + d_2}{N}}\right) \leq \eta \exp(-(d_1 + d_2)), \quad (3.6)$$

where  $\tau \asymp \sqrt{N/(d_1 + d_2)}$  and  $\nu$  and  $\eta$  are constants.

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

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

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

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

**Theorem 3.** Suppose Conditions (C1) and (C2) in Lemma 3 hold and  $\mathcal{B}_q(\boldsymbol{\Theta}^*) \leq \rho$ . We further assume that  $\text{vec}(\mathbf{X}_i)$  is Gaussian. Choose  $\tau \asymp \sqrt{N/(d_1 + d_2)}$  and  $\lambda_N = 2\nu \sqrt{(d_1 + d_2)/N}$ , where  $\nu$  is the same as in Lemma 3. There exist constants  $\{C_i\}_{i=1}^4$  such that once  $\rho((d_1 + d_2)/N)^{1-\frac{q}{2}} \leq C_1$ , we have

$$P\left(\|\widehat{\boldsymbol{\Theta}}(\tau, \lambda_N) - \boldsymbol{\Theta}^*\|_F^2 \geq C_2 \rho \left(\frac{d_1 + d_2}{N}\right)^{1-\frac{q}{2}}\right) \leq \eta \exp(-(d_1 + d_2)) + 2 \exp(-N/32),$$

and

$$P\left(\|\widehat{\boldsymbol{\Theta}}(\tau, \lambda_N) - \boldsymbol{\Theta}^*\|_N \geq C_3 \rho \left(\frac{d_1 + d_2}{N}\right)^{\frac{1-q}{2}}\right) \leq \eta \exp(-(d_1 + d_2)) + 2 \exp(-N/32),$$

where  $\eta$  is the same constant as in Lemma 3.

**Remark 5.** The Frobenius norm rate here is the same rate as established under sub-Gaussian noise in Negahban and Wainwright (2011). When  $q = 0$ ,  $\rho$  is the upper bound of the rank of



$\Theta^*$  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  $\Theta^*$ , rather than the ambiem number of parameters  $d_1 * d_2$

### 3.3 Matrix Completion

In this section, we consider the matrix completion problem with heavy-tailed noises. Under a conventional setting,  $\mathbf{X}_i$  is a singleton,  $\|\Theta^*\|_{\max} = O(1)$  and  $\|\Theta^*\|_F = O(\sqrt{d_1 d_2})$ . If we rescale the original model as

$$Y_i = \langle \mathbf{X}_i, \Theta^* \rangle + \epsilon_i = \langle \sqrt{d_1 d_2} \mathbf{X}_i, \Theta^* / \sqrt{d_1 d_2} \rangle + \epsilon_i$$

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

For the matrix completion problem, in order to achieve consistent estimation, we require the true coefficient matrix  $\Theta^*$  not to be overly spiky, i.e.,  $\|\Theta^*\|_{\max} \leq R \|\Theta^*\|_F / \sqrt{d_1 d_2} \leq R / \sqrt{d_1 d_2}$ . We put a similar constraint in seeking the corresponding M-estimator:

$$\hat{\Theta} \in \underset{\|\Theta\|_{\max} \leq R / \sqrt{d_1 d_2}}{\operatorname{argmin}} -\langle \hat{\Sigma}_{Y\mathbf{X}}(\tau), \Theta \rangle + \frac{1}{2} \operatorname{vec}(\Theta)^T \hat{\Sigma}_{\mathbf{X}\mathbf{X}} \operatorname{vec}(\Theta) + \lambda_N \|\Theta\|_N. \quad (3.7)$$

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  $\tilde{Y}_i = \operatorname{sgn}(Y_i)(|Y_i| \wedge \tau)$  and  $\tilde{\mathbf{X}}_i = \mathbf{X}_i$  in (2.7). Then,  $\hat{\Sigma}_{Y\mathbf{X}}$  and  $\hat{\Sigma}_{\mathbf{X}\mathbf{X}}$  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.** *Under the following conditions:*

- (C1)  $\|\Theta^*\|_F \leq 1$  and  $\|\Theta^*\|_{\max} \leq R / \sqrt{d_1 d_2}$ , where  $0 < R < \infty$ ;
- (C2)  $\mathbf{X}_i$  is uniformly sampled from  $\{\sqrt{d_1 d_2} \cdot \mathbf{e}_j \mathbf{e}_k^T\}_{1 \leq j \leq d_1, 1 \leq k \leq d_2}$ ;
- (C3)  $\forall i = 1, \dots, N$ ,  $E(E(\epsilon_i^2 | \mathbf{X}_i))^k \leq M < \infty$ , where  $k > 1$ ;

there exists a constant  $\gamma > 0$  such that for any  $\delta > 0$ , as long as  $(d_1 \vee d_2) \log(d_1 + d_2)/N < \gamma$ ,

$$P\left(\left\|\widehat{\Sigma}_{Y\mathbf{X}}(\tau) - \frac{1}{N} \sum_{i=1}^N \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle \mathbf{X}_i\right\|_{op} > \nu \sqrt{\frac{\delta(d_1 \vee d_2) \log(d_1 + d_2)}{N}}\right) \leq 2(d_1 + d_2)^{1-\delta}, \quad (3.8)$$

where  $\tau \asymp (\sqrt{N/((d_1 \vee d_2) \log(d_1 + d_2))})$  and  $\nu$  is a universal constant.

**Remark 6.** Again, for  $\bar{\Sigma}_{Y\mathbf{X}} = \frac{1}{N} \sum_{i=1}^N Y_i \mathbf{X}_i$ , Negahban and Wainwright (2012) proved that  $\|\bar{\Sigma}_{Y\mathbf{X}} - \frac{1}{N} \sum_{i=1}^N \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle \mathbf{X}_i\|_{op} = O_P(\sqrt{(d_1 + d_2) \log(d_1 + d_2)/N})$  for sub-exponential noise. Compared with this result, Lemma 4 achieves the same rate of convergence. By Jessen's inequality, condition (C3) is implied by  $E\epsilon_i^{2k} \leq M < \infty$ .

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

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

$$\tau \asymp \sqrt{N/((d_1 \vee d_2) \log(d_1 + d_2))} \quad \text{and} \quad \lambda_N = 2\nu \sqrt{\delta(d_1 \vee d_2) \log(d_1 + d_2)/N}$$

and assume  $(d_1 \vee d_2) \log(d_1 + d_2)/N < \gamma$ , where  $\nu$  and  $\gamma$  are the same as in Lemma 4. There exist universal constants  $\{C_i\}_{i=1}^4$  such that with probability at least  $1 - 2(d_1 + d_2)^{1-\delta} - C_1 \exp(-C_2(d_1 + d_2))$  we have

$$\|\widehat{\boldsymbol{\Theta}}(\tau, \lambda_N) - \boldsymbol{\Theta}^*\|_F^2 \leq C_3 \max\left\{\rho \left(\frac{\delta R^2(d_1 + d_2) \log(d_1 + d_2)}{N}\right)^{1-\frac{q}{2}}, \frac{R^2}{N}\right\}$$

and

$$\|\widehat{\boldsymbol{\Theta}}(\tau, \lambda_N) - \boldsymbol{\Theta}^*\|_N \leq C_4 \max\left\{\rho \left(\frac{\delta R^2(d_1 + d_2) \log(d_1 + d_2)}{N}\right)^{\frac{1-q}{2}}, \left(\frac{\rho R^{2-2q}}{N^{1-q}}\right)^{\frac{1}{2-q}}\right\},$$

where  $\widehat{\boldsymbol{\Theta}}$  is defined in (3.7).

**Remark 7.** Theorem 4 achieves the same statistical error rate of Frobenius norm as established in Negahban and Wainwright (2012), which also matches the information-theoretic lower bound established in Negahban and Wainwright (2012).

### 3.4 Multi-Task Regression

Before presenting the theoretical results, we first simplify (2.6) under the setting of multi-task regression. According to (2.8), (2.6) can be reduced to the following form:

$$\hat{\Theta} \in \operatorname{argmin}_{\Theta \in \mathcal{S}} \frac{1}{d_2} (-\langle \hat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}, \Theta \rangle + \frac{1}{n} \sum_{i=1}^n \|\Theta^T \tilde{\mathbf{x}}_i\|_2^2) + \lambda_N \|\Theta\|_N. \quad (3.9)$$

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

Under the sub-Gaussian design, we only need to shrink the response vector  $\mathbf{y}_i$ . We choose for (2.8)  $\tilde{\mathbf{x}}_i = \mathbf{x}_i$  and  $\tilde{\mathbf{y}}_i = (\|\mathbf{y}_i\|_2 \wedge \tau) \mathbf{y}_i / \|\mathbf{y}_i\|_2$ , where  $\tau$  is some threshold value that depends on  $n, d_1$  and  $d_2$ . In other words, we keep the original design, but shrink the Euclidean norm of the response. 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 is only of bounded moments, we need to shrink both the design vector  $\mathbf{x}_i$  and response vector  $\mathbf{y}_i$  by their  $\ell_4$  norm instead, i.e., we choose  $\tilde{\mathbf{x}}_i = (\|\mathbf{x}_i\|_4 \wedge \tau_1) \mathbf{x}_i / \|\mathbf{x}_i\|_4$  and  $\tilde{\mathbf{y}}_i = (\|\mathbf{y}_i\|_4 \wedge \tau_2) \mathbf{y}_i / \|\mathbf{y}_i\|_4$ , where  $\tau_1$  and  $\tau_2$  are two thresholds. Here shrinking by the fourth order moment in fact accelerates the convergence rate of the induced bias so that it will match the final statistical error rate. Again, we will not repeat stating these choices in the following lemmas and theorems for less redundancy.

**Lemma 5.** *Convergence of gradients of robustified quadratic loss.*

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

- (C1)  $\lambda_{\max}(\mathbb{E} \mathbf{y}_i \mathbf{y}_i^T) \leq R < \infty$ ;
- (C2)  $\{\mathbf{x}_i\}_{i=1}^n$  are i.i.d. sub-Gaussian vectors with  $\|\mathbf{x}_i\|_{\psi_2} \leq \kappa_0 < \infty$ ,  $\mathbb{E} \mathbf{x}_i = \mathbf{0}$  and  $\lambda_{\min}(\mathbb{E} \mathbf{x}_i \mathbf{x}_i^T) \geq \kappa_{\mathcal{L}} > 0$ .
- (C3)  $\forall i = 1, \dots, n$ ,  $j_1, j_2 = 1, \dots, d_2$  and  $j_1 \neq j_2$ ,  $\epsilon_{ij_1} \perp \epsilon_{ij_2} | \mathbf{x}_i$ , and  $\forall j = 1, \dots, d_1$ ,  $\mathbb{E}(\mathbb{E}(\epsilon_{ij}^2 | \mathbf{x}_i))^k \leq M < \infty$ , where  $k > 1$ ;

there exists some constant  $\gamma > 0$  such that if  $(d_1 + d_2) \log(d_1 + d_2)/n < \gamma$ , we have for any  $\delta > 0$ ,

$$P\left(\|\hat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}(\tau) - \hat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \Theta^*\|_{op} \geq \sqrt{\frac{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n}}\right) \leq 2(d_1 + d_2)^{1-\eta_1 \delta},$$

where  $\tau \asymp \sqrt{n/((d_1 + d_2) \log(d_1 + d_2))}$  and  $\nu_1$  and  $\eta_1$  are universal constants.

(b) **Bounded moment design.** Consider Condition (C2') that for any  $\mathbf{v} \in \mathcal{S}^{d_1-1}$ ,  $\mathbb{E}(\mathbf{v}^T \mathbf{x}_i)^4 \leq M < \infty$ . Under Conditions (C1), (C2') and (C3), it holds for any  $\delta > 0$

$$P\left(\|\widehat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}(\tau_1, \tau_2) - \widehat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(\tau_1) \Theta^*\|_{op} \geq \sqrt{\frac{(\nu_2 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n}}\right) \leq 2(d_1 + d_2)^{1-\eta_2\delta},$$

where  $\tau_1, \tau_2 \asymp (n/((d_1 + d_2) \log(d_1 + d_2)))^{\frac{1}{4}}$  and  $\nu_2$  and  $\eta_2$  are universal constants.

**Remark 8.** When the noise and design are sub-Gaussian, Negahban and Wainwright (2011) used the covering argument to show that for regular sample covariance matrices  $\bar{\Sigma}_{\mathbf{xy}}$  and  $\bar{\Sigma}_{\mathbf{xx}}$ ,

$$\|\bar{\Sigma}_{\mathbf{xy}} - \bar{\Sigma}_{\mathbf{xx}} \Theta^*\|_{op} = \|\bar{\Sigma}_{\mathbf{xy}} - \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^T \Theta^*\|_{op} = \left\| \frac{1}{n} \sum_{j=1}^n \epsilon_j \mathbf{x}_j^T \right\|_{op} = O_P(\sqrt{(d_1 + d_2)/n}).$$

Lemma 5 shows that up to just 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 rate for the low-rank multi-tasking regression.

**Theorem 5.** Statistical error rate for multitask regression. Assume  $\mathcal{B}_q(\Theta^*) \leq \rho$ .

(a) **Sub-Gaussian design.** Suppose that Conditions (C1), (C2) and (C3) in Lemma 5 hold. For any  $\delta > 0$ , choose

$$\tau \asymp \sqrt{n/((d_1 + d_2) \log(d_1 + d_2))} \quad \text{and} \quad \lambda_N = \frac{2}{d_2} \sqrt{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)/n},$$

where  $\nu_1$  is the same as in Lemma 5. There exist constants  $\gamma_1, \gamma_2 > 0$  such that if  $(d_1 + d_2) \log(d_1 + d_2)/n < \gamma_1$  and  $d_1 + d_2 \geq \gamma_2$ , then with probability at least  $1 - 3(d_1 + d_2)^{1-\eta_1\delta}$  we have

$$\|\widehat{\Theta}(\tau, \lambda_N) - \Theta^*\|_F^2 \leq C_1 \rho \left( \frac{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{1-\frac{q}{2}}$$

and

$$\|\widehat{\Theta}(\tau, \lambda_N) - \Theta^*\|_N \leq C_2 \rho \left( \frac{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{\frac{1-q}{2}},$$

where  $C_1$  and  $C_2$  are universal constants and  $\eta_1$  is the same as in Lemma 5.

(b) **Bounded moment design.** Suppose instead that Conditions (C1), (C2') and (C3) in Lemma 5 hold. For any  $\delta > 0$ , choose

$$\tau_1, \tau_2 \asymp \left( n / ((d_1 + d_2) \log(d_1 + d_2)) \right)^{\frac{1}{4}} \quad \text{and} \quad \lambda_N = \frac{2}{d_2} \sqrt{(\nu_2 + \delta)(d_1 + d_2) \log(d_1 + d_2) / n},$$

where  $\nu_2$  is the same as in Lemma 5. There exist constants  $\gamma_3, \gamma_4 > 0$  such that if  $(d_1 + d_2) \log(d_1 + d_2) / n < \gamma_3$  and  $d_1 + d_2 \geq \gamma_4$ , then with probability at least  $1 - 3(d_1 + d_2)^{1-\eta_2\delta}$ ,

$$\|\widehat{\Theta}(\tau_1, \tau_2, \lambda_N) - \Theta^*\|_F^2 \leq C_3 \rho \left( \frac{(\nu_2 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{1-\frac{q}{2}}$$

and

$$\|\widehat{\Theta}(\tau_1, \tau_2, \lambda_N) - \Theta^*\|_N \leq C_4 \rho \left( \frac{(\nu_2 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{\frac{1-q}{2}},$$

where  $C_3$  and  $C_4$  are universal constants and  $\eta_2$  is the same as in Lemma 5.

## 4 Robust Covariance Estimation

In derivation of  $\|\widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}}(\tau_1, \tau_2) - \widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}(\tau_1)\Theta^*\|_{op}$  in multi-tasking regression, we find that when the random sample has only bounded moments, the  $\ell_4$ -norm shrinkage sample covariance achieves the spectral norm convergence rate of order  $O_P(\sqrt{d \log d / n})$  in estimating the true covariance. This is nearly the optimal rate with sub-Gaussian random samples, up to a logarithmic factor. Here we formulate the problem and the result, whose proof is relegated to the appendix.

Suppose we have  $n$  i.i.d.  $d$ -dimensional random vectors  $\{\mathbf{x}_i\}_{i=1}^n$  with  $E \mathbf{x}_i = \mathbf{0}$ . Our goal is to estimate the covariance matrix  $\Sigma = E(\mathbf{x}_i \mathbf{x}_i^T)$  when the distribution of  $\{\mathbf{x}_i\}_{i=1}^n$  has only fourth bounded 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  $\ell_4$ -norm. We propose the following shrinkage sample covariance to estimate  $\Sigma$ .

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

**Theorem 6.** Suppose  $E(\mathbf{v}^T \mathbf{x}_i)^4 \leq R$  for any  $\mathbf{v} \in \mathcal{S}^{d-1}$ , then it holds that for any  $\delta > 0$ ,

$$P\left(\|\tilde{\Sigma}_n(\tau) - \Sigma\|_{op} \geq \sqrt{\frac{\delta R d \log d}{n}}\right) \leq d^{1-C\delta}, \quad (4.2)$$

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

We have several comments for the above theorem. First of all, unlike the bounded moment conditions in the previous section,  $R$  here can go to infinity with certain rates. Hence we also put this quantity into the rate of convergence. If indeed  $R < \infty$ , we recover the true covariance matrix with statistical error of  $O_P(\sqrt{d \log d/n})$ . In comparison with the robust covariance matrix estimator given in Fan et al. (2016), our estimator here is positive semidefinite and is very simple to implement. Our concentration inequality here is on the operator norm, while their result is on the element-wise max norm.

If we are concerned with statistical error in terms of elementwise max norm, then we need to apply elementwise truncation to the random samples rather than  $\ell_4$  norm shrinkage. Let  $\check{\mathbf{x}}_i$  satisfy  $\check{x}_{ij} = \text{sgn}(x_{ij})(|x_{ij}| \wedge \tau)$  for  $1 \leq j \leq d_1$  and  $\check{\Sigma}_n = \sum_{i=1}^n \check{\mathbf{x}}_i \check{\mathbf{x}}_i^T$ . It is not hard to derive  $\|\check{\Sigma}_n - \Sigma\|_{\max} = O_P(\sqrt{\log d/n})$  as in Fan et al. (2016) with  $\tau \asymp (n/\log d)^{\frac{1}{4}}$ . Further regularization can be applied to  $\check{\Sigma}_n$  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 et al. (2013), among others.

Theorem 6 is non-asymptotic as all the results in the previous sections and can be applied to both low-dimensional and high-dimensional regimes. In fact,  $\tilde{\Sigma}_n(\tau)$  can outperform the sample covariance  $\bar{\Sigma}_n$  even with Gaussian samples if the dimension is high. The reason is that according to Theorem 5.39 in Vershynin (2010),  $\|\bar{\Sigma}_n - \Sigma\|_{op} = O_P(\sqrt{d/n} \vee (d/n))$ . 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 retains the statistical error of order  $\sqrt{d \log d/n}$  regardless of ratio between the dimension and sample size. 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.

## 5 Simulation Study

In this section, we first compare the numerical performance of the robust procedure and standard procedure in compressed sensing, matrix completion and multi-tasking regression. For each setup, we investigate three noise settings: log-normal noise, truncated Cauchy noise and Gaussian noise. They represent heavy-tailed asymmetric distributions, heavy-tailed symmetric distributions and light-tailed symmetric distributions. The results from the standard and robust methods are shown in the same color for each scenario in the following figures so that they can be compared more easily.

All the robust procedures proposed in our work are very easy to implement; we only need to truncate or shrink the data appropriately, and then apply the standard procedure to the transformed data. As for the parameter tuning, we refer to the rate developed in our theories for robust procedures and theories in Negahban and Wainwright (2011) and Negahban and Wainwright (2012) for standard procedures. The constants before the rate are tuned for best performance. The main message is that the robust procedure outperforms the standard procedure under the setting with bounded moment noise, and it performs equally well as the standard procedure under the Gaussian noise. The simulations are based on 100 independent Monte Carlo replications. There are no standard algorithms for solving penalized trace regression problems. Hence, besides presenting the numerical performance, we also elucidate the algorithms that we use to solve the corresponding optimization problems, which might be of interest to readers concerned with implementations. These algorithms are not necessarily the same as those used in the literature.

We also compare the numerical performance of the regular sample covariance and shrinkage sample covariance as proposed in (4.1) in estimating the true covariance. We choose  $d/n = 0.2, 0.5, 1$  and for each ratio, we let  $n = 100, 200, \dots, 500$ . 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 can not only overcome the heavy-tailed corruption, but also mitigate the curse of high dimensions.

### 5.1 Compressed Sensing

We first specify the parameters in the true model:  $Y = \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle + \epsilon_i$ . We always set  $d_1 = d_2 = d$ ,  $\|\boldsymbol{\Theta}^*\|_F = \sqrt{5}$  and  $\text{rank}(\boldsymbol{\Theta}^*) = 5$ . In the simulation, we construct  $\boldsymbol{\Theta}^*$  to be

$\sum_{i=1}^5 \mathbf{v}_i \mathbf{v}_i^T$ , where  $\mathbf{v}_i$  is the  $i$ th top eigenvector of the sample covariance of 100 i.i.d. centered Gaussian random vectors with covariance  $\mathbf{I}_d$ . The design matrix  $\mathbf{X}_i$  has i.i.d. standard Gaussian entries. The noise distributions are characterized as follows:

- Log-normal:  $\epsilon_i \stackrel{d}{=} (Z - \mathbb{E} Z)/50$ , where  $Z \sim \ln \mathcal{N}(0, \sigma^2)$  and  $\sigma^2 = 6.25$ ;
- Truncated Cauchy:  $\epsilon_i \stackrel{d}{=} \min(Z, 10^3)/10$ , where  $Z$  follows Cauchy distribution;
- Gaussian:  $\epsilon_i \sim N(0, \sigma^2)$ , where  $\sigma^2 = 0.25$ .

The constants above are chosen to ensure appropriate signal-to-noise ratio for better presentation. We present the numerical results in Figure 2. As we can observe from the plots, the robust estimator has much smaller statistical error than the standard estimator under the heavy-tailed noise, i.e., the log-normal and truncated Cauchy noise. When  $d = 40$  or  $60$ , 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 the setting of Gaussian noise, the robust estimator has nearly the same statistical performance as the standard one, which shows that it does not hurt to use the robust procedure under the light-tailed noise setting.

As for the implementation, we exploit the contractive Peaceman-Rachford splitting method (PRSM) to solve the compressed sensing problem. Here we briefly introduce the general scheme of the contractive PRSM for clarity. The contractive PRSM is for minimizing the summation of two convex functions under linear constraint:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^{p_1}, \mathbf{y} \in \mathbb{R}^{p_2}} \quad & f_1(\mathbf{x}) + f_2(\mathbf{y}), \\ \text{subject to} \quad & \mathbf{C}_1 \mathbf{x} + \mathbf{C}_2 \mathbf{y} - \mathbf{c} = \mathbf{0}, \end{aligned} \tag{5.1}$$

where  $\mathbf{C}_1 \in \mathbb{R}^{p_3 \times p_1}$ ,  $\mathbf{C}_2 \in \mathbb{R}^{p_3 \times p_2}$  and  $\mathbf{c} \in \mathbb{R}^{p_3}$ . The general iteration scheme of the contractive



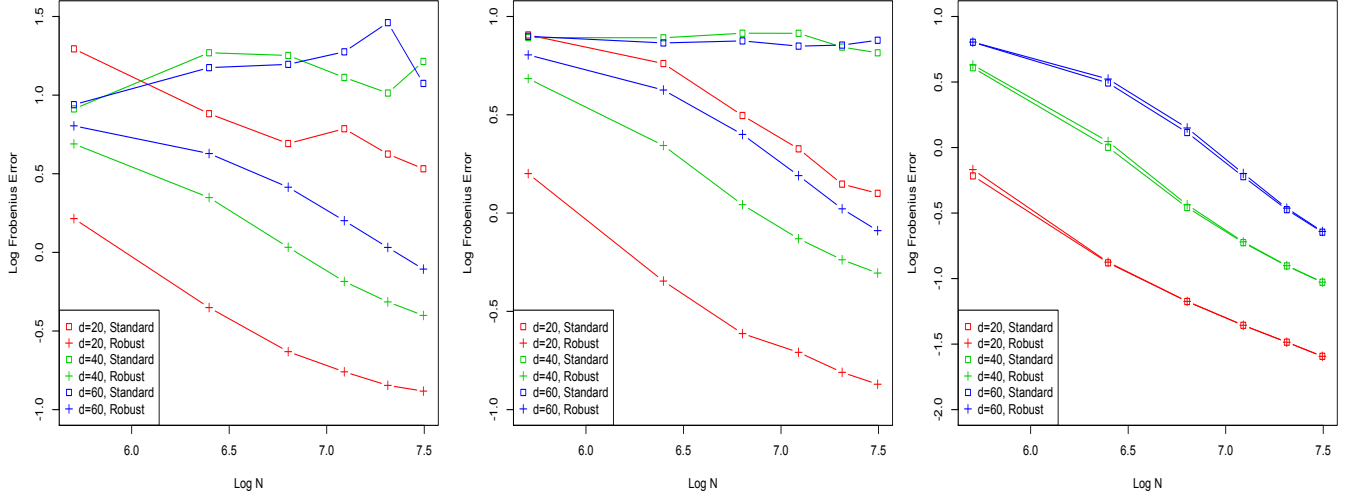


Figure 2: Statistical errors of  $\ln\|\hat{\Theta} - \Theta^*\|_F$  v.s. logarithmic sample size  $\ln N$  for different dimensions  $d$  in matrix compressed sensing.

PRSM is

$$\begin{cases} \mathbf{x}^{(k+1)} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ f_1(\mathbf{x}) - (\boldsymbol{\rho}^{(k)})^T (\mathbf{C}_1 \mathbf{x} + \mathbf{C}_2 \mathbf{y}^{(k)} - \mathbf{c}) + \frac{\beta}{2} \|\mathbf{C}_1 \mathbf{x} + \mathbf{C}_2 \mathbf{y}^{(k)} - \mathbf{c}\|_2^2 \right\}, \\ \boldsymbol{\rho}^{(k+\frac{1}{2})} = \boldsymbol{\rho}^{(k)} - \alpha \beta (\mathbf{C}_1 \mathbf{x}^{(k+1)} + \mathbf{C}_2 \mathbf{y}^{(k)} - \mathbf{c}), \\ \mathbf{y}^{(k+1)} = \underset{\mathbf{y}}{\operatorname{argmin}} \left\{ f_2(\mathbf{y}) - (\boldsymbol{\rho}^{(k+\frac{1}{2})})^T (\mathbf{C}_1 \mathbf{x}^{(k+1)} + \mathbf{C}_2 \mathbf{y} - \mathbf{c}) + \frac{\beta}{2} \|\mathbf{C}_1 \mathbf{x}^{(k+1)} + \mathbf{C}_2 \mathbf{y} - \mathbf{c}\|_2^2 \right\}, \\ \boldsymbol{\rho}^{(k+1)} = \boldsymbol{\rho}^{(k+\frac{1}{2})} - \alpha \beta (\mathbf{C}_1 \mathbf{x}^{(k+1)} + \mathbf{C}_2 \mathbf{y}^{(k+1)} - \mathbf{c}), \end{cases} \quad (5.2)$$

where  $\boldsymbol{\rho} \in \mathbb{R}^{p_3}$  is the Lagrangian multiplier,  $\beta$  is the penalty parameter and  $\alpha$  is the relaxation factor (Eckstein and Bertsekas (1992)). Since the parameter of interest in our work is  $\Theta^*$ , now we use  $\boldsymbol{\theta}_x \in \mathbb{R}^{d_1 d_2}$  and  $\boldsymbol{\theta}_y \in \mathbb{R}^{d_1 d_2}$  to replace  $\mathbf{x}$  and  $\mathbf{y}$  respectively in (5.2). By substituting

$$f_1(\boldsymbol{\theta}_x) = \frac{1}{N} \sum_{i=1}^N (Y_i - \langle \operatorname{vec}(\mathbf{X}_i), \boldsymbol{\theta}_x \rangle)^2, \quad f_2(\boldsymbol{\theta}_y) = \lambda \|\operatorname{mat}(\boldsymbol{\theta}_y)\|_N, \quad \mathbf{C}_1 = \mathbf{I}, \quad \mathbf{C}_2 = -\mathbf{I} \text{ and } \mathbf{c} = \mathbf{0}$$

into (5.2), we can obtain the PRSM algorithm for the compressed sensing problem. Let  $\mathbb{X}$  be a  $N$ -by- $d_1 d_2$  matrix whose rows are i.i.d. random designs  $\{\operatorname{vec}(\mathbf{X}_i)\}_{i=1}^N$  and  $\mathbb{Y}$  be the

$N$ -dimensional response vector. Then we have the following iteration scheme specifically for compressed sensing.

$$\begin{cases} \boldsymbol{\theta}_x^{(k+1)} = (2\mathbb{X}^T\mathbb{X}/N + \beta \cdot \mathbf{I})^{-1}(\beta \cdot \boldsymbol{\theta}_y^{(k)} + \boldsymbol{\rho}^{(k)} + 2\mathbb{X}^T\mathbb{Y}/N), \\ \boldsymbol{\rho}^{(k+\frac{1}{2})} = \boldsymbol{\rho}^{(k)} - \alpha\beta(\boldsymbol{\theta}_x^{(k+1)} - \boldsymbol{\theta}_y^{(k)}), \\ \boldsymbol{\theta}_y^{(k+1)} = \text{vec}(S_{\lambda/\beta}(\text{mat}(\boldsymbol{\theta}_x - \boldsymbol{\rho}^{(k+\frac{1}{2})}))), \\ \boldsymbol{\rho}^{(k+1)} = \boldsymbol{\rho}^{(k+\frac{1}{2})} - \alpha\beta(\boldsymbol{\theta}_x^{(k+1)} - \boldsymbol{\theta}_y^{(k+1)}), \end{cases} \quad (5.3)$$

where we choose  $\alpha = 0.9$  and  $\beta = 1$  according to Eckstein and Bertsekas (1992) and He et al. (2014),  $\boldsymbol{\rho} \in \mathbb{R}^{d_1 d_2}$  is the Lagrangian multiplier and  $S_\tau(\mathbf{z})$  is the singular value soft-thresholding function for matrix version of  $\mathbf{z} \in \mathbb{R}^{d_1 d_2}$ . To be more specific, let  $\mathbf{Z} = \text{mat}(\mathbf{z}) \in \mathbb{R}^{d_1 \times d_2}$  and  $\mathbf{Z} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T = \mathbf{U} \text{diag}(\lambda_1, \dots, \lambda_r) \mathbf{V}^T$  be its singular value decomposition. Then  $S_\tau(\mathbf{z}) = \text{vec}(\mathbf{U} \text{diag}((\lambda_1 - \tau)_+, (\lambda_2 - \tau)_+, \dots, (\lambda_r - \tau)_+) \mathbf{V}^T)$ , where  $(x)_+ = \max(x, 0)$ . The algorithm stops if  $\|\boldsymbol{\theta}_x - \boldsymbol{\theta}_y\|_2$  is smaller than some predetermined threshold, and returns  $\text{mat}(\boldsymbol{\theta}_y)$  as the final estimator of  $\boldsymbol{\Theta}^*$ .

## 5.2 Matrix Completion

We again set  $d_1 = d_2 = d$  and construct  $\boldsymbol{\Theta}^*$  to be  $\sum_{i=1}^5 \mathbf{v}_i \mathbf{v}_i^T / \sqrt{5}$ , where  $\mathbf{v}_i$  is the  $i$ th top eigenvector of the sample covariance of 100 i.i.d. centered Gaussian random vectors with covariance  $\mathbf{I}_d$ . Each design matrix  $\mathbf{X}_i$  takes the singleton form, which is uniformly sampled from  $\{\mathbf{e}_j \mathbf{e}_k^T\}_{1 \leq j, k \leq d}$ . The noise distributions are

- Log-normal:  $\epsilon_i \stackrel{d}{=} (Z - \mathbb{E} Z)/250$ , where  $Z \sim \ln \mathcal{N}(0, \sigma^2)$  and  $\sigma^2 = 9$ ;
- Truncated Cauchy:  $\epsilon_i \stackrel{d}{=} \min(Z, 10^3)/16$ , where  $Z$  follows Cauchy distribution;
- Gaussian:  $\epsilon_i \sim N(0, \sigma^2)$ , where  $\sigma^2 = 0.25$ .

Again, the constants above are set for an appropriate signal-to-noise ratio for better presentation. We present the numerical results in Figure 3. 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  $\boldsymbol{\Theta}^*$  under the log-normal and truncated Cauchy noise. Under Gaussian noise, the robust procedure has nearly the same statistical performance as the standard procedure.

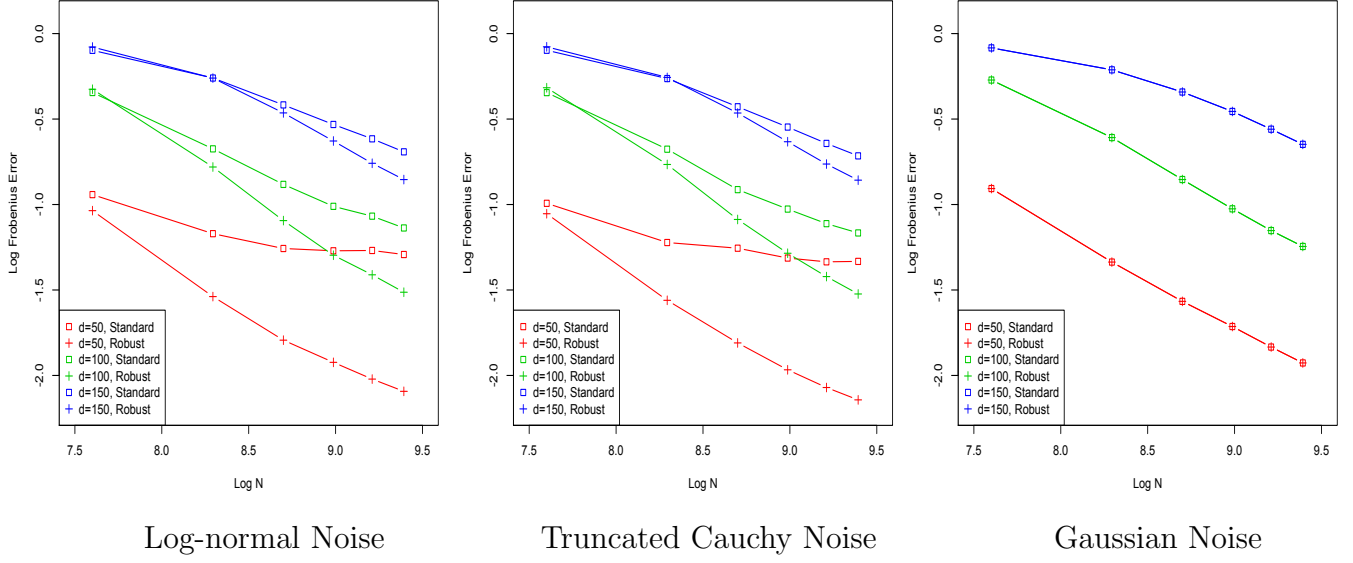


Figure 3: Statistical errors of  $\ln\|\hat{\Theta} - \Theta^*\|_F$  v.s. logarithmic sample size  $\ln N$  for different dimensions  $d$  in matrix completion.

To solve the matrix completion problem in (3.7), we adapt the ADMM method inspired by Fang et al. (2015). They propose to recover the matrix by minimizing the square loss plus both nuclear norm and matrix max-norm penalizations under the entrywise max-norm constraint. By simply setting the penalization coefficient for the matrix max-norm to be zero, we can apply their algorithm to our problem. Let  $\mathbf{L}, \mathbf{R}, \mathbf{W} \in \mathbb{R}^{(d_1+d_2) \times (d_1+d_2)}$ , which are variables in our algorithm. Define  $\Theta^n, \Theta^s \in \mathbb{R}^{d_1 \times d_2}$  such that  $\Theta_{ij}^n = \sum_{t=1}^N 1_{\{\mathbf{x}_t = \mathbf{e}_i \mathbf{e}_j^T\}}$  and  $\Theta_{ij}^s = \sum_{t=1}^N Y_t 1_{\{\mathbf{x}_t = \mathbf{e}_i \mathbf{e}_j^T\}}$ , so  $\Theta^n$  and  $\Theta^s$  are constants given the data. Below we present the iteration scheme of our algorithm to solve the matrix completion problem (3.7). Readers who are interested in the derivation of the algorithm can refer to Fang et al. (2015) for the technical details.

$$\left\{ \begin{array}{l} \mathbf{L}^{(k+1)} = \Pi_{\mathcal{S}_+^{d_1+d_2}} \{ \mathbf{R}^{(k)} - \rho^{-1}(\mathbf{W}^{(k)} + \lambda_N \mathbf{I}) \}, \\ \mathbf{C} = \begin{pmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{pmatrix} = \mathbf{L}^{(k+1)} + \mathbf{W}^{(k)}/\rho, \\ \mathbf{R}_{ij}^{12} = \Pi_{[-R,R]} \{ (\rho \mathbf{C}_{ij}^{12} + 2\boldsymbol{\Theta}_{ij}^s/N)/(\rho + 2\boldsymbol{\Theta}_{ij}^n/N) \}, \quad 1 \leq i \leq d_1, 1 \leq j \leq d_2 \\ \mathbf{R}^{(k+1)} = \begin{pmatrix} \mathbf{C}^{11} & \mathbf{R}^{12} \\ (\mathbf{R}^{12})^T & \mathbf{C}^{22} \end{pmatrix}, \\ \mathbf{W}^{(k+1)} = \mathbf{W}^{(k)} + \gamma \rho (\mathbf{L}^{(k+1)} - \mathbf{R}^{(k+1)}). \end{array} \right. \quad (5.4)$$

In the algorithm above,  $\Pi_{\mathcal{S}_+^{d_1+d_2}}(\cdot)$  is the projection operator onto the space of positive semidefinite matrices  $\mathcal{S}_+^{d_1+d_2}$ ,  $\rho$  is the penalization parameter which we set to be 0.1 in our simulation and  $\gamma$  is the step length which is typically set to be 1.618 according to Fang et al. (2015). We omit the stopping criteria for this algorithm here since it is complex. Readers who are interested in the implementation details of this algorithm can refer to Fang et al. (2015) for detailed instruction. Once the stopping criteria is satisfied, the algorithm returns  $\mathbf{R}^{12}$  as the final estimator of  $\boldsymbol{\Theta}^*$ .

### 5.3 Multi-Tasking Regression

We again set  $d_1 = d_2 = d$  and construct  $\boldsymbol{\Theta}^*$  to be  $\sum_{i=1}^5 \mathbf{v}_i \mathbf{v}_i^T$ , where  $\mathbf{v}_i$  is the  $i$ th top eigenvector of the sample covariance of 100 i.i.d centered Gaussian random vectors with covariance  $\mathbf{I}_d$ . The design vectors  $\{\mathbf{x}_i\}_{i=1}^n$  are i.i.d. Gaussian vectors with covariance matrix  $\mathbf{I}_d$ . The noise distributions are characterized as follows:

- Log-normal:  $\epsilon_i \stackrel{d}{=} (Z - \mathbb{E} Z)/50$ , where  $Z \sim \ln \mathcal{N}(0, \sigma^2)$  and  $\sigma^2 = 4$ ;
- Truncated Cauchy:  $\epsilon_i \stackrel{d}{=} \min(Z, 10^4)/10$ , where  $Z$  follows Cauchy distribution;
- Gaussian:  $\epsilon_i \sim N(0, \sigma^2)$ , where  $\sigma^2 = 0.25$ .

We present the numerical results in Figure 4. Similar to the two examples before, the robust procedure has sharper statistical accuracy in estimating  $\boldsymbol{\Theta}^*$  than the standard procedure under both heavy-tailed noises. Under Gaussian noise, the robust procedure has nearly the same statistical performance as the standard procedure.

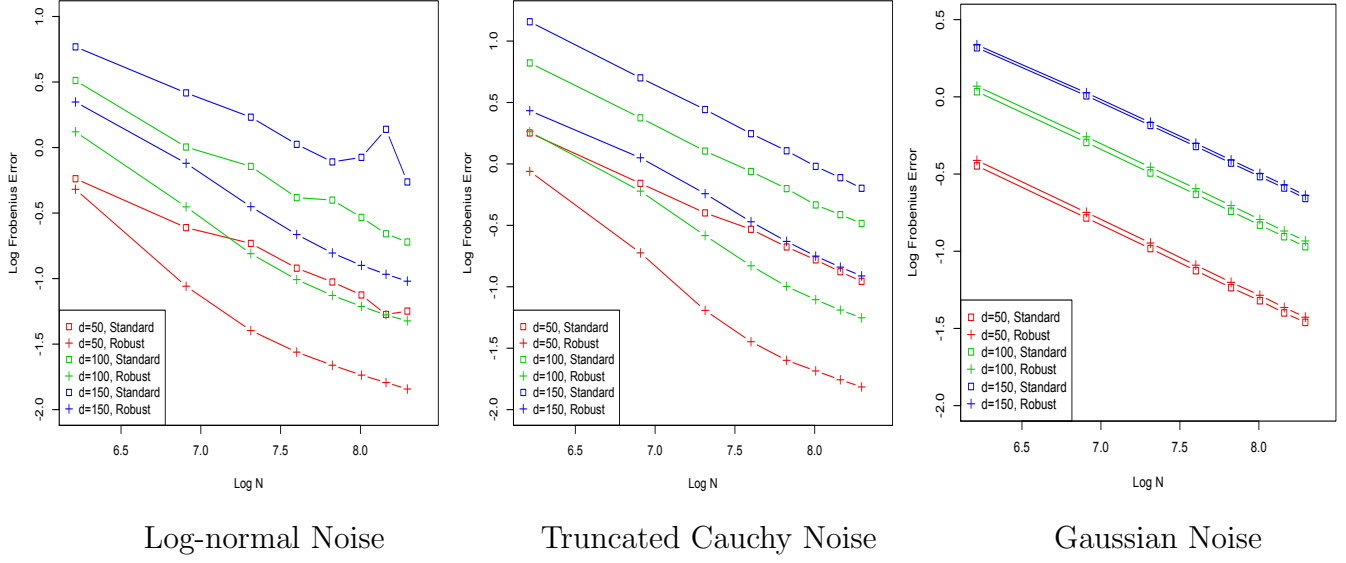


Figure 4: Statistical errors of  $\ln\|\hat{\Theta} - \Theta^*\|_F$  v.s. logarithmic sample size  $\ln N$  for different dimensions  $d$  in multi-tasking regression.

For multi-tasking regression, we exploit the contractive PRSM method again. Let  $\mathbb{X}$  be the  $n$ -by- $d_1$  design matrix and  $\mathbb{Y}$  be the  $n$ -by- $d_2$  response matrix. By following the general iteration scheme (5.2), we can develop the iteration steps for the multi-tasking regression.

$$\begin{cases} \Theta_x^{(k+1)} = (2\mathbb{X}^T\mathbb{X}/n + \beta \cdot \mathbf{I})^{-1}(\beta \cdot \Theta_y^{(k)} + \rho^{(k)} + 2\mathbb{X}^T\mathbb{Y}/n), \\ \rho^{(k+\frac{1}{2})} = \rho^{(k)} - \alpha\beta(\Theta_x^{(k+1)} - \Theta_y^{(k)}), \\ \Theta_y^{(k+1)} = S_{\lambda/\beta}(\Theta_x - \rho^{(k+\frac{1}{2})}), \\ \rho^{(k+1)} = \rho^{(k+\frac{1}{2})} - \alpha\beta(\Theta_x^{(k+1)} - \Theta_y^{(k+1)}), \end{cases} \quad (5.5)$$

where  $\Theta_x, \Theta_y, \rho \in \mathbb{R}^{d_1 \times d_2}$  and  $S_\tau(\cdot)$  is the same singular value soft thresholding function as in (5.3). Analogous to the compressed sensing, we choose  $\alpha = 0.9$  and  $\beta = 1$ . As long as  $\|\Theta_x - \Theta_y\|_F$  is smaller than some predetermined threshold, the iteration stops and the algorithm returns  $\Theta_y$  as the final estimator of  $\Theta^*$ .

## 5.4 Covariance Estimation

In this subsection, we investigate the statistical error of the sample covariance  $\bar{\Sigma}_n$  and shrinkage sample covariance  $\tilde{\Sigma}_n(\tau)$  proposed in Section 4 when the random samples are heavy-tailed. We only consider two simple distributions: Gaussian and Student's  $t_3$  random samples. The dimension is set to be proportional to sample size, i.e.,  $d/n = \alpha$  with  $\alpha$  being 0.2, 0.5, 1.  $n$  will range from 100 to 500 for each case. Regardless of how large the dimension  $d$  is, the true covariance  $\Sigma$  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. We present our results in Figure 5. The statistical errors are measured in terms of the spectral norm gap between the estimator and true covariance, and our simulation is based on 1,000 independent Monte Carlo replications.

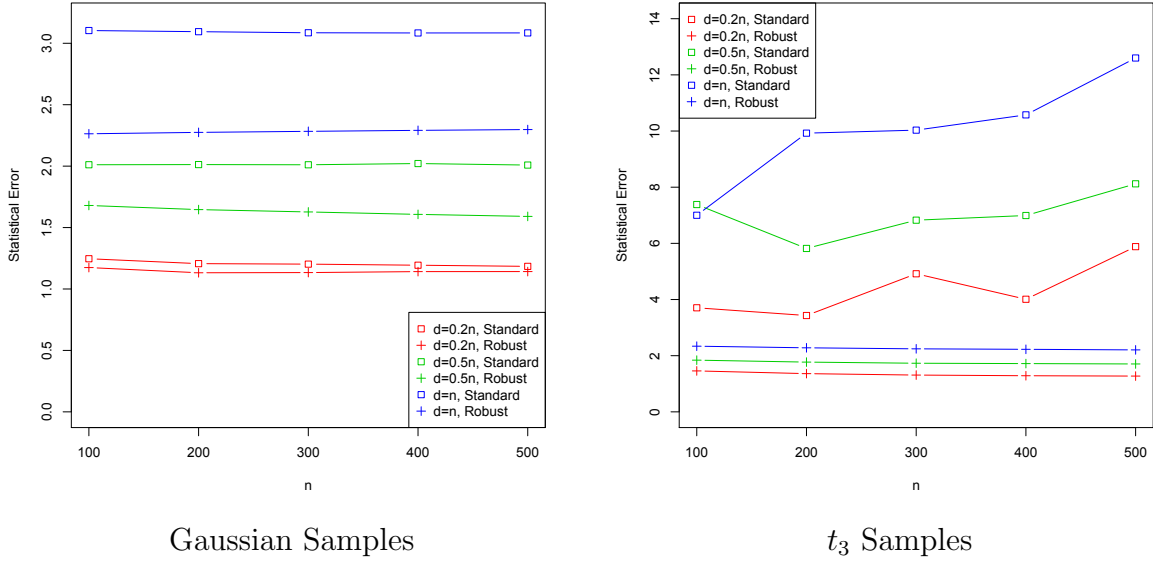


Figure 5: Statistical errors of  $\|\hat{\Sigma} - \Sigma\|_{op}$  v.s. sample size  $n$  for different dimensions  $d$  in covariance estimation.

As we can see, for Gaussian samples, as long as we fix  $d/n$ , the statistical error of both  $\bar{\Sigma}_n$  and  $\tilde{\Sigma}_n(\tau)$  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  $\tilde{\Sigma}_n(\tau)$  is over  $\bar{\Sigma}_n$ . This validates our remark after Theorem 6 that the shrinkage ameliorates the impact 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 stable statistical error and enjoys much higher accuracy than the regular sample covariance. This strongly supports the sharp statistical error rate we derived for  $\tilde{\Sigma}_n(\tau)$  in Theorem 6.

## References

- BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Annals of Statistics* **36** 2577–2604.
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* **37** 1705–1732.
- BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford.
- BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Annals of Statistics* **43** 2507–2536.
- CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684.
- CAI, T. T. and ZHANG, A. (2014). Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Transactions on Information Theory* **60** 122–132.
- CAI, T. T. and ZHANG, A. (2015). Rop: Matrix recovery via rank-one projections. *Annals of Statistics* **43** 102–138.
- CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics* **40** 2389–2420.
- CANDES, E. and TAO, T. (2007). The dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ . *Annals of Statistics* **35** 2313–2351.

- CANDES, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathématique* **346** 589–592.
- CANDES, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proceedings of the 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 Transactions on Information Theory* **57** 2342–2359.
- CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics* **9** 717–772.
- CANDES, E. J. and TAO, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory* **52** 5406–5425.
- CATONI, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l’Institut Henri Poincaré* **48** 1148–1185.
- CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review* **43** 129–159.
- DONOHO, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory* **52** 1289–1306.
- DONOHO, D. L., JOHNSTONE, I. and MONTANARI, A. (2013). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Transactions on Information Theory* **59** 3396–3433.
- ECKSTEIN, J. and BERTSEKAS, D. P. (1992). On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming* **55** 293–318.
- FAN, J., LI, Q. and WANG, Y. (2016). Robust estimation of high-dimensional mean regression. *Journal of Royal Statistical Society, Series B* .
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.



- FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 603–680.
- FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911.
- FAN, J. and YAO, Q. (2015). *Elements of Financial Econometrics*. Science Press, Beijing.
- FANG, X. E., LIU, H., TOH, K. C. and ZHOU, W.-X. (2015). Max-norm optimization for robust matrix recovery. Tech. rep.
- 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. *Nature communications* **5**.
- HE, B., LIU, H., WANG, Z. and YUAN, X. (2014). A strictly contractive peaceman–rachford splitting method for convex programming. *SIAM Journal on Optimization* **24** 1011–1040.
- HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research* **17** 1–40.
- IBRAGIMOV, R. and SHARAKHMETOV, S. (1998). Short communications: On an exact constant for the rosenthal inequality. *Theory of Probability & Its Applications* **42** 294–302.
- KIM, S. and XING, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. *Annals of Applied Statistics* **6** 1095–1117.
- KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics* **39** 2302–2329.
- LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37** 4254–4278.

- LOH, P.-L. (2015). Statistical consistency and asymptotic normality for high-dimensional robust  $m$ -estimators. *arXiv preprint arXiv:1501.00312* .
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34** 1436–1462.
- MINSKER, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli* **21** 2308–2335.
- NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics* **39** 1069–1097.
- NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research* **13** 1665–1697.
- NEGAHBAN, S., YU, B., WAINWRIGHT, M. J. and RAVIKUMAR, P. K. (2012). A unified framework for high-dimensional analysis of  $m$ -estimators with decomposable regularizers. *Statistical Science* **24** 538–577.
- NEMIROVSKY, A.-S., YUDIN, D.-B. and DAWSON, E.-R. (1982). Problem complexity and method efficiency in optimization .
- NOWAK, R. D., WRIGHT, S. J. ET AL. (2007). Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing* **1** 586–597.
- OLIVEIRA, R. I. (2016). The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields* 1–20.
- RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over-balls. *IEEE Transactions on Information Theory* **57** 6976–6994.
- RECHT, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research* **12** 3413–3430.

- RECHT, B., FAZEL, M. and PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* **52** 471–501.
- ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Annals of Statistics* **39** 887–930.
- RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59** 3434–3447.
- STOCK, J. H. and WATSON, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* **20** 147–162.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288.
- TROPP, J. A. (2012). User-friendly tools for random matrices: An introduction. Tech. rep., DTIC Document.
- TROPP, J. A. (2015). An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571* .
- VELU, R. and REINSEL, G. C. (2013). *Multivariate reduced-rank regression: theory and applications*, vol. 136. Springer Science & Business Media.
- VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .
- ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of statistics* **38** 894–942.
- ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36** 1509–1533.

## A Proofs

*Proof of Theorem 1.* This Lemma is just a simple application of the theoretical framework established in Negahban et al. (2012), but for completeness and clarity, we present the whole proof here. We start from the optimality of  $\hat{\Theta}$ :

$$-\langle \hat{\Sigma}_{YX}, \hat{\Theta} \rangle + \frac{1}{2} \text{vec}(\hat{\Theta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Theta}) + \lambda_N \|\hat{\Theta}\|_N \leq -\langle \hat{\Sigma}_{YX}, \Theta^* \rangle + \frac{1}{2} \text{vec}(\Theta^*)^T \hat{\Sigma}_{XX} \text{vec}(\Theta^*) + \lambda_N \|\Theta^*\|_N.$$

Note that  $\hat{\Delta} = \hat{\Theta} - \Theta^*$ . Simple algebra delivers that

$$\begin{aligned} \frac{1}{2} \text{vec}(\hat{\Delta})^T \hat{\Sigma}_{XX} \text{vec}(\hat{\Delta}) &\leq \langle \hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*)), \hat{\Delta} \rangle + \lambda_N \|\hat{\Delta}\|_N \\ &\leq \|\hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op} \cdot \|\hat{\Delta}\|_N + \lambda_N \|\hat{\Delta}\|_N \leq 2\lambda_N \|\hat{\Delta}\|_N, \end{aligned} \quad (\text{A.1})$$

if  $\lambda_N \geq 2\|\hat{\Sigma}_{YX} - \text{mat}(\hat{\Sigma}_{XX} \text{vec}(\Theta^*))\|_{op}$ . To bound the RHS of (A.1), we need to decompose  $\hat{\Delta}$  as Negahban and Wainwright (2011) did. Let  $\Theta^* = \mathbf{U}\mathbf{D}\mathbf{V}^T$  be the SVD of  $\Theta^*$ , where the diagonals of  $\mathbf{D}$  are in the decreasing order. Denote the first  $r$  columns of  $\mathbf{U}$  and  $\mathbf{V}$  by  $\mathbf{U}^r$  and  $\mathbf{V}^r$  respectively, and define

$$\begin{aligned} \mathcal{M} &:= \{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \text{row}(\Theta) \subseteq \text{col}(\mathbf{V}^r), \text{col}(\Theta) \subseteq \text{col}(\mathbf{U}^r)\}, \\ \overline{\mathcal{M}}^\perp &:= \{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \text{row}(\Theta) \perp \text{col}(\mathbf{V}^r), \text{col}(\Theta) \perp \text{col}(\mathbf{U}^r)\}, \end{aligned} \quad (\text{A.2})$$

where  $\text{col}(\cdot)$  and  $\text{row}(\cdot)$  denote the column space and row space respectively. For any  $\Delta \in \mathbb{R}^{d_1 \times d_2}$  and Hilbert space  $\mathcal{W} \subseteq \mathbb{R}^{d_1 \times d_2}$ , let  $\Delta_{\mathcal{W}}$  be the projection of  $\Delta$  onto  $\mathcal{W}$ . We first clarify here what  $\Delta_{\mathcal{M}}$ ,  $\Delta_{\overline{\mathcal{M}}}$  and  $\Delta_{\overline{\mathcal{M}}^\perp}$  are. Write  $\Delta$  as

$$\Delta = [\mathbf{U}^r, \mathbf{U}^{r^\perp}] \begin{bmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22} \end{bmatrix} [\mathbf{V}^r, \mathbf{V}^{r^\perp}]^T,$$

then the following equalities hold:

$$\Delta_{\mathcal{M}} = \mathbf{U}^r \mathbf{\Gamma}_{11} (\mathbf{V}^r)^T, \quad \Delta_{\overline{\mathcal{M}}^\perp} = \mathbf{U}^{r^\perp} \mathbf{\Gamma}_{22} (\mathbf{V}^{r^\perp})^T, \quad \Delta_{\overline{\mathcal{M}}} = [\mathbf{U}^r, \mathbf{U}^{r^\perp}] \begin{bmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & \mathbf{0} \end{bmatrix} [\mathbf{V}^r, \mathbf{V}^{r^\perp}]^T. \quad (\text{A.3})$$

Applying Lemma 1 in Negahban et al. (2012) to our new loss function implies that if  $\lambda_N \geq 2\|\hat{\Sigma}_{Y\mathbf{X}} - \text{mat}(\hat{\Sigma}_{\mathbf{X}\mathbf{X}}\text{vec}(\Theta^*))\|_{op}$ , it holds that

$$\|\hat{\Delta}_{\mathcal{M}^\perp}\|_N \leq 3\|\hat{\Delta}_{\mathcal{M}}\|_N + 4 \sum_{j \geq r+1} \sigma_j(\Theta^*). \quad (\text{A.4})$$

Note that  $\text{rank}(\hat{\Delta}_{\mathcal{M}}) \leq 2r$ ; we thus have

$$\|\hat{\Delta}\|_N \leq \|\hat{\Delta}_{\mathcal{M}}\|_N + \|\hat{\Delta}_{\mathcal{M}^\perp}\|_N \leq 4\|\hat{\Delta}_{\mathcal{M}}\|_N + 4 \sum_{j \geq r+1} \sigma_j(\Theta^*) \leq 4\sqrt{2r}\|\hat{\Delta}\|_F + 4 \sum_{j \geq r+1} \sigma_j(\Theta^*). \quad (\text{A.5})$$

Following the proof of Corollary 2 in Negahban and Wainwright (2011), we determine the value of  $r$  here. For a threshold  $\tau > 0$ , we choose

$$r = \#\{j \in \{1, 2, \dots, (d_1 \wedge d_2)\} | \sigma_j(\Theta^*) \geq \tau\}.$$

Then it follows that

$$\sum_{j \geq r+1} \sigma_j(\Theta^*) \leq \tau \sum_{j \geq r+1} \frac{\sigma_j(\Theta^*)}{\tau} \leq \tau \sum_{j \geq r+1} \left(\frac{\sigma_j(\Theta^*)}{\tau}\right)^q \leq \tau^{1-q} \sum_{j \geq r+1} \sigma_j(\Theta^*)^q \leq \tau^{1-q} \rho. \quad (\text{A.6})$$

On the other hand,  $\rho \geq \sum_{j \leq r} \sigma_j(\Theta^*)^q \geq r\tau^q$ , so  $r \leq \rho\tau^{-q}$ . Combining (A.1), (A.5) and  $\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{\mathbf{X}\mathbf{X}} \text{vec}(\hat{\Delta}) \geq \kappa_{\mathcal{L}} \|\hat{\Delta}\|_F^2$ , we have

$$\frac{1}{2} \kappa_{\mathcal{L}} \|\hat{\Delta}\|_F^2 \leq 2\lambda_N (4\sqrt{2r}\|\hat{\Delta}\|_F + 4\tau^{1-q}\rho),$$

which implies that

$$\|\hat{\Delta}\|_F \leq 4\sqrt{\frac{\lambda_N \rho}{\kappa_{\mathcal{L}}}} \left( \sqrt{\frac{32\lambda_N \tau^{-q}}{\kappa_{\mathcal{L}}}} + \sqrt{\tau^{1-q}} \right).$$

Choosing  $\tau = \lambda_N / \kappa_{\mathcal{L}}$ , we have for some constant  $C_1$ ,

$$\|\hat{\Delta}\|_F \leq C_1 \sqrt{\rho} \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{1-\frac{q}{2}}.$$

Combining this result with (A.5) and (A.6), we can further derive the statistical error rate in terms of the nuclear norm as follows.

$$\|\widehat{\Delta}\|_N \leq C_2 \rho \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{1-q},$$

where  $C_2$  is certain positive constant.  $\square$

*Proof of Lemma 1.* (a) We first prove for the case of the sub-Gaussian design. Recall that we use  $\widetilde{Y}_i$  to denote  $\text{sgn}(Y_i)(|Y_i| \wedge \tau)$  and  $\widetilde{\mathbf{x}}_i = \mathbf{x}_i$  in this case. Let  $\widehat{\sigma}_{x_j \widetilde{Y}}(\tau) = \frac{1}{N} \sum_{i=1}^N \widetilde{Y}_i x_{ij}$ . Note that

$$\text{Var}(\widetilde{Y}_i x_{ij}) \leq \mathbb{E}(\widetilde{Y}_i^2 x_{ij}^2) \leq \mathbb{E}(Y_i^2 x_{ij}^2) \leq (\mathbb{E} Y_i^{2k})^{\frac{1}{k}} (\mathbb{E} x_{ij}^{\frac{2k}{k-1}})^{\frac{k-1}{k}} \leq 2M^{\frac{1}{k}} \kappa_0^2 k / (k-1) < \infty, \quad (\text{A.7})$$

which is a constant that we denote by  $v_1$ . In addition, for  $p > 2$ ,

$$\mathbb{E} |\widetilde{Y}_i x_{ij}|^p \leq \tau^{p-2} \mathbb{E}(Y_i^2 |x_{ij}|^p) \leq \tau^{p-2} M^{\frac{1}{k}} (\mathbb{E} |x_{ij}|^{\frac{kp}{k-1}})^{\frac{k-1}{k}} \leq \tau^{p-2} M^{\frac{1}{k}} \left( \kappa_0 \sqrt{\frac{kp}{k-1}} \right)^p.$$

By the Jensen's inequality and then Stirling approximation, it follows that for some constants  $c_1$  and  $c_2$ ,

$$\begin{aligned} \mathbb{E} |\widetilde{Y}_i x_{ij} - \mathbb{E} \widetilde{Y}_i x_{ij}|^p &\leq 2^{p-1} (\mathbb{E} |\widetilde{Y}_i x_{ij}|^p + |\mathbb{E} \widetilde{Y}_i x_{ij}|^p) \leq 2^{p-1} (\mathbb{E} |Y_i x_{ij}|^p + (\mathbb{E} |Y_i x_{ij}|)^p) \\ &\leq \frac{c_1 p! (c_2 \tau)^{p-2}}{2}. \end{aligned}$$

Define  $v := c_1 \vee v_1$ . According to Bernstein's Inequality (Theorem 2.10 in Boucheron et al. (2013)), we have for  $j = 1, \dots, d$ ,

$$P\left(|\widehat{\sigma}_{x_j \widetilde{Y}}(\tau) - \mathbb{E}(\widetilde{Y}_i x_{ij})| \geq \sqrt{\frac{2vt}{N}} + \frac{c_2 \tau t}{N}\right) \leq 2 \exp(-t).$$

Also note that by Markov's inequality,

$$\mathbb{E}((\widetilde{Y}_i - Y_i)x_{ij}) \leq \mathbb{E}(|Y_i x_{ij}| \cdot 1_{|Y_i| > \tau}) \leq \sqrt{\mathbb{E}(Y_i^2 x_{ij}^2) P(|Y_i| > \tau)} \leq \sqrt{\frac{v \mathbb{E} Y_i^2}{\tau^2}} \leq \frac{\sqrt{v M^{\frac{1}{k}}}}{\tau}.$$

Note that since  $\text{Var}(Y_i) = \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}_{\mathbf{xx}} \boldsymbol{\beta}^* + \mathbb{E} \epsilon^2 \leq M^{\frac{1}{k}}$  and  $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{xx}}) \geq \kappa_{\mathcal{L}}$ ,  $\|\boldsymbol{\beta}^*\|_2 \leq M^{\frac{1}{k}} / \kappa_{\mathcal{L}}$ .

Therefore,  $\|\mathbf{x}_i^T \boldsymbol{\beta}^*\|_{\psi_2} \leq M^{\frac{1}{k}} \kappa_0 / \kappa_{\mathcal{L}}$  and  $\|x_{ij} \mathbf{x}_i^T \boldsymbol{\beta}^*\|_{\psi_1} \leq 2M^{\frac{1}{k}} \kappa_0^2 / \kappa_{\mathcal{L}}$ . By Proposition 5.16 (Bernstein-type inequality) in Vershynin (2010), we have for sufficiently small  $t$ ,

$$P\left(|(\widehat{\boldsymbol{\Sigma}}_{x_j \mathbf{x}})^T \boldsymbol{\beta}^* - \mathbb{E}(Y_i x_{ij})| \geq c_2 \sqrt{\frac{t}{N}}\right) \leq \exp(-t)$$

for some constant  $c_2$ . Choose  $\tau \asymp \sqrt{N/\log d}$ . An application of the triangle inequality and the union bound yields that as long as  $\log d/N < \gamma_1$  for certain  $\gamma_1$ , we have for some constant  $\nu_1 > 0$ ,

$$P\left(\|\widehat{\boldsymbol{\Sigma}}_{Y\mathbf{x}}(\tau) - \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i^T \boldsymbol{\theta}^*) \mathbf{x}_i\|_{\max} \geq \nu_1 \sqrt{\frac{\delta \log d}{N}}\right) \leq 2d^{1-\delta}.$$

(b) Now we switch to the case where both the noise and the design have only bounded moments. Note that

$$\begin{aligned} \|\widehat{\boldsymbol{\Sigma}}_{\widetilde{Y}\widetilde{\mathbf{x}}} - \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} \boldsymbol{\theta}^*\|_{\max} &\leq \|\widehat{\boldsymbol{\Sigma}}_{\widetilde{Y}\widetilde{\mathbf{x}}} - \boldsymbol{\Sigma}_{\widetilde{Y}\widetilde{\mathbf{x}}}\|_{\max} + \|\boldsymbol{\Sigma}_{\widetilde{Y}\widetilde{\mathbf{x}}} - \boldsymbol{\Sigma}_{Y\mathbf{x}}\|_{\max} + \|\boldsymbol{\Sigma}_{Y\mathbf{x}} - \boldsymbol{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} \boldsymbol{\theta}^*\|_{\max} \\ &\quad + \|(\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} - \boldsymbol{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}) \boldsymbol{\theta}^*\|_{\max} = T_1 + T_2 + T_3 + T_4. \end{aligned}$$

We bound the four terms one by one. For  $1 \leq j \leq d$ , analogous to (A.7),

$$\text{Var}(\widetilde{Y}_i \widetilde{x}_{ij}) \leq \mathbb{E}(\widetilde{Y}_i \widetilde{x}_{ij})^2 \leq \mathbb{E}(Y_i x_{ij})^2 \leq \sqrt{\mathbb{E} Y_i^4 \mathbb{E} x_{ij}^4} =: v_1 < \infty.$$

In addition,  $\mathbb{E} |\widetilde{Y}_i \widetilde{x}_{ij}|^p \leq (\tau_1 \tau_2)^{p-2} v_1$ . Therefore according to Bernstein's Inequality (Theorem 2.10 in Boucheron et al. (2013)), we have

$$P\left(|\widehat{\sigma}_{\widetilde{Y}\widetilde{x}_j} - \sigma_{\widetilde{Y}\widetilde{x}_j}| \geq \sqrt{\frac{2v_1 t}{N}} + \frac{c\tau_1 \tau_2 t}{N}\right) \leq \exp(-t),$$

where  $\widehat{\sigma}_{\widetilde{Y}\widetilde{x}_j} = \frac{1}{N} \sum_{i=1}^N \widetilde{Y}_i \widetilde{x}_{ij}$ ,  $\sigma_{\widetilde{Y}\widetilde{x}_j} = \mathbb{E} \widetilde{Y}_i \widetilde{x}_{ij}$  and  $c$  is certain constant. Then by the union bound, we have

$$P\left(|T_1| > \sqrt{\frac{2v_1 t}{N}} + \frac{c\tau_1 \tau_2 t}{N}\right) \leq d \exp(-t).$$

Next we bound  $T_2$ . Note that for  $1 \leq j \leq d$ ,

$$\begin{aligned} \mathbb{E} \tilde{Y}_i \tilde{x}_{ij} - \mathbb{E} Y_i x_{ij} &= \mathbb{E} \tilde{Y}_i \tilde{x}_{ij} - \mathbb{E} \tilde{Y}_i x_{ij} + \mathbb{E} \tilde{Y}_i x_{ij} - \mathbb{E} Y_i x_{ij} = \mathbb{E} \tilde{Y}_i (\tilde{x}_{ij} - x_{ij}) + \mathbb{E} (\tilde{Y}_i - Y_i) x_{ij} \\ &\leq \sqrt{\mathbb{E} (Y_i^2 (\tilde{x}_{ij} - x_{ij})^2) P(|x_{ij}| \geq \tau_2)} + \sqrt{\mathbb{E} ((\tilde{Y}_i - Y_i)^2 x_{ij}^2) P(|Y_i| \geq \tau_1)} \\ &\leq \sqrt{M v_1} \left( \frac{1}{\tau_2^2} + \frac{1}{\tau_1^2} \right), \end{aligned}$$

which delivers that  $T_2 \leq \sqrt{M v_1} (1/\tau_1^2 + 1/\tau_2^2)$ . Then we bound  $T_3$ . For  $1 \leq j \leq d$ ,

$$\begin{aligned} \mathbb{E} ((\tilde{\mathbf{x}}_i^T \boldsymbol{\theta}^*) \tilde{x}_{ij}) - \mathbb{E} ((\mathbf{x}_i^T \boldsymbol{\theta}^*) x_{ij}) &= \sum_{k=1}^d \mathbb{E} (\theta_k^* (\tilde{x}_{ik} \tilde{x}_{ij} - x_{ik} x_{ij})) \leq \sum_{k=1}^d |\theta_k^*| \mathbb{E} |\tilde{x}_{ij} \tilde{x}_{ik} - x_{ij} x_{ik}| \\ &\leq \sum_{k=1}^d |\theta_k^*| (\mathbb{E} |x_{ij} (\tilde{x}_{ik} - x_{ik})| + \mathbb{E} |(\tilde{x}_{ij} - x_{ij}) x_{ik}|) \\ &\leq \sum_{k=1}^d |\theta_k^*| \left( \mathbb{E} (|x_{ij} (\tilde{x}_{ik} - x_{ik})| 1_{\{|x_{ik}| > \tau_2\}}) + \mathbb{E} (|(\tilde{x}_{ij} - x_{ij}) x_{ik}| 1_{\{|x_{ij}| > \tau_2\}}) \right) \leq \frac{CR}{\tau_2^2}. \end{aligned}$$

Finally we bound  $T_4$ . For  $1 \leq j, k \leq d$ , we have  $|\tilde{x}_{ik} \tilde{x}_{ij}| \leq \tau_2^2$  and  $\text{Var}(\tilde{x}_{ij} \tilde{x}_{ik}) \leq M =: v_2$ . Therefore according to Bernstein's inequality (Theorem 2.10 in Boucheron et al. (2013)),

$$P\left(|\hat{\sigma}_{\tilde{x}_j \tilde{x}_k} - \sigma_{\tilde{x}_j \tilde{x}_k}| \geq \sqrt{\frac{2v_2 t}{N}} + \frac{c\tau_2^2 t}{N}\right) \leq \exp(-t), \quad (\text{A.8})$$

where  $\hat{\sigma}_{\tilde{x}_j \tilde{x}_k} = (1/N) \sum_{i=1}^N \tilde{x}_{ij} \tilde{x}_{ik}$ ,  $\sigma_{\tilde{x}_j \tilde{x}_k} = \mathbb{E} \tilde{x}_{ij} \tilde{x}_{ik}$  and  $c$  is certain constant. We therefore have

$$P\left(|T_4| \geq R \left( \sqrt{\frac{2v_2 t}{N}} + \frac{c\tau_2^2 t}{N} \right)\right) \leq d^2 \exp(-t).$$

Now we choose  $\tau_1, \tau_2 \asymp (N/\log d)^{1/4}$ , then it follows that for some constant  $\nu_2 > 0$ ,

$$P\left(\|\hat{\boldsymbol{\Sigma}}_{\tilde{\mathbf{Y}}\tilde{\mathbf{X}}} - \hat{\boldsymbol{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \boldsymbol{\theta}^*\|_{\max} > \nu_2 \sqrt{\frac{\delta \log d}{N}}\right) \leq 2d^{1-\delta}.$$

□

*Proof of Lemma 2.* **(a)** We first consider the sub-Gaussian design. Again, since  $\tilde{\mathbf{x}}_i = \mathbf{x}_i$  in this case, we do not add any tilde above  $\mathbf{x}_i$  in this proof. Define  $\hat{\mathbf{D}}_{\mathbf{xx}}$  to be the diagonal of



$\widehat{\Sigma}_{\mathbf{xx}}$ . Note that given  $\|\mathbf{x}_i\|_{\psi_2} \leq \kappa_0$  and  $\lambda_{\min}(\Sigma_{\mathbf{xx}}) \geq \kappa_{\mathcal{L}} > 0$ , it follows that

$$\forall \mathbf{v} \in \mathcal{S}^{d-1}, \sqrt{\mathbb{E}(\mathbf{v}^T \mathbf{x}_i)^4} \leq 4\kappa_0^2 \leq (4\kappa_0^2/\kappa_{\mathcal{L}}) \mathbf{v}^T \Sigma_{\mathbf{xx}} \mathbf{v}.$$

In addition, since  $\|x_{ij}\|_{\psi_2} \leq \kappa_0$ ,  $(\mathbb{E} x_{ij}^8)^{\frac{1}{4}} \leq 8\kappa_0^2 \leq (8\kappa_0^2/\kappa_{\mathcal{L}}) \mathbb{E} x_{ij}^2$ . According to Lemma 5.2 in Oliveira (2016), for any  $\eta_1 > 1$ ,

$$P\left(\forall \mathbf{v} \in \mathbb{R}^d : \mathbf{v}^T \widehat{\Sigma}_{\mathbf{xx}} \mathbf{v} \geq \frac{1}{2} \mathbf{v}^T \Sigma_{\mathbf{xx}} \mathbf{v} - \frac{c_1(1 + 2\eta_1 \log d)}{N} \|\widehat{\mathbf{D}}_{\mathbf{xx}}^{1/2} \mathbf{v}\|_1^2\right) \geq 1 - \frac{d^{1-\eta_1}}{3}, \quad (\text{A.9})$$

where  $c_1$  is some universal constant. For any  $1 \leq j \leq d$ , we know that  $\|x_{ij}\|_{\psi_1} \leq 2\|x_{ij}\|_{\psi_2}^2 = 2\kappa_0^2$ , therefore by the Bernstein-type inequality we have for sufficiently small  $t$ ,

$$P\left(\left|\frac{1}{N} \sum_{i=1}^n x_{ij}^2 - \mathbb{E} x_{ij}^2\right| \geq t\right) \leq 2 \exp(-cNt^2),$$

where  $c$  depends on  $\kappa_0$ . Note that  $\mathbb{E} x_{ij}^2 \leq 2\kappa_0^2$ . An application of the union bound delivers that when  $t$  is sufficiently small,

$$P(\|\widehat{\mathbf{D}}_{\mathbf{xx}}\|_{op} \geq 2\kappa_0^2 + t) \leq 2d \exp(-cNt^2).$$

Let  $t < \kappa_0^2$ , then the inequality above yields that for a new constant  $C_2 > 0$ ,

$$P(\|\widehat{\mathbf{D}}_{\mathbf{xx}}\|_{op} \geq 3\kappa_0^2) \leq 2d \exp(-C_2N).$$

Combining the inequality above with (A.9), it follows for a new constant  $C_1 > 0$  that

$$P\left(\forall \mathbf{v} \in \mathbb{R}^d : \mathbf{v}^T \widehat{\Sigma}_{\mathbf{xx}} \mathbf{v} \geq \frac{1}{2} \mathbf{v}^T \Sigma_{\mathbf{xx}} \mathbf{v} - \frac{C_1(1 + 2\eta_1 \log d)}{N} \|\mathbf{v}\|_1^2\right) \geq 1 - \frac{d^{1-\eta_1}}{3} - 2d \exp(-C_2N).$$

**(b)** Now we switch to the case of designs with only bounded moments. We will show that the sample covariance of the truncated design has the restricted eigenvalue property. Recall that for  $i = 1, \dots, N, j = 1, \dots, d$ ,  $\tilde{x}_{ij} = \text{sgn}(x_{ij})(|x_{ij}| \wedge \tau_2)$ . Note that

$$\mathbf{v}^T \widehat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \mathbf{v} = \mathbf{v}^T (\widehat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} - \Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}) \mathbf{v} + \mathbf{v}^T (\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} - \Sigma_{\mathbf{xx}}) \mathbf{v} + \mathbf{v}^T \Sigma_{\mathbf{xx}} \mathbf{v}.$$

According to (A.8), we have for some  $c_1 > 0$ ,

$$P\left(\|\widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} - \Sigma_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}\|_{\max} \geq \sqrt{\frac{2Mt}{N}} + \frac{c_1\tau_2^2 t}{N}\right) \leq d^2 \exp(-t).$$

Given  $\tau_2 \asymp (N/\log d)^{1/4}$ , we have for some constant  $c_2 > 0$  and any  $\eta_2 > 2$ ,

$$P\left(\|\widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} - \Sigma_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}\|_{\max} \geq c_2 \sqrt{\frac{\eta_2 \log d}{N}}\right) \leq d^{2-\eta_2}.$$

In addition, for any  $1 \leq j_1, j_2 \leq d$ , we have

$$\begin{aligned} \mathbb{E} x_{ij_1} x_{ij_2} - \mathbb{E} \widetilde{x}_{ij_1} \widetilde{x}_{ij_2} &\leq \mathbb{E}(|x_{ij_1} x_{ij_2}| (1_{\{|x_{ij_1}| \geq \tau_2\}} + 1_{\{|x_{ij_2}| \geq \tau_2\}})) \\ &\leq \sqrt{M}(\sqrt{P(|x_{ij_1}| \geq \tau_2)} + \sqrt{P(|x_{ij_2}| \geq \tau_2)}) \\ &\leq \sqrt{M}\left(\sqrt{\frac{\mathbb{E} x_{ij_1}^4}{\tau_2^4}} + \sqrt{\frac{\mathbb{E} x_{ij_2}^4}{\tau_2^4}}\right) \leq c_3 \sqrt{\frac{\log d}{N}} \end{aligned}$$

for some  $c_3 > 0$ , which implies that  $\|\Sigma_{\mathbf{xx}} - \Sigma_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}\|_{\max} = O_P(\sqrt{\log d/N})$ . Therefore we have

$$P\left(\forall \mathbf{v} \in \mathbb{R}^d, \mathbf{v}^T \widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} \mathbf{v} \geq \mathbf{v}^T \Sigma_{\mathbf{xx}} \mathbf{v} - (c_2 \eta_2 + c_3) \sqrt{\frac{\log d}{N}} \|\mathbf{v}\|_1^2\right) \leq d^{2-\eta_2}.$$

□

Proof of Theorem 2. Suppose  $\|\widehat{\Delta}\|_2 \geq c_1 \sqrt{\rho} \left(\frac{\lambda_N}{\kappa_{\mathcal{L}}}\right)^{1-\frac{q}{2}}$  for some  $c_1 > 0$ , then by (A.5) and (A.6),

$$\begin{aligned} \|\widehat{\Delta}\|_1 &\leq 4\sqrt{2r} \|\widehat{\Delta}\|_2 + 4\tau^{1-q} \rho \leq 4\sqrt{2} \kappa_{\mathcal{L}}^{\frac{q}{2}} \sqrt{\rho} \lambda_N^{-\frac{q}{2}} \|\widehat{\Delta}\|_2 + 4\kappa_{\mathcal{L}}^{q-1} \rho \lambda_N^{1-q} \\ &\leq (4\sqrt{2} + 4c_1^{-1}) \sqrt{\rho} \kappa_{\mathcal{L}}^{\frac{q}{2}} \lambda_N^{-\frac{q}{2}} \|\widehat{\Delta}\|_2. \end{aligned} \tag{A.10}$$

For any  $\delta > 1$ , since  $\lambda_N = 2\nu \sqrt{\delta \log d/N}$ , it is easy to verify that as long as  $\rho(\log d/N)^{1-\frac{q}{2}} \leq C_1$  for some constant  $C_1$ , we will have by (A.10) and (3.3) in Lemma 2

$$P\left(\widehat{\Delta}^T \widehat{\Sigma}_{\mathbf{xx}} \widehat{\Delta} \geq \frac{\kappa_{\mathcal{L}}}{4} \|\widehat{\Delta}\|_2^2\right) \geq 1 - d^{1-\delta}.$$

An application of Theorem 1 delivers that for constants  $c_2, c_3 > 0$ ,

$$\|\widehat{\Delta}\|_2 \leq c_2 \sqrt{\rho} \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{1-\frac{q}{2}} \quad \text{and} \quad \|\widehat{\Delta}\|_1 \leq c_3 \rho \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{\frac{1-q}{2}}.$$

When  $\|\widehat{\Delta}\|_2 \leq c_1 \sqrt{\rho} \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{1-\frac{q}{2}}$ , we can still obtain the  $\ell_1$  norm bound that  $\|\widehat{\Delta}\|_1 \leq c_4 \rho \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{\frac{1-q}{2}}$  for some constant  $c_4$  through (A.5) and (A.6). Overall, we can achieve the conclusion for some constants  $C_2$  and  $C_3$  that with probability at least  $1 - 3d^{1-\delta}$ ,

$$\|\widehat{\Delta}\|_2^2 \leq C_2 \rho \left( \frac{\delta \log d}{N} \right)^{1-\frac{q}{2}} \quad \text{and} \quad \|\widehat{\Delta}\|_1 \leq C_3 \rho \left( \frac{\delta \log d}{N} \right)^{\frac{1-q}{2}}.$$

When the design only satisfies bounded moment conditions, again we first assume that  $\|\widehat{\Delta}\|_2 \geq c_5 \sqrt{\rho} (\lambda_N / \kappa_{\mathcal{L}})^{1-\frac{q}{2}}$  for some constant  $c_5$ . Analogous to the case of the sub-Gaussian design, it is easy to verify that for any  $\delta > 1$ , as long as  $\rho (\log d / N)^{\frac{1-q}{2}} \leq C_4$  for some constant  $C_4$ , we will have by (A.10) and (3.3) in Lemma 2

$$P\left(\widehat{\Delta}^T \widehat{\Sigma}_{\mathbf{xx}} \widehat{\Delta} \geq \frac{\kappa_{\mathcal{L}}}{2} \|\widehat{\Delta}\|_2^2\right) \geq 1 - d^{1-\delta},$$

An application of Lemma 1 delivers that for some constants  $c_6, c_7 > 0$ ,

$$\|\widehat{\Delta}\|_2 \leq c_7 \sqrt{\rho} \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{1-\frac{q}{2}} \quad \text{and} \quad \|\widehat{\Delta}\|_1 \leq c_8 \sqrt{\rho} \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{\frac{1-q}{2}}.$$

When  $\|\widehat{\Delta}\|_2 \leq c_5 \sqrt{\rho} (\lambda_N / \kappa_{\mathcal{L}})^{1-\frac{q}{2}}$ , we have exactly the same steps as those in the case of the sub-Gaussian design.  $\square$

Proof of Lemma 3. Recall that  $\widetilde{Y}_i = \text{sgn}(Y_i)(|Y_i| \wedge \tau)$ , then

$$\begin{aligned} & \|\widehat{\Sigma}_{Y\mathbf{X}}(\tau) - \frac{1}{N} \sum_{i=1}^N \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle \mathbf{X}_i\|_{op} \\ & \leq \|\widehat{\Sigma}_{Y\mathbf{X}}(\tau) - \mathbb{E}(\widetilde{Y}\mathbf{X})\|_{op} + \|\mathbb{E}((\widetilde{Y} - Y)\mathbf{X})\|_{op} + \left\| \frac{1}{n} \sum_{i=1}^n \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle \mathbf{X}_i - \mathbb{E} Y \mathbf{X} \right\|_{op}. \end{aligned} \tag{A.11}$$

Next we use the covering argument to bound each term of the RHS. Let  $\mathcal{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|u\|_2 = 1\}$ ,  $\mathcal{N}^{d-1}$  be the  $1/4$ -net on  $\mathcal{S}^{d-1}$  and  $\Phi(\mathbf{A}) = \sup_{\mathbf{u} \in \mathcal{N}^{d_1-1}, \mathbf{v} \in \mathcal{N}^{d_2-1}} \mathbf{u}^T \mathbf{A} \mathbf{v}$  for any

matrix  $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$ , then we claim

$$\|\mathbf{A}\|_{op} \leq \frac{16}{7} \Phi(\mathbf{A}). \quad (\text{A.12})$$

To establish the claim above, note that by the definition of the  $1/4$ -net, for any  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ , there exist  $\mathbf{u}_1 \in \mathcal{N}^{d_1-1}$  and  $\mathbf{v}_1 \in \mathcal{N}^{d_2-1}$  such that  $\|\mathbf{u} - \mathbf{u}_1\|_2 \leq 1/4$  and  $\|\mathbf{v} - \mathbf{v}_1\|_2 \leq 1/4$ . Then it follows that

$$\begin{aligned} \mathbf{u}^T \mathbf{A} \mathbf{v} &= \mathbf{u}_1^T \mathbf{A} \mathbf{v}_1 + (\mathbf{u} - \mathbf{u}_1)^T \mathbf{A} \mathbf{v}_1 + \mathbf{u}_1^T \mathbf{A} (\mathbf{v} - \mathbf{v}_1) + (\mathbf{u} - \mathbf{u}_1)^T \mathbf{A} (\mathbf{v} - \mathbf{v}_1) \\ &\leq \Phi(\mathbf{A}) + \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{16}\right) \sup_{\mathbf{u} \in \mathcal{S}^{d_1-1}, \mathbf{v} \in \mathcal{S}^{d_2-1}} \mathbf{u}^T \mathbf{A} \mathbf{v}. \end{aligned}$$

Taking the superlative over  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$  on the LHS yields (A.12).

Now fix  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ . Later on we always write  $\mathbf{u}^T \mathbf{X}_i \mathbf{v}$  as  $Z_i$  and  $\mathbf{u}^T \mathbf{X} \mathbf{v}$  as  $Z$  for convenience. Consider

$$\mathbf{u}^T (\widehat{\Sigma}_{Y\mathbf{X}}(\tau) - \mathbb{E}(\widetilde{Y}\mathbf{X}))\mathbf{v} = \frac{1}{n} \sum_{i=1}^n \widetilde{Y}_i Z_i - \mathbb{E}(\widetilde{Y}Z).$$

Note that

$$\begin{aligned} \mathbb{E}(\widetilde{Y}_i Z_i - \mathbb{E} \widetilde{Y}_i Z_i)^2 &\leq \mathbb{E}(\widetilde{Y}_i Z_i)^2 \leq \mathbb{E}(Y_i Z_i)^2 = \mathbb{E}(Z_i^2 \mathbb{E}(Y_i^2 | \mathbf{X}_i)) = \mathbb{E}\left(Z_i^2 (\langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle)^2\right) + \mathbb{E}\left(Z_i^2 \mathbb{E}(\epsilon_i^2 | \mathbf{X}_i)\right) \\ &\leq 16R^2 \kappa_0^4 + (2k\kappa_0^2/(k-1)) \left(\mathbb{E}(\mathbb{E}(\epsilon_i^2 | \mathbf{X}_i))^k\right)^{\frac{1}{k}} \leq 16R^2 \kappa_0^4 + 2k\kappa_0^2 M^{\frac{1}{k}}/(k-1) < \infty, \end{aligned} \quad (\text{A.13})$$

which we denote by  $v_1$  for convenience. Also we have

$$\begin{aligned} \mathbb{E} |\widetilde{Y}_i Z_i|^p &\leq \tau^{p-2} \mathbb{E}(Y_i^2 |Z_i|^p) = \tau^{p-2} \mathbb{E}(\langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle^2 |Z_i|^p + \epsilon_i^2 |Z_i|^p) \\ &\leq \tau^{p-2} \left( \sqrt{\mathbb{E} \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle^4 \mathbb{E} |Z_i|^{2p}} + \left(\mathbb{E}(\mathbb{E}(\epsilon_i^2 | \mathbf{X}_i)^k)\right)^{\frac{1}{k}} \left(\mathbb{E} x_{ij}^{\frac{pk}{k-1}}\right)^{1-\frac{1}{k}} \right) \\ &\leq \tau^{p-2} \left( 4R^2 \kappa_0^2 (\kappa_0 \sqrt{2p})^p + M^{\frac{1}{k}} (\kappa_0 \sqrt{\frac{pk}{k-1}})^p \right) \end{aligned}$$

and it holds for constants  $c_1$  and  $c_2$  that

$$\begin{aligned} \mathbb{E} |\tilde{Y}_i Z_i - \mathbb{E} \tilde{Y}_i Z_i|^p &\leq 2^{p-1} (\mathbb{E} |\tilde{Y}_i Z_i|^p + |\mathbb{E} \tilde{Y}_i Z_i|^p) \leq 2^{p-1} (\mathbb{E} |\tilde{Y}_i Z_i|^p + (\mathbb{E} |Y_i Z_i|)^p) \\ &\leq c_1 p! (c_2 \tau)^{p-2}. \end{aligned} \quad (\text{A.14})$$

where the last inequality uses the Stirling approximation of  $p!$ . Define  $v := c_1 \vee v_1$ . Then an application of Bernstein's Inequality (Theorem 2.10 in Boucheron et al. (2013)) to  $\{\tilde{Y}_i Z_i\}_{i=1}^N$  delivers

$$P\left(\left|\frac{1}{N} \sum_{i=1}^N \tilde{Y}_i Z_i - \mathbb{E} \tilde{Y}_i Z_i\right| > \sqrt{\frac{2vt}{N}} + \frac{c_2 \tau t}{N}\right) \leq \exp(-t).$$

By taking the union bound over all  $(\mathbf{u}, \mathbf{v}) \in \mathcal{N}^{d_1-1} \times \mathcal{N}^{d_2-1}$  and (A.12) it follows that

$$P\left(\|\hat{\Sigma}_{\tilde{Y}\mathbf{X}}(\tau) - \mathbb{E}(\tilde{Y}\mathbf{X})\|_{op} \geq \frac{16}{7} \left(\sqrt{\frac{2vt}{N}} + \frac{c_2 \tau t}{N}\right)\right) \leq \exp((d_1 + d_2) \log 8 - t), \quad (\text{A.15})$$

where  $c_1$  is a constant.

Next we aim to bound  $\|\mathbb{E}((\tilde{Y} - Y)\mathbf{X})\|_{op}$ . For any  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ , by the Cauchy-Schwartz inequality and the Markov inequality,

$$\mathbb{E}((\tilde{Y} - Y)Z) \leq \mathbb{E}(|YZ| \cdot 1_{|Y|>\tau}) \leq \sqrt{\mathbb{E}(Y^2 Z^2) P(|Y| > \tau)} \leq \sqrt{\frac{v \mathbb{E} Y^2}{\tau^2}} \leq \frac{\sqrt{v(2R^2 \kappa_0^2 + M^{\frac{1}{k}})}}{\tau}.$$

Note that the inequality above holds for any  $(\mathbf{u}, \mathbf{v})$ . Therefore

$$\|\mathbb{E}((\tilde{Y} - Y)\mathbf{X})\|_{op} \leq \frac{\sqrt{v(2R^2 \kappa_0^2 + M^{\frac{1}{k}})}}{\tau}. \quad (\text{A.16})$$

Now we give an upper bound of the third term on the RHS of (A.11). For any  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ ,  $\|\langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle Z_i\|_{\psi_1} \leq R\kappa_0^2$ , so by Proposition 5.16 (Bernstein-type inequality) in Vershynin (2010) it follows that for sufficiently small  $t$ ,

$$P\left(\left|\frac{1}{N} \sum_{i=1}^N \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle Z_i - \mathbb{E} Y Z\right| > t\right) \leq 2 \exp\left(-\frac{c_3 N t^2}{R^2 \kappa_0^4}\right),$$

where  $c_3$  is a constant. Then an combination of the union bound over all points on  $\mathcal{N}^{d_1-1} \times$

$\mathcal{N}^{d_2-1}$  and (A.12) delivers

$$P\left(\left\|\frac{1}{N}\sum_{i=1}^N\langle\mathbf{X}_i,\boldsymbol{\Theta}^*\rangle\mathbf{X}_i-\mathbb{E}Y\mathbf{X}\right\|_{op}\geq\frac{16}{7}t\right)\leq 2\exp\left((d_1+d_2)\log 8-\frac{c_3Nt^2}{R^2\kappa_0^4}\right), \quad (\text{A.17})$$

Finally we choose  $\tau \asymp \sqrt{N/(d_1+d_2)}$ . Combining (A.15), (A.16) and (A.17), we can find a constant  $\gamma > 0$  such that as long as  $(d_1+d_2)/N < \gamma$ , it holds that

$$P\left(\left\|\widehat{\boldsymbol{\Sigma}}_{Y\mathbf{X}}(\tau)-\frac{1}{N}\sum_{i=1}^N\langle\mathbf{X}_i,\boldsymbol{\Theta}^*\rangle\mathbf{X}_i\right\|_{op}>\nu\sqrt{\frac{d_1+d_2}{N}}\right)\leq\eta\exp(-(d_1+d_2)),$$

where  $\nu$  and  $\eta$  are constants. □

Proof of Theorem 3. We first verify the RSC property. According to Proposition 1 in Negahban and Wainwright (2011), the following inequality holds for all  $\boldsymbol{\Delta} \in \mathbb{R}^{d_1 \times d_2}$  with probability at least  $1 - 2\exp(-N/32)$ ,

$$\sqrt{\text{vec}(\boldsymbol{\Delta})^T \overline{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{X}} \text{vec}(\boldsymbol{\Delta})} \geq \frac{1}{4} \|\sqrt{\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{X}}} \text{vec}(\boldsymbol{\Delta})\|_2 - c \left( \sqrt{\frac{d_1}{N}} + \sqrt{\frac{d_2}{N}} \right) \|\boldsymbol{\Delta}\|_N. \quad (\text{A.18})$$

Let  $\kappa_{\mathcal{L}} = (1/32)\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{X}}) > 0$ . For ease of notations, write  $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$  as  $\widehat{\boldsymbol{\Delta}}$ . Suppose  $\|\widehat{\boldsymbol{\Delta}}\|_F \geq c_1 \sqrt{\rho} \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{1-\frac{q}{2}}$  for some  $c_1 > 0$ , then by (A.5) and (A.6),

$$\begin{aligned} \|\widehat{\boldsymbol{\Delta}}\|_N &\leq 4\sqrt{2r}\|\widehat{\boldsymbol{\Delta}}\|_F + 4\tau^{1-q}\rho \leq 4\sqrt{2\rho\kappa_{\mathcal{L}}^{\frac{q}{2}}\lambda_N^{-\frac{q}{2}}}\|\widehat{\boldsymbol{\Delta}}\|_F + 4\kappa_{\mathcal{L}}^{q-1}\rho\lambda_N^{1-q} \\ &\leq (4\sqrt{2} + 4c_1^{-1})\sqrt{\rho\kappa_{\mathcal{L}}^{\frac{q}{2}}\lambda_N^{-\frac{q}{2}}}\|\widehat{\boldsymbol{\Delta}}\|_F. \end{aligned} \quad (\text{A.19})$$

Since we choose  $\lambda_N = 2\nu\sqrt{(d_1+d_2)/N}$ , there exists a constant  $C_1$  such that as long as  $\rho((d_1+d_2)/N)^{1-\frac{q}{2}} \leq C_1$ , we have by (A.18) and (A.19)

$$\text{vec}(\widehat{\boldsymbol{\Delta}})^T \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{X}} \text{vec}(\widehat{\boldsymbol{\Delta}}) \geq \kappa_{\mathcal{L}} \|\widehat{\boldsymbol{\Delta}}\|_F^2$$

with probability at least  $1 - 2\exp(-N/32)$ . An application of Theorem 1 delivers that for

some constants  $c_2, c_3 > 0$ , it holds with high probability that

$$\|\widehat{\Delta}\|_F \leq c_2 \sqrt{\rho} \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{1-\frac{q}{2}} \quad \text{and} \quad \|\widehat{\Delta}\|_N \leq c_3 \rho \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{\frac{1-q}{2}}.$$

When  $\|\widehat{\Delta}\|_F \leq c_1 \sqrt{\rho} \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{1-\frac{q}{2}}$ , we can still obtain the  $\ell_1$  norm bound that  $\|\widehat{\Delta}\|_1 \leq c_4 \rho \left( \frac{\lambda_N}{\kappa_{\mathcal{L}}} \right)^{\frac{1-q}{2}}$  through (A.5) and (A.6), where  $c_4$  is some constant. Overall, we can achieve the conclusion that,

$$\|\widehat{\Delta}\|_F^2 \leq C_2 \rho \left( \frac{d_1 + d_2}{N} \right)^{1-\frac{q}{2}} \quad \text{and} \quad \|\widehat{\Delta}\|_N \leq C_3 \rho \left( \frac{d_1 + d_2}{N} \right)^{\frac{1-q}{2}}$$

with probability at least  $1 - C_4 \exp(-(d_1 + d_2))$  for constants  $C_2, C_3$  and  $C_4$ . □

*Proof of Lemma 4.* We follow essentially the same strategies as in proof of Lemma 3. The only difference is that we do not use the covering argument to bound the first term in (A.11). Instead we apply the Matrix Bernstein inequality (Theorem 6.1.1 in Tropp (2015)) to take advantage of the singleton design under the matrix completion setting.

For any fixed  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ , write  $\mathbf{u}^T \mathbf{X}_i \mathbf{v}$  as  $Z_i$ . Then we have

$$\begin{aligned} \mathbb{E}(\widetilde{Y}_i Z_i) &= \sqrt{d_1 d_2} \mathbb{E}(\widetilde{Y}_i u_{j(i)} v_{k(i)}) = \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} \mathbb{E}(|\widetilde{Y}_i| | j(i) = j_0, k(i) = k_0) \cdot |u_{j_0} v_{k_0}| \\ &\leq \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} \mathbb{E}(|Y_i| | j(i) = j_0, k(i) = k_0) \cdot |u_{j_0} v_{k_0}| \\ &\leq \sqrt{R^2 + M^{\frac{1}{k}}} \cdot \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} |u_{j_0} v_{k_0}| \leq \sqrt{R^2 + M^{\frac{1}{k}}}. \end{aligned}$$

Since the above argument holds for all  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ , we have  $\|\mathbb{E}(\widetilde{Y}_i \mathbf{X}_i)\|_{op} \leq \sqrt{R^2 + M^{\frac{1}{k}}}$ . In addition,

$$\begin{aligned} \|\mathbb{E} \widetilde{Y}_i^2 \mathbf{X}_i^T \mathbf{X}_i\|_{op} &= d_1 d_2 \|\mathbb{E} \widetilde{Y}_i^2 \mathbf{e}_{k(i)} \mathbf{e}_{j(i)}^T \mathbf{e}_{j(i)} \mathbf{e}_{k(i)}^T\|_{op} = d_1 d_2 \|\mathbb{E} \widetilde{Y}_i^2 \mathbf{e}_{k(i)} \mathbf{e}_{k(i)}^T\|_{op} \\ &= d_1 d_2 \left\| \mathbb{E} \left( \mathbb{E}(\widetilde{Y}_i^2 | \mathbf{X}_i) \mathbf{e}_{k(i)} \mathbf{e}_{k(i)}^T \right) \right\|_{op} = \left\| \sum_{k_0=1}^{d_2} \sum_{j_0=1}^{d_1} \mathbb{E}(\widetilde{Y}_i^2 | k(i) = k_0, j(i) = j_0) \mathbf{e}_{k_0} \mathbf{e}_{k_0}^T \right\|_{op} \\ &= \max_{k_0=1, \dots, d_2} \sum_{j_0=1}^{d_1} \mathbb{E}(\widetilde{Y}_i^2 | k(i) = k_0, j(i) = j_0) \leq d_1 \mathbb{E} Y_i^2 \leq d_1 (R^2 + M^{\frac{1}{k}}) \end{aligned}$$

Similarly we can get  $\|E \tilde{Y}_i^2 \mathbf{X}_i \mathbf{X}_i^T\|_{op} \leq d_2(R^2 + M^{\frac{1}{k}})$ . Write  $\tilde{Y}_i \mathbf{X}_i - E \tilde{Y}_i \mathbf{X}_i$  as  $\mathbf{A}_i$ . Therefore by the triangle inequality,  $\max(\|E \mathbf{A}_i^T \mathbf{A}_i\|_{op}, \|E \mathbf{A}_i \mathbf{A}_i^T\|_{op}) \leq (d_1 \vee d_2)(R^2 + M^{\frac{1}{k}})$ , which we denote by  $v$  for convenience. Since  $\|\mathbf{A}_i\|_{op} \leq 2\sqrt{d_1 d_2} \tau$ , an application of the Matrix Bernstein inequality delivers,

$$P\left(\|\hat{\Sigma}_{\tilde{Y}\mathbf{X}}(\tau) - E(\tilde{Y}\mathbf{X})\|_{op} \geq t\right) \leq (d_1 + d_2) \exp\left(\frac{-Nt^2/2}{v + \sqrt{d_1 d_2} \tau t/3}\right), \quad (\text{A.20})$$

Next we aim to bound  $\|E((\tilde{Y}_i - Y_i)\mathbf{X}_i)\|_{op}$ . Fix  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ . By the Cauchy-Schwartz inequality and the Markov inequality,

$$\begin{aligned} E((\tilde{Y}_i - Y_i)Z_i) &= \sqrt{d_1 d_2} E((\tilde{Y}_i - Y_i)u_{j(i)}v_{k(i)}) \\ &= \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} E(|\tilde{Y}_i - Y_i| |j(i) = j_0, k(i) = k_0) \cdot |u_{j_0} v_{k_0}| \\ &\leq \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} E(|Y_i| \cdot 1_{|Y_i| > \tau} |j(i) = j_0, k(i) = k_0) \cdot |u_{j_0} v_{k_0}| \\ &\leq \frac{R^2 + M^{\frac{1}{k}}}{\tau} \cdot \frac{1}{\sqrt{d_1 d_2}} \sum_{j_0=1}^{d_1} \sum_{k_0=1}^{d_2} |u_{j_0} v_{k_0}| \leq \frac{R^2 + M^{\frac{1}{k}}}{\tau}. \end{aligned}$$

Note that the inequality above holds for any  $(\mathbf{u}, \mathbf{v})$ . Therefore

$$\|E((\tilde{Y}_i - Y_i)\mathbf{X}_i)\|_{op} \leq \frac{R^2 + M^{\frac{1}{k}}}{\tau}. \quad (\text{A.21})$$

Now we give an upper bound of the third term on the RHS of (A.11). Denote  $\langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle \mathbf{X}_i - E \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle \mathbf{X}_i$  by  $\mathbf{B}_i$ . It is not hard to verify that  $\|\mathbf{B}_i\|_{op} \leq 2R\sqrt{d_1 d_2}$  and  $\max(\|E \mathbf{B}_i^T \mathbf{B}_i\|_{op}, \|E \mathbf{B}_i \mathbf{B}_i^T\|_{op}) \leq (d_1 \vee d_2)R^2$ , it follows that for any  $t \in \mathbb{R}$ ,

$$P\left(\left\|\frac{1}{N} \sum_{i=1}^N \langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle \mathbf{X}_i - E Y_i \mathbf{X}_i\right\|_{op} \geq t\right) \leq (d_1 + d_2) \exp\left(\frac{-Nt^2/2}{(d_1 \vee d_2)R^2 + 2R\sqrt{d_1 d_2}t/3}\right). \quad (\text{A.22})$$

Finally, choose  $\tau = \sqrt{N/((d_1 \vee d_2) \log(d_1 + d_2))}$ . Combining (A.20), (A.21) and (A.22), it is easy to verify that for any  $\delta > 0$ , there exist constants  $\nu$  and  $\gamma$  such that the conclusion holds as long as  $(d_1 \vee d_2) \log(d_1 + d_2)/N < \gamma$ .



□

*Proof of Theorem 4.* This proof essentially follows the proof of Lemma 1 in Negahban and Wainwright (2012). Write  $(d_1 + d_2)/2$  as  $d$ . Define a constraint set

$$\mathcal{C}(N; c_0) = \left\{ \Delta \in \mathbb{R}^{d_1 \times d_2}, \Delta \neq \mathbf{0} \mid \sqrt{d_1 d_2} \frac{\|\Delta\|_{\max} \cdot \|\Delta\|_N}{\|\Delta\|_F^2} \leq \frac{1}{c_0 L} \sqrt{\frac{N}{(d_1 + d_2) \log(d_1 + d_2)}} \right\}.$$

According to Case 1 in the proof of Lemma 1 in Negahban and Wainwright (2012), if  $\hat{\Delta} \notin \mathcal{C}(N; c_0)$ , we have

$$\|\hat{\Delta}\|_F^2 \leq 2c_0 R \sqrt{\frac{d \log d}{N}} \{8\sqrt{r} \|\hat{\Delta}\|_F + 4 \sum_{j=r+1}^{d_1 \wedge d_2} \sigma_j(\Theta^*)\},$$

Following the same strategies in the proof of Theorem 1 in our work, we have

$$\|\hat{\Delta}\|_F \leq c_1 \sqrt{\rho} \left( 2c_0 R \sqrt{\frac{d \log d}{N}} \right)^{1-\frac{q}{2}}$$

for some constant  $c_1$ . If  $\hat{\Delta} \in \mathcal{C}(N; c_0)$ , according to Case 2 in proof of Lemma 1 in Negahban and Wainwright (2012), with probability at least  $1 - C_1 \exp(-C_2 d \log d)$ , either  $\|\hat{\Delta}\|_F \leq 512R/\sqrt{N}$ , or  $\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{\mathbf{X}\mathbf{X}} \text{vec}(\hat{\Delta}) \geq (1/256) \|\hat{\Delta}\|_F^2$ , where  $C_1$  and  $C_2$  are certain constants. For the case where  $\|\hat{\Delta}\|_F \leq 512R/\sqrt{N}$ , combining this fact with (A.5) and (A.6) delivers that

$$\|\hat{\Delta}\|_N \leq 4\sqrt{2\rho}\tau^{-\frac{q}{2}} \frac{512R}{\sqrt{N}} + 4\tau^{1-q}\rho.$$

We minimize the RHS of the inequality above by plugging in  $\tau = \left(\frac{R^2}{\rho N}\right)^{\frac{1}{2-q}}$ . Then we have for some constant  $c_3 > 0$ ,

$$\|\hat{\Delta}\|_N \leq c_3 \left( \rho \left( \frac{R^2}{N} \right)^{1-q} \right)^{\frac{1}{2-q}}.$$

For the case where  $\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{\mathbf{X}\mathbf{X}} \text{vec}(\hat{\Delta}) \geq (1/256) \|\hat{\Delta}\|_F^2$ , it is implied by Theorem 1 in our work that

$$\|\hat{\Delta}\|_F \leq c_4 \sqrt{\rho} \lambda_N^{1-\frac{q}{2}} \quad \text{and} \quad \|\hat{\Delta}\|_N \leq c_5 \rho \lambda_N^{\frac{1-q}{2}}$$

for some constants  $c_4$  and  $c_5$ . Since  $\lambda_N = \nu \sqrt{\delta(d_1 \vee d_2) \log(d_1 + d_2)/N}$  and  $R \geq 1$ , by

Lemma 4, it holds with probability at least  $1 - C_1 \exp(-C_2(d_1 + d_2)) - 2(d_1 + d_2)^{1-\delta}$  that

$$\|\widehat{\Delta}\|_F^2 \leq C_3 \max\left\{\rho\left(\frac{\delta R^2(d_1 + d_2) \log(d_1 + d_2)}{N}\right)^{1-\frac{q}{2}}, \frac{R^2}{N}\right\}$$

and

$$\|\widehat{\Delta}\|_N \leq C_4 \max\left\{\rho\left(\frac{\delta R^2(d_1 + d_2) \log(d_1 + d_2)}{N}\right)^{\frac{1-q}{2}}, \left(\frac{\rho R^{2-2q}}{N^{1-q}}\right)^{\frac{1}{2-q}}\right\}$$

for constants  $C_3$  and  $C_4$ .

□

*Proof of Lemma 5.* (a) First we study the sub-Gaussian design. Since  $\tilde{\mathbf{x}}_i = \mathbf{x}_i$ , we do not add tildes above  $\mathbf{x}, \mathbf{x}_i$  or  $x_{ij}$  in this proof. Denote  $\mathbf{y}_i(\|\mathbf{y}_i\|_2 \wedge \tau)/\|\mathbf{y}_i\|_2$  by  $\tilde{\mathbf{y}}_i$ , then we have

$$\begin{aligned} \|\widehat{\Sigma}_{\mathbf{x}\tilde{\mathbf{y}}}(\tau) - \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^T \Theta^*\|_{op} &= \|\widehat{\Sigma}_{\tilde{\mathbf{y}}\mathbf{x}}(\tau) - \frac{1}{n} \sum_{j=1}^n \Theta^{*T} \mathbf{x}_j \mathbf{x}_j^T\|_{op} \\ &\leq \|\widehat{\Sigma}_{\tilde{\mathbf{y}}\mathbf{x}}(\tau) - \mathbb{E}(\tilde{\mathbf{y}}\mathbf{x}^T)\|_{op} + \|\mathbb{E}(\tilde{\mathbf{y}}\mathbf{x}^T - \mathbf{y}\mathbf{x}^T)\|_{op} + \|\widehat{\Sigma}_{\mathbf{x}\mathbf{x}}\Theta^* - \mathbb{E}(\mathbf{y}\mathbf{x}^T)\|_{op} \end{aligned} \quad (\text{A.23})$$

Let

$$\mathbf{S}_i = \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T \\ \mathbf{x}_i \tilde{\mathbf{y}}_i^T - \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T & \mathbf{0} \end{bmatrix}.$$

Now we bound  $\|\mathbb{E} \mathbf{S}_i^p\|_{op}$  for  $p > 2$ . When  $p$  is even, i.e.,  $p = 2l$ , we have

$$\mathbf{S}_i^{2l} = \begin{bmatrix} ((\tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T)(\mathbf{x}_i \tilde{\mathbf{y}}_i - \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i))^l & \mathbf{0} \\ \mathbf{0} & ((\mathbf{x}_i \tilde{\mathbf{y}}_i - \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i)(\tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T))^l \end{bmatrix}.$$

For any  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ ,

$$\begin{aligned} \mathbf{v}^T (\mathbb{E}(\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T))^l \mathbf{v} &= \mathbb{E}((\mathbf{x}_i^T \mathbf{x}_i)^l (\tilde{\mathbf{y}}_i^T \tilde{\mathbf{y}}_i)^{l-1} (\mathbf{v}^T \tilde{\mathbf{y}}_i)^2) \leq \tau^{2l-2} \mathbb{E}((\mathbf{x}_i^T \mathbf{x}_i)^l (\mathbf{v}^T \tilde{\mathbf{y}}_i)^2) \\ &\leq \tau^{2l-2} \mathbb{E}((\mathbf{x}_i^T \mathbf{x}_i)^l (\mathbf{v}^T \mathbf{y}_i)^2) = \tau^{2l-2} \mathbb{E}(\mathbb{E}[(\mathbf{v}^T \mathbf{y}_i)^2 | \mathbf{x}_i] (\mathbf{x}_i^T \mathbf{x}_i)^l) \\ &= \tau^{2l-2} \mathbb{E}\left((\mathbf{v}^T \Theta^{*T} \mathbf{x}_i)^2 (\mathbf{x}_i^T \mathbf{x}_i)^l + \mathbb{E}[(\mathbf{v}^T \boldsymbol{\epsilon}_i)^2 | \mathbf{x}_i] (\mathbf{x}_i^T \mathbf{x}_i)^l\right). \end{aligned} \quad (\text{A.24})$$

Also note that for any  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ , we have

$$R\|\mathbf{v}\|_2^2 \geq \mathbf{v}^T \mathbb{E} \mathbf{y}_i \mathbf{y}_i^T \mathbf{v} \geq (\Theta^* \mathbf{v})^T \Sigma_{\mathbf{x}\mathbf{x}} (\Theta^* \mathbf{v}) \geq \kappa_{\mathcal{L}} \|\Theta^{*T} \mathbf{v}\|_2^2. \quad (\text{A.25})$$

Therefore it follows that  $\|\Theta^*\|_{op} \leq \sqrt{R/\kappa_{\mathcal{L}}}$ . Combining this with the fact that  $\|\mathbf{x}_i^T \mathbf{x}_i\|_{\psi_1} \leq 2d_1\kappa_0^2$ , it follows that

$$\mathbb{E} \left( (\mathbf{v}^T \Theta^{*T} \mathbf{x}_i)^2 (\mathbf{x}_i^T \mathbf{x}_i)^l \right) \leq \sqrt{\mathbb{E}(\mathbf{v}^T \Theta^{*T} \mathbf{x}_i)^4} \sqrt{\mathbb{E}(\mathbf{x}_i^T \mathbf{x}_i)^{2l}} \leq (2\sqrt{R/\kappa_{\mathcal{L}}\kappa_0})^2 (4d_1\kappa_0^2 l)^l.$$

Also by Hölder Inequality, we have

$$\begin{aligned} \mathbb{E} \left( \mathbb{E}[(\mathbf{v}^T \boldsymbol{\epsilon}_i)^2 | \mathbf{x}_i] (\mathbf{x}_i^T \mathbf{x}_i)^l \right) &\leq \mathbb{E} \left( \mathbb{E}[(\mathbf{v}^T \mathbf{x}_i)^2 | \mathbf{x}_i]^k \right)^{\frac{1}{k}} \left( \mathbb{E}[(\mathbf{x}_i^T \mathbf{x}_i)^{\frac{lk}{k-1}}] \right)^{1-\frac{1}{k}} \\ &\leq M^{\frac{1}{k}} (2d_1\kappa_0^2 \frac{lk}{k-1})^l. \end{aligned}$$

Therefore we have

$$\|\mathbb{E}(\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T)\|_{op} \leq c'_1 (c_1 \tau)^{2l-2} (d_1 l)^l \quad (\text{A.26})$$

for  $l \geq 1$ , where  $c_1$  and  $c'_1$  are some constants. Letting  $l = 1$  in the equation above implies that  $\|\mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T\|_{op} \leq \|\mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T\|_{op} \leq c'_1 d_1$ . Therefore it holds that  $(\tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T)(\mathbf{x}_i \tilde{\mathbf{y}}_i - \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T) \preceq 2\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T + 2\mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T \preceq 2\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T + 2c'_1 d_1 \mathbf{I}$ . In addition, for any commutable positive semi-definite (PSD) matrices  $\mathbf{A}$  and  $\mathbf{B}$ , it is true that  $(\mathbf{A} + \mathbf{B})^l \preceq 2^{l-1}(\mathbf{A}^l + \mathbf{B}^l)$  for  $l > 1$ . Then it follows that

$$\begin{aligned} ((\tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T)(\mathbf{x}_i \tilde{\mathbf{y}}_i^T - \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T))^l &\preceq (2\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T + 2c'_1 d_1 \mathbf{I})^l \\ &\preceq 2^{2l-1}((\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^l + (c'_1 d_1)^l \mathbf{I}). \end{aligned} \quad (\text{A.27})$$

Therefore we have  $\|\mathbb{E}((\tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T)(\mathbf{x}_i \tilde{\mathbf{y}}_i^T - \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T))^l\|_{op} \leq c'_2 (c_2 \tau)^{2l-2} (d_1 l)^l$  for some constants  $c_2$  and  $c'_2$ . Using similar methods, we can derive  $\|\mathbb{E}((\mathbf{x}_i \tilde{\mathbf{y}}_i^T - \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T)(\tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T))^l\|_{op} \leq c'_3 (c_3 \tau)^{2l-2} (d_2 l)^l$  for some constants  $c_3$  and  $c'_3$ . So we can achieve for some constants  $c_4$  and  $c'_4$ ,

$$\|\mathbb{E} \mathbf{S}_i^{2l}\|_{op} \leq c'_4 (c_4 \tau \sqrt{d_1 \vee d_2})^{2l-2} l^l (d_1 \vee d_2) \quad (\text{A.28})$$

When  $p$  is an odd number, i.e.,  $p = 2l + 1$ , by (A.27) we have

$$\begin{aligned} \mathbf{S}_i^{2l+1} &= \sqrt{\mathbf{S}_i^{4l+2}} \\ &\preceq 2^{2l+1/2} \text{diag}(\sqrt{(\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^{2l+1} + (\mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^{2l+1}}, \sqrt{(\mathbf{x}_i \tilde{\mathbf{y}}_i^T \tilde{\mathbf{y}}_i \mathbf{x}_i^T)^{2l+1} + (\mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T \mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T)^{2l+1}}) \\ &\preceq 2^{2l+1/2} \text{diag}((\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^{l+1/2} + (\mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^{l+1/2}, (\mathbf{x}_i \tilde{\mathbf{y}}_i^T \tilde{\mathbf{y}}_i \mathbf{x}_i^T)^{l+1/2} + (\mathbb{E} \tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbb{E} \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^{l+1/2}). \end{aligned}$$

Note that  $(\mathbf{x}_i \tilde{\mathbf{y}}_i^T \tilde{\mathbf{y}}_i \mathbf{x}_i^T)^{\frac{1}{2}} = \mathbf{x}_i \tilde{\mathbf{y}}_i^T \tilde{\mathbf{y}}_i \mathbf{x}_i^T / (\|\mathbf{x}_i\|_2 \|\tilde{\mathbf{y}}_i\|_2)$  and  $(\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^{\frac{1}{2}} = \tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T / (\|\mathbf{x}_i\|_2 \|\tilde{\mathbf{y}}_i\|_2)$ , so for any  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ , we have

$$\mathbf{v}^T \mathbb{E}(\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^{l+1/2} \mathbf{v} = \mathbb{E}((\mathbf{x}_i^T \mathbf{x}_i)^{l+\frac{1}{2}} (\tilde{\mathbf{y}}_i^T \tilde{\mathbf{y}}_i)^{l-\frac{1}{2}} (\mathbf{v}^T \tilde{\mathbf{y}}_i)^2) \leq \tau^{2l-1} \mathbb{E}((\mathbf{x}_i^T \mathbf{x}_i)^{l+\frac{1}{2}} (\mathbf{v}^T \tilde{\mathbf{y}}_i)^2).$$

Following the same steps as in the case of  $l$  being even, we can derive  $\|\mathbb{E}(\tilde{\mathbf{y}}_i \mathbf{x}_i^T \mathbf{x}_i \tilde{\mathbf{y}}_i^T)^{l+1/2}\|_{op} \leq c'_5 (c_5 \tau)^{2l-1} (d_1(l+1/2))^{l+1/2}$ ,  $\|\mathbb{E}(\mathbf{x}_i \tilde{\mathbf{y}}_i^T \tilde{\mathbf{y}}_i \mathbf{x}_i^T)^{l+1/2}\|_{op} \leq c'_6 (c_6 \tau)^{2l-1} (d_2(l+1/2))^{l+1/2}$  and finally

$$\|\mathbb{E} \mathbf{S}_i^{2l+1}\|_{op} \leq c'_7 (c_7 \tau \sqrt{d_1 \vee d_2})^{2l-1} (l+1/2)^{l+1/2} (d_1 \vee d_2). \quad (\text{A.29})$$

Define  $\sigma^2 = (c_7 \vee c'_7)(d_1 \vee d_2)$ . Then by (A.28), we have  $\|\mathbb{E} \mathbf{S}_i^2\|_{op} \leq \sigma^2$ . Also by combining (A.28) and (A.29), we can verify the moment constraints of Lemma 6; for  $p > 2$ , we have

$$\|\mathbb{E} \mathbf{S}_i^p\|_{op} \leq p! ((c_4 \vee c_7) \tau \sqrt{d_1 \vee d_2})^{p-2} \sigma^2. \quad (\text{A.30})$$

Therefore by Lemma 6, we have

$$\begin{aligned} P\left(\|\widehat{\Sigma}_{\tilde{\mathbf{y}}\mathbf{x}}(\tau) - \mathbb{E}(\tilde{\mathbf{y}}\mathbf{x}^T)\|_{op} \geq t\right) &\leq P\left(\left\|\frac{1}{n} \sum_{i=1}^n \mathbf{S}_i\right\|_{op} \geq t\right) \\ &\leq (d_1 + d_2) \cdot \exp\left(-c \min\left(\frac{nt^2}{\sigma^2}, \frac{nt}{(c_4 \vee c_7) \tau \sqrt{d_1 \vee d_2}}\right)\right). \end{aligned} \quad (\text{A.31})$$

Next we aim to bound  $\|\mathbb{E}(\tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbf{y}_i \mathbf{x}_i^T)\|_{op}$ . Note that for any  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ ,

$$\begin{aligned} \mathbb{E}((\mathbf{v}^T \tilde{\mathbf{y}}_i)^2 (\mathbf{u}^T \mathbf{x}_i)^2) &= \mathbb{E}(\mathbb{E}((\mathbf{v}^T \tilde{\mathbf{y}}_i)^2 | \mathbf{x}_i) (\mathbf{u}^T \mathbf{x}_i)^2) = \mathbb{E}\left(\mathbb{E}((\mathbf{v}^T \boldsymbol{\Theta}^* \mathbf{x}_i)^2 (\mathbf{u}^T \mathbf{x}_i)^2) + \mathbb{E}((\mathbf{v}^T \boldsymbol{\epsilon}_i)^2 | \mathbf{x}_i) (\mathbf{u}^T \mathbf{x}_i)^2\right) \\ &:= v^2 < \infty. \end{aligned}$$

For any  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ , by the Cauchy-Schwartz inequality and the Markov inequality,

$$\begin{aligned} \mathbb{E}((\mathbf{v}^T \tilde{\mathbf{y}}_i)(\mathbf{u}^T \mathbf{x}_i) - (\mathbf{v}^T \mathbf{y}_i)(\mathbf{u}^T \mathbf{x}_i)) &\leq \mathbb{E}(|(\mathbf{v}^T \tilde{\mathbf{y}}_i)(\mathbf{u}^T \mathbf{x}_i)| \cdot 1_{\|\mathbf{y}\|_2 > \tau}) \leq \sqrt{\mathbb{E}(|(\mathbf{v}^T \tilde{\mathbf{y}}_i)(\mathbf{u}^T \mathbf{x}_i)|^2) P(\|\mathbf{y}\|_2 > \tau)} \\ &\leq \sqrt{\frac{v^2 \mathbb{E}\|\mathbf{y}\|_2^2}{\tau^2}} = \frac{v}{\tau} \sqrt{R^2 \mathbb{E}\|\mathbf{x}\|_2^2 + \mathbb{E}\|\boldsymbol{\epsilon}\|_2^2} \leq \frac{v \sqrt{d_2}}{\tau} \sqrt{R^2 \kappa_0^2 + M^{\frac{1}{k}}}. \end{aligned}$$

Since  $\mathbf{u}$  and  $\mathbf{v}$  are arbitrary, we have

$$\|\mathbb{E}(\tilde{\mathbf{y}}_i \mathbf{x}_i^T - \mathbf{y}_i \mathbf{x}_i^T)\|_{op} \leq \frac{v\sqrt{d_2}}{\tau} \sqrt{R^2 \kappa_0^2 + M^{\frac{1}{k}}}. \quad (\text{A.32})$$

Now we give an upper bound of the third term on the RHS of (A.23). For this term, we still follow the covering argument. Let  $\mathcal{N}^{d-1}$  be the  $1/4$ -net on  $\mathcal{S}^{d-1}$ . For any  $\mathbf{u} \in \mathcal{S}^{d_1-1}$  and  $\mathbf{v} \in \mathcal{S}^{d_2-1}$ ,  $\|\mathbf{u}^T \mathbf{x}_i \mathbf{x}_i^T \boldsymbol{\Theta}^* \mathbf{v}\|_{\psi_1} \leq \sqrt{R/\kappa_{\mathcal{L}} \kappa_0^2}$ , so by Proposition 5.16 (Bernstein-type inequality) in Vershynin (2010) it follows that for sufficiently small  $t$ ,

$$P\left(\left|\mathbf{u}^T \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}} \boldsymbol{\Theta}^* \mathbf{v} - \mathbb{E}(\mathbf{u}^T \mathbf{x}_i)(\mathbf{v}^T \mathbf{y}_i)\right| > t\right) \leq 2 \exp(-c_8 n t^2),$$

where  $c_8$  is some positive constant. Then an combination of the union bound over all points on  $\mathcal{N}^{d_1-1} \times \mathcal{N}^{d_2-1}$  and (A.12) delivers

$$P\left(\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}} \boldsymbol{\Theta}^* - \mathbb{E} \mathbf{y} \mathbf{x}^T\|_2 \geq \frac{16}{7} t\right) \leq 2 \exp((d_1 + d_2) \log 8 - c_8 n t^2), \quad (\text{A.33})$$

where  $c_8$  is a constant.

Finally we choose  $\tau = O(\sqrt{n/\log(d_1 + d_2)})$ , and combining (A.31), (A.32) and (A.33) delivers that for any  $\delta > 0$ , as long as  $(d_1 + d_2) \log(d_1 + d_2)/n < \gamma$  for some constant  $\gamma > 0$ ,

$$P\left(\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\tilde{\mathbf{y}}}(\tau_1) - \frac{1}{n} \sum_{j=1}^n \boldsymbol{\Theta}^{*T} \mathbf{x}_j \mathbf{x}_j^T\|_{op} \geq \sqrt{\frac{(\nu + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n}}\right) \leq 2(d_1 + d_2)^{1-\eta\delta},$$

where  $\nu$  and  $\eta$  are universal constants.

(b) Now we switch to the case of designs with only bounded moments. We first show that for two constants  $c_3$  and  $c_4$ ,

$$P\left(\|\widehat{\boldsymbol{\Sigma}}_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}} - \boldsymbol{\Sigma}_{\mathbf{xy}}\|_{op} \geq \sqrt{\frac{(c_3 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n}}\right) \leq (d_1 + d_2)^{1-c_4\delta}. \quad (\text{A.34})$$

Note that

$$\mathbb{E}(\mathbf{v}^T \mathbf{y}_i)^4 = \mathbb{E}(\mathbf{v}^T \boldsymbol{\Theta}^{*T} \mathbf{x}_i + \mathbf{v}^T \boldsymbol{\epsilon}_i)^4 \leq 16(\mathbb{E}(\mathbf{v}^T \boldsymbol{\Theta}^{*T} \mathbf{x}_i)^4 + \mathbb{E}(\mathbf{v}^T \boldsymbol{\epsilon}_i)^4).$$

By Rosenthal-type inequality in Ibragimov and Sharakhmetov (1998), we have

$$\mathbb{E}(\mathbf{v}^T \mathbf{y}_i)^4 \leq 16(16R^4 \kappa_0^4 + M) =: V_1.$$

In addition,  $\mathbb{E}(\mathbf{v}^T \mathbf{x}_i)^4 \leq 16\kappa_0^4 =: V_2$ . Let  $V := \max(V_1, V_2)$ ,  $\mathbf{Z}_i := \mathbf{x}_i \mathbf{y}_i^T$  and  $\tilde{\mathbf{X}}_i := \tilde{\mathbf{x}}_i \tilde{\mathbf{y}}_i^T$ . Following the proof strategy for Theorem 6, we have

$$\|\tilde{\mathbf{Z}}_i - \mathbb{E} \tilde{\mathbf{Z}}_i\|_{op} \leq \|\tilde{\mathbf{Z}}_i\|_{op} + \|\mathbb{E} \tilde{\mathbf{Z}}_i\|_{op} = \|\tilde{\mathbf{x}}_i\|_2 \|\tilde{\mathbf{y}}_i\|_2 + \sqrt{R} \leq (d_1 d_2)^{\frac{1}{4}} \tau^2 + \sqrt{V}.$$

Also for any  $\mathbf{v} \in \mathcal{S}^{d-1}$ , we have

$$\begin{aligned} \mathbb{E}(\mathbf{v}^T \tilde{\mathbf{Z}}_i^T \tilde{\mathbf{Z}}_i \mathbf{v}) &= \mathbb{E}(\|\tilde{\mathbf{x}}_i\|_2^2 (\mathbf{v}^T \tilde{\mathbf{y}}_i)^2) \leq \mathbb{E}(\|\mathbf{x}_i\|_2^2 (\mathbf{v}^T \mathbf{y}_i)^2) \\ &= \sum_{j=1}^d \mathbb{E}(x_{ij}^2 (\mathbf{v}^T \mathbf{y}_i)^2) \leq \sum_{j=1}^d \sqrt{\mathbb{E}(x_{ij}^4) \mathbb{E}(\mathbf{v}^T \mathbf{y}_i)^4} \leq V d_1. \end{aligned}$$

Then it follows that  $\|\mathbb{E} \tilde{\mathbf{Z}}_i^T \tilde{\mathbf{Z}}_i\|_{op} \leq V d_1$ . Similarly we can obtain  $\|\mathbb{E} \tilde{\mathbf{Z}}_i \tilde{\mathbf{Z}}_i^T\|_{op} \leq V d_2$ . Denote  $d_1 + d_2$  by  $d$ . Since  $\max(\|(\mathbb{E} \tilde{\mathbf{X}}_i)^T \mathbb{E} \tilde{\mathbf{X}}_i\|_{op}, \|(\mathbb{E} \tilde{\mathbf{Z}}_i) \mathbb{E} \tilde{\mathbf{Z}}_i^T\|_{op}) \leq \|\mathbb{E} \mathbf{Z}_i\|_{op}^2 \leq V$ ,

$$\max(\|\mathbb{E}((\tilde{\mathbf{Z}}_i - \mathbb{E} \tilde{\mathbf{Z}}_i)^T (\tilde{\mathbf{Z}}_i - \mathbb{E} \tilde{\mathbf{Z}}_i))\|_{op}, \|\mathbb{E}((\tilde{\mathbf{Z}}_i - \mathbb{E} \tilde{\mathbf{Z}}_i)(\tilde{\mathbf{Z}}_i - \mathbb{E} \tilde{\mathbf{Z}}_i)^T)\|_{op}) \leq V(d+1).$$

By Corollary 6.2.1 in Tropp (2012), we have for some constant  $c_1$ ,

$$P\left(\left\|\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{Z}}_i - \mathbb{E} \tilde{\mathbf{Z}}_i\right\|_{op} > t\right) \leq d \exp\left(-c_1\left(\frac{nt^2}{V(d+1)} \wedge \frac{nt}{\sqrt{d\tau_1\tau_2} + \sqrt{V}}\right)\right). \quad (\text{A.35})$$

Now we bound the bias  $\mathbb{E}\|\mathbf{Z}_i - \tilde{\mathbf{Z}}_i\|_{op}$ . For any  $\mathbf{v} \in \mathcal{S}^{d-1}$ , it holds that

$$\begin{aligned} \mathbb{E}(\mathbf{v}^T (\mathbf{Z}_i - \tilde{\mathbf{Z}}_i) \mathbf{v}) &= \mathbb{E}(|(\mathbf{v}^T \mathbf{x}_i)(\mathbf{v}^T \mathbf{y}_i) - (\mathbf{v}^T \tilde{\mathbf{x}}_i)(\mathbf{v}^T \tilde{\mathbf{y}}_i)| 1_{\{\|\mathbf{x}_i\|_4 \geq \tau_1 \text{ or } \|\mathbf{y}_i\|_4 \geq \tau_2\}}) \\ &\leq \mathbb{E}(|(\mathbf{v}^T \mathbf{x}_i)(\mathbf{v}^T \mathbf{y}_i)| (1_{\{\|\mathbf{x}_i\|_4 > \tau_1\}} + 1_{\{\|\mathbf{y}_i\|_4 > \tau_2\}})) \\ &\leq \sqrt{\mathbb{E}((\mathbf{v}^T \mathbf{x}_i)^2 (\mathbf{v}^T \mathbf{y}_i)^2) P(\|\mathbf{x}_i\|_4 > \tau_1)} + \sqrt{\mathbb{E}((\mathbf{v}^T \mathbf{x}_i)^2 (\mathbf{v}^T \mathbf{y}_i)^2) P(\|\mathbf{y}_i\|_4 > \tau_2)} \\ &\leq \frac{V\sqrt{d_1}}{\tau_1^2} + \frac{V\sqrt{d_2}}{\tau_2^2}. \end{aligned}$$

Therefore we have  $\|\mathbb{E}(\mathbf{Z}_i - \tilde{\mathbf{Z}}_i)\|_{op} \leq V\sqrt{d}(1/\tau_1^2 + 1/\tau_2^2)$ . Choose  $\tau_1 \asymp (n/\log d_1)^{\frac{1}{4}}$ ,  $\tau_2 \asymp$

$(n/\log d_2)^{\frac{1}{4}}$  and substitute  $t$  with  $\sqrt{\delta d \log d/n}$ . Then we obtain (A.34) by combining the bound of bias and (A.35).

Finally, note that

$$\|\widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}} - \widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} \Theta^*\|_{op} \leq \|\widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{y}}} - \Sigma_{\mathbf{x}\mathbf{y}}\|_{op} + \|(\widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}} - \Sigma_{\mathbf{x}\mathbf{x}}) \Theta^*\|_{op} + \|\Sigma_{\mathbf{x}\mathbf{y}} - \Sigma_{\mathbf{x}\mathbf{x}} \Theta^*\|_{op}.$$

Combining consistency of  $\widehat{\Sigma}_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}$  established in Theorem 6, Condition (C1) and (A.34), we can reach the conclusion of the lemma.  $\square$

**Lemma 6** (Matrix Bernstein Inequality with Moment Constraint). *Consider a finite sequence  $\{\mathbf{S}_i\}_{i=1}^n$  of independent, random, Hermitian matrices with dimensions  $d \times d$ . Assume that  $\mathbb{E} \mathbf{S}_i = 0$  and  $\|\mathbb{E} \mathbf{S}_i^2\|_{op} \leq \sigma^2$ . Also the following moment conditions hold for all  $1 \leq i \leq n$  and  $p \geq 2$ :*

$$\|\mathbb{E} \mathbf{S}_i^p\|_{op} \leq p! L^{p-2} \sigma^2,$$

where  $L$  is a constant. Then for every  $t \geq 0$  we have

$$P\left(\left\|\frac{1}{n} \sum_{i=1}^n \mathbf{S}_i\right\|_{op} \geq t\right) \leq d \exp\left(\frac{-nt^2}{4\sigma^2 + 2Lt}\right) \leq d \cdot \exp\left(-c \min\left(\frac{nt^2}{\sigma^2}, \frac{nt}{L}\right)\right).$$

Proof of Lemma 6. Given the moment constraints, we have for  $0 < \theta < 1/L$ ,

$$\mathbb{E} e^{\theta \mathbf{S}_i} = \mathbf{I} + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E} \mathbf{S}_i^p}{p!} \preceq \mathbf{I} + \sum_{p=2}^{\infty} \sigma^2 L^{p-2} \theta^p \mathbf{I} = \mathbf{I} + \frac{\theta^2 \sigma^2}{1 - \theta L} \mathbf{I} \preceq \exp\left(\frac{\theta^2 \sigma^2}{1 - \theta L}\right) \mathbf{I}.$$

Let  $g(\theta) = \theta^2/(1 - \theta L)$ . Owing to the master tail inequality (Theorem 3.6.1 in Tropp (2015)), we have

$$\begin{aligned} P\left(\left\|\sum_{i=1}^n \mathbf{S}_i\right\|_{op} \geq t\right) &\leq \inf_{\theta > 0} e^{-\theta t} \operatorname{tr} \exp\left(\sum_i \log \mathbb{E} e^{\theta \mathbf{S}_i}\right) \\ &\leq \inf_{0 < \theta < 1/L} e^{-\theta t} \operatorname{tr} \exp\left(n \sigma^2 g(\theta) \mathbf{I}\right) \\ &\leq \inf_{0 < \theta < 1/L} d e^{-\theta t} \exp\left(n \sigma^2 g(\theta)\right). \end{aligned}$$

Choosing  $\theta = t/(2n\sigma^2 + Lt)$ , we can reach the conclusion.  $\square$

Proof of Theorem 5. (a) First consider the case of the sub-Gaussian design. Denote  $\frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^T$  by  $\overline{\Sigma}_{\mathbf{x}\mathbf{x}}$ . Since  $\mathbf{x}_j$  is a sub-Gaussian vector, an application of Theorem 5.39 in Vershynin

(2010) implies that with probability at least  $1 - \exp(-c_1 d_1)$ ,  $\|\bar{\Sigma}_{\mathbf{xx}} - \Sigma_{\mathbf{xx}}\|_{op} \leq c_2 \sqrt{d_1/n}$  and furthermore  $\lambda_{\min}(\bar{\Sigma}_{\mathbf{xx}}) \geq \frac{1}{2} \lambda_{\min}(\Sigma_{\mathbf{xx}}) > 0$  as long as  $d_1/n$  is sufficiently small. Therefore

$$\begin{aligned}
\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{\mathbf{xx}} \text{vec}(\hat{\Delta}) &= \frac{1}{N} \sum_{i=1}^N \langle \hat{\Delta}, \mathbf{X}_i \rangle^2 = \frac{1}{N} \sum_{j=1}^n \sum_{k=1}^{d_2} \langle \hat{\Delta}, \mathbf{x}_j \mathbf{e}_k^T \rangle^2 \\
&= \frac{1}{N} \sum_{j=1}^n \sum_{k=1}^{d_2} \text{Tr}(\mathbf{x}_j^T \hat{\Delta} \mathbf{e}_k)^2 = \frac{1}{N} \sum_{k=1}^{d_2} \sum_{j=1}^n (\mathbf{x}_j^T \hat{\Delta} \mathbf{e}_k)^2 \\
&= \frac{1}{d_2} \sum_{k=1}^{d_2} (\hat{\Delta} \mathbf{e}_k)^T \bar{\Sigma}_{\mathbf{xx}} (\hat{\Delta} \mathbf{e}_k) \geq \frac{1}{2d_2} \lambda_{\min}(\Sigma_{\mathbf{xx}}) \|\hat{\Delta}\|_F^2,
\end{aligned} \tag{A.36}$$

with probability at least  $1 - \exp(-c_1 d_1)$ . In addition, by Lemma 5, as long as we choose  $\lambda_N = \frac{2}{d_2} \sqrt{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)/n}$ ,  $\lambda_N \geq 2 \|\hat{\Sigma}_{Y\mathbf{X}} - \frac{1}{N} \sum_{i=1}^N Y_i \mathbf{X}_i\|_{op}$  with probability at least  $1 - 2(d_1 + d_2)^{1-\eta\delta}$ . Finally, by Theorem 1 we establish the statistical error rate: there exist  $\gamma_1, \gamma_2 > 0$  that as long as  $(d_1 + d_2) \log(d_1 + d_2)/n < \gamma_1$  and  $d_1 + d_2 > \gamma_2$ , we have

$$\|\hat{\Theta}(\lambda_N, \tau) - \Theta^*\|_F^2 \leq C_1 \rho \left( \frac{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{1-q/2}$$

and

$$\|\hat{\Theta}(\lambda_N, \tau) - \Theta^*\|_N \leq C_2 \rho \left( \frac{(\nu_1 + \delta)(d_1 + d_2) \log(d_1 + d_2)}{n} \right)^{\frac{1-q}{2}}$$

for some constants  $C_1$  and  $C_2$ .

(b) Now we switch to the designs with bounded moments. According to Theorem 6, with probability at least  $1 - d_1^{1-C\delta}$ ,

$$\|\hat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} - \Sigma_{\mathbf{xx}}\|_{op} \leq \sqrt{\frac{\delta d_1 \log d_1}{n}},$$

which furthermore implies that  $\lambda_{\min}(\hat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}) \geq \frac{1}{2} \lambda_{\min}(\Sigma_{\mathbf{xx}}) > 0$  as long as  $d_1/n$  is sufficiently small. Analogous to (A.36), we therefore have

$$\text{vec}(\hat{\Delta})^T \hat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \text{vec}(\hat{\Delta}) \geq \frac{1}{2d_2} \lambda_{\min}(\Sigma_{\mathbf{xx}}) \|\hat{\Delta}\|_F^2.$$

Combining Theorem 1 with the inequality above, we can reach the final conclusion. □



Proof of Theorem 6. We denote  $\mathbf{x}_i \mathbf{x}_i^T$  by  $\mathbf{X}_i$  and  $\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T$  by  $\tilde{\mathbf{X}}_i$  for ease of notations. Note that

$$\|\tilde{\mathbf{X}}_i - \mathbb{E} \tilde{\mathbf{X}}_i\|_{op} \leq \|\tilde{\mathbf{X}}_i\|_{op} + \|\mathbb{E} \tilde{\mathbf{X}}_i\|_{op} = \|\tilde{\mathbf{x}}_i\|_2^2 + \sqrt{R} \leq \sqrt{d}\tau^2 + \sqrt{R}$$

Also for any  $\mathbf{v} \in \mathcal{S}^{d-1}$ , we have

$$\begin{aligned} \mathbb{E}(\mathbf{v}^T \tilde{\mathbf{X}}_i^T \tilde{\mathbf{X}}_i \mathbf{v}) &= \mathbb{E}(\|\tilde{\mathbf{x}}_i\|_2^2 (\mathbf{v}^T \tilde{\mathbf{x}}_i)^2) \leq \mathbb{E}(\|\mathbf{x}_i\|_2^2 (\mathbf{v}^T \mathbf{x}_i)^2) \\ &= \sum_{j=1}^d \mathbb{E}(x_{ij}^2 (\mathbf{v}^T \mathbf{x}_i)^2) \leq \sum_{j=1}^d \sqrt{\mathbb{E}(x_{ij}^4) \mathbb{E}(\mathbf{v}^T \mathbf{x}_i)^4} \leq Rd \end{aligned}$$

Then it follows that  $\|\mathbb{E} \tilde{\mathbf{X}}_i^T \tilde{\mathbf{X}}_i\|_{op} \leq Rd$ . Since  $\|(\mathbb{E} \tilde{\mathbf{X}}_i)^T \mathbb{E} \tilde{\mathbf{X}}_i\|_{op} \leq \|\mathbb{E} \mathbf{X}_i\|_{op}^2 \leq R$ ,  $\|\mathbb{E}((\tilde{\mathbf{X}}_i - \mathbb{E} \tilde{\mathbf{X}}_i)^T (\tilde{\mathbf{X}}_i - \mathbb{E} \tilde{\mathbf{X}}_i))\|_{op} \leq R(d+1)$ . By Theorem 5.29 (Non-commutative Bernstein-type inequality) in Vershynin (2010), we have for some constant  $c$ ,

$$P\left(\left\|\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{X}}_i - \mathbb{E} \tilde{\mathbf{X}}_i\right\|_{op} > t\right) \leq 2d \exp\left(-c\left(\frac{nt^2}{R(d+1)} \wedge \frac{nt}{\sqrt{d}\tau^2 + \sqrt{R}}\right)\right). \quad (\text{A.37})$$

Now we bound the bias  $\mathbb{E}\|\mathbf{X}_i - \tilde{\mathbf{X}}_i\|_{op}$ . For any  $\mathbf{v} \in \mathcal{S}^{d-1}$ , it holds that

$$\begin{aligned} \mathbb{E}(\mathbf{v}^T (\mathbf{X}_i - \tilde{\mathbf{X}}_i) \mathbf{v}) &= \mathbb{E}(((\mathbf{v}^T \mathbf{x}_i)^2 - (\mathbf{v}^T \tilde{\mathbf{x}}_i)^2) 1_{\{\|\mathbf{x}_i\|_4 > \tau\}}) \\ &\leq \mathbb{E}((\mathbf{v}^T \mathbf{x}_i)^2 1_{\{\|\mathbf{x}_i\|_4 > \tau\}}) \leq \sqrt{\mathbb{E}(\mathbf{v}^T \mathbf{x}_i)^4 P(\|\mathbf{x}_i\|_4 > \tau)} \\ &\leq \sqrt{\frac{R^2 d}{\tau^4}} = \frac{R\sqrt{d}}{\tau^2}. \end{aligned}$$

Therefore we have  $\|\mathbb{E}(\mathbf{X}_i - \tilde{\mathbf{X}}_i)\|_{op} \leq R\sqrt{d}/\tau^2$ . Choose  $\tau \asymp (nR/\delta \log d)^{\frac{1}{4}}$  and substitute  $t$  with  $\sqrt{\delta R d \log d/n}$ . Then we reach the final conclusion by combining the bound of bias and (A.37). □